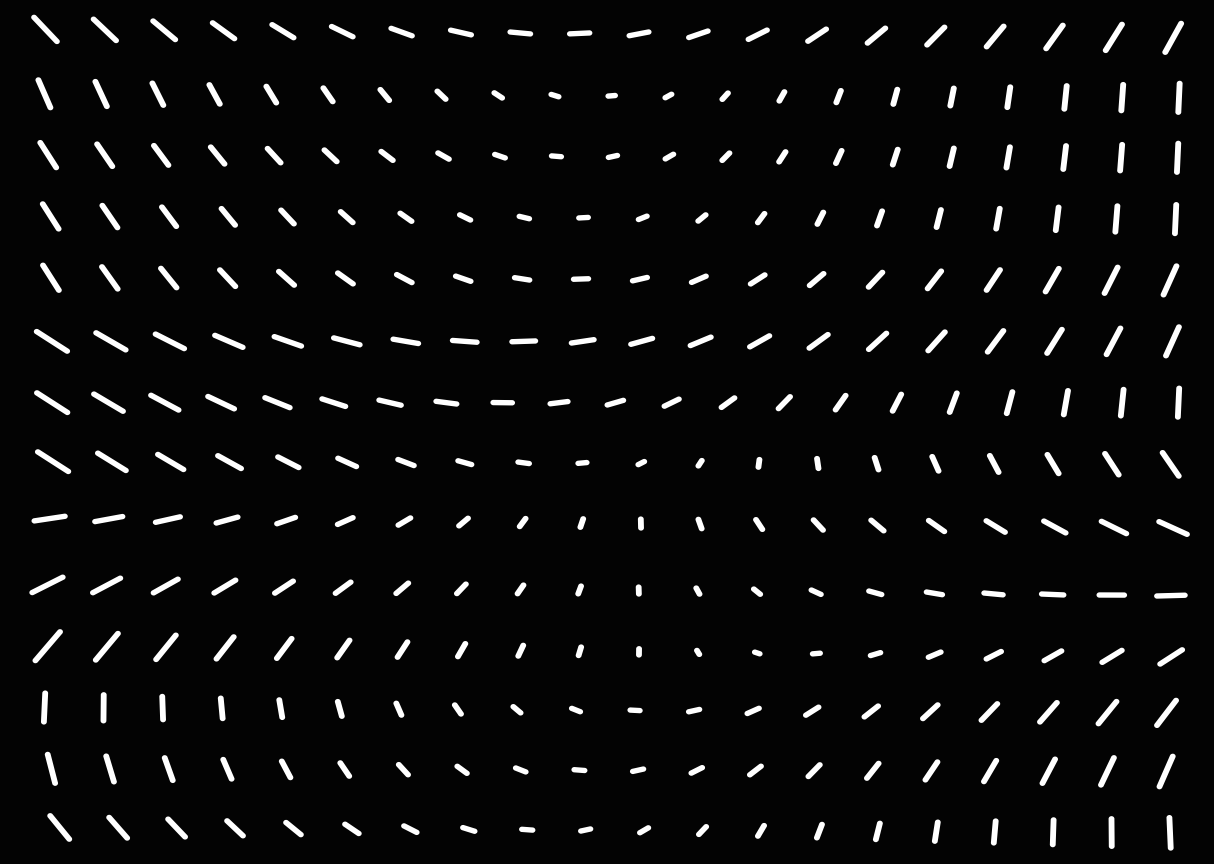
CISC 271

Pre- Req



Pre Req Lecture Video

Lecture 1- Vectors

- A vector is a finite column of real numbers $v = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$
- Vector operations:

  Scaling vectors: $v = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}$, $-3v = \begin{bmatrix} -9 \\ -15 \\ -6 \end{bmatrix}$

  Addition vectors: adding two vectors of the same size, component wise

Note: Euclidean norm or L_2 is just the magnitude used for machine learning

- Vector Norm (Magnitude)
  - The length of a vector is the squared rooted sum of all its components

$$\vec{v} = \begin{bmatrix} 5 \\ 4 \\ 1 \end{bmatrix}$$

$$\|\vec{v}\| = \sqrt{5^2 + 4^2 + 1^2}$$

$$= \sqrt{38}$$

- Vector dot Product: $u \cdot v = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n$

  Results in a scalar
- Two vectors are orthogonal if $u \cdot v = 0$
- Unit vector is a vector with magnitud equal to 1
- To make a vector, a unit vector do $\vec{u} = \frac{\vec{v}}{\|v\|}$

Linear Combinations

- A linear combination is $w = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n$, where c is scalars
- Multiply each vector component by a scalar, then add together.
- usefull for solving systems of linear equations, span, linear independence

---

Linear Independence

- A set of vectors is linearly indep if no vector can be expressed as a lin combo of the others. For a set the only solution is

$$c_1 v_1 + c_2 v_2 + c_n v_n = 0$$

Ex $v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are linearly indep

Note: Full rank means linearly indep

All c must be 0

Linearly dependent

A set of vectors is linearly dependent if atleast one vector can be expressed as a lin combo of the others

So scalars c_1, c_2, c_n atleast 1 is non-zero

At least one c is non-zero

Lesson 2- Vector Spaces

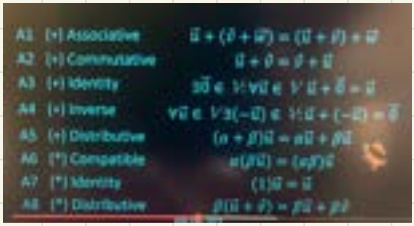
Abstract Vector Spaces

- A vector space V over a field F, is a set of vectors with two operations.

1. Vector addition:
   - $u + v \in V$ for all u, v
2. scalar multiplication:
   $cv \in V$ for all $c \in F$

Note: Dimensions are minimal number K vectors needed to construct V.

Vector space: Axioms



Interpretation of a Vector Space

For any $\vec{u} \in V$ and $\vec{v} \in V$, any $\alpha \in R$, $\beta \in R$

$$a\vec{u} + \beta\vec{v} \in V$$

Related to all axioms

Subspaces

- A subspace of a vector space V is a subset $W \subseteq V$ that is itself a vector space under same operations.

3 conditions:

1.) $0 \in W$

2.) closed under addition $u + v \in W$ → if v is in vector space, and v is in vector space $u + v$ is in vector space.

3.) closed under multiplication $cv \in W$

Math 110 Notes

Linear Combinations

a linear combination is when you are able to multiply each component of a vector by a scalar and make it equal to another vector.

Ex $v1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ $v2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$; some scalar c

① $cv1 + cv2 = \vec{B}$ ← If this is true, v1, v2 are a linear combination of vector B

Unit vectors

- unit vectors in $\mathbb{R}^2, \mathbb{R}^3$

$$\vec{U} = \frac{\vec{v}}{|\vec{v}|}$$

- have magnitude of 1:
- to get a unit vector, $\vec{v} = (3,4)$

$\vec{v} = (3,4)$, $|v| = \sqrt{3^2+4^2}$, $|v| = 5$ $\Longrightarrow$ $\vec{v} = \frac{(3,4)}{5}$

$\vec{v} = \left(\frac{3}{5}, \frac{4}{5}\right)$

$\vec{v}$ is a unit vector of $\vec{v}$.

Definitions

Linear System: is a system of equations, where the highest degree on any variable is less than or equal to one.

Rank: The rank of a matrix is the # of leading ones in RREF.

Parametrized in vector Form: $\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix} + t\begin{pmatrix} 7 \\ 7 \\ 5 \end{pmatrix}$

Parametric Form:

$x = 3 + 7t$
$y = 2 + 7t$
$z = 1 + 5t$

Span: $S = \{v_1 \ldots v_n\} \in \mathbb{R}^n$. The set of all linear combinations of vectors of S is called the span of S

REF - first leading entry is 1
- All other entries are 0.
- All zeroes, go bottom

Ex $\begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 5 \\ 0 & 0 & 6 \end{bmatrix}$

RREF - Same thing, just more reduced.

Ex $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

REF (Row Echelon Form) Tells us:

- number of free variables
- number of solutions to linear system
- rank
- system consistent/inconsistent

RREF (Reduced Row Echelon Form) Tells us

- Directly find the solutions of a linear system
- rank
- Clearer picture of relationship of variables
- number of free variables
- linear dependency / independency.

Matrix Multiplication:

Let $A = [2, 5, 6]$ Let $B = \begin{bmatrix} 3 \\ 4 \\ -5 \end{bmatrix}$

1 x 3 × 3 x 1

column of A should equal column of B

should result in a 1 by 1 matrix.

$AB \neq BA$ Order matters

Always Multiply Row by column:

① Example

$\begin{bmatrix} -2 & 1 \\ 10 & 5 \end{bmatrix}$ $\begin{bmatrix} 4 & -2 \\ 1 & -8 \end{bmatrix}$

2 by 2 x 2 by 2 will be 2 by 2.

$\begin{bmatrix} -7 & -4 \\ 45 & -60 \end{bmatrix}$ row 1 x column 1
row 1 x c2
r2, c1
r2 x c2

A square matrix A is symmetric if $A^T = A$

Example

$$A = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 5 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

When the transpose is the same as the original.

$$A^T = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 5 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

$\therefore$ A is symmetric

Determinant of a 2 by 2:

$$\begin{bmatrix} 5 & 3 \\ -1 & 4 \end{bmatrix} = 5\times4 - 3\times(-1) = 23$$

Det $= ad - bc$

$Ax = b$ has a unique solution if A has an inverse, $x = A^{-1}b$

$A(A^{-1}b) = (AA^{-1})b = Ib = b$

Find Inverse of:

$$\left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 2 & 2 & 4 & 0 & 1 & 0 \\ 1 & 3 & -3 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & -2 & 6 & 0 & 1 & 0 \\ 0 & 1 & -2 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 2 & -1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 0 & -\frac{1}{2} & 0 \\ 0 & 1 & -2 & 0 & 0 & 1 \end{array}\right]$$

$R2 = R2 - 2R1$, $R3 = R3 - R1$; $R2 = -\frac{1}{2}R2$

$R1 = R1 - 2R2$, $R3 = R3 - R2$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 5 & 1 & 0 & 0 \\ 0 & 1 & -3 & 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & -\frac{5}{2} & -5 \\ 0 & 1 & 0 & 0 & -\frac{3}{2} & 3 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \to \begin{bmatrix} 9 & -\frac{5}{2} & -5 \\ -5 & 1 & 3 \\ -2 & \frac{1}{2} & 1 \end{bmatrix}$$

$R1 = R1 - 5R3$ $R2 = R2 + 3R3$

Invertible matrices

If a matrix is invertible, it's determinant can NOT be equal to 0.

$AA^{-1} = I_n$

Ex

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

F.T.I.M

a) A is invertible

b) $A\vec{x} = \vec{b}$ has a unique solution $x = A^{-1}\vec{b}$

c) $A\vec{x} = 0$ has a unique solution.

d) $RREF(A) = I$

e) A is product of elementary matrices

$AE_1E_2E_3 = I_n$

A subspace of $\mathbb{R}^n$ is any collection S of vectors in $\mathbb{R}^n$ such that:

1. contain $\vec{0}$
2. closed under scalar addition
3. closed under multiplication

Basis is always associated with a subspace. Basis is a set of linearly independent vectors that span V.

$$\left\{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right\}$$

Linearly independent
Spans $\mathbb{R}^3$
$\therefore$ Basis of $\mathbb{R}^3$

How to check a basis of $\mathbb{R}^3$.

① Check if the given vectors are the same dimension as $\mathbb{R}^3$.
② Setup matrix, row reduce it
③ If matrix has 3 pivots and is square matrix that means it is invertible. Which means it is a basis for $\mathbb{R}^3$.

To find a basis for a set of vectors:

① Setup augmented matrix equal to 0.
② Choose linearly independent columns.
③ Look at original columns and those form a basis.

Basis - minimum vectors needed to span
Column space - set of all lin combinations.

Finding the Basis of Null Space

① Set up Augmented matrix $[A|0$
② then RREF
③ Parameterize solutions

What is null space?

- set of all vectors when multiplied by the matrix give you the $\vec{0}$.

(null(A)) $Ax = 0$

What is Standard Basis?

- set of vectors, where each component is 0, except 1.

Standard Basis in $\mathbb{R}^3$: $e1 = \begin{bmatrix}1\\0\\0\end{bmatrix}, e2 = \begin{bmatrix}0\\1\\0\end{bmatrix}, e3 = \begin{bmatrix}0\\0\\1\end{bmatrix}$

The rowspace of A is the set of all possible linear combinations of it's rows

The columnspace of A is the set of all possible linear combinations of it's columns

How to find rowspace/columnspace

① Put Matrix into RREF

Row space
Identify non zero rows. These rows gives you the basis for the row space.

column spac
Identify Pivot columns. Take the original columns gives you basis for the column space.

* row rank = col rank = rank (A)

To check if a given vector is in the rowspace:

① Set augmented matrix equal to it. Then see if it is a unique sol.

Nullity of a matrix A, is the dimension of it's null Space.

S is a subspace:
The number of vectors in a basis for S is called the dimension of S.

$\dim \mathbb{R}^n = n$

$\dim(\text{col}(A)) + \dim(\text{null } A) = \#$ of columns

The rows and columns spaces of matrix A always have the same dimension.

$\dim(\text{row}(A)) = \dim(\text{col}(A))$
$=$ number of leading 1's in RREF

The Rank Theorem:

$\text{rank}(A) + \text{nullity}(A) = r + (n-r)$
$= n$

$\text{rank}(A) = \dim \text{col}(A)$ # of pivots

$\text{nullity}(A) = \dim \text{Nul}(A)$ # of free vars
# of columns without Pivots

$\text{rank}(A) + \text{nullity}(A) = n$

- $\text{rank}(A)$: # of Pivot columns
- $\text{nullity}(A)$: # of non Pivot columns
- n: # of columns Total

Fundamental Theorem of Invertible matrices

Let A be an nxn matrix

1a) A is invertible
1b) $Ax = b$ has a unique sol
1c) $Ax = 0$ has a trivial sol
d) RREF of $A = I_n$
e) rank $(A) = n$
f) nullity $(A) = 0$
g) Column vectors of A are lin indep, Span $\mathbb{R}^3$ and form a basis
h) row vectors of A are lin indep, Span $\mathbb{R}^3$ and form a basis

Find a basis for a subspace

1. Find Coordinates by Parametrize
2. RREF
3. Check lin independent

Injective (one to on)

T is injective only when the null space only contains the $\vec{0}$

If there is a free variables, this tells us the kernel has infitly many solutions.

for any $y \in Y \rightarrow$ at most 1 x such that $T(x) = y$

one x gives one y.

Rank(A) = n

Surjective (on to)

L.T $\mathbb{R}^4 \rightarrow \mathbb{R}^3$

$\dim(\text{Im } T) = 3 = \dim(\mathbb{R}^3)$

L.T is surjective when $\text{Im } T = \mathbb{R}^3$, or when it spans the codomain.

All y's have atleast 1 x.

These 2 events are completely independent.

Every element of y has some x.

Rank(A) = m

Eigenvalue - a scalar multiple of a matrix, represents stretch/compression

Eigen vector - a vector that corresponds to an eigenvalue. Must be Non zero

Eigenspace - set of all vectors x such that $(A-\lambda I)x=0$. (Parametrize and express relation as

Note

eigenvalue can be 0

eigenvalue can be complex #

eigenvectors can not be $\vec{0}$

Key Ideas

- If a matrix is invertible its det $\neq 0$.
- Swapping rows, swaps the sign of the determinant

- Any matrix with an all 0 row or column has a det $=0$.
- Any triangular matrix has a det equal to product of the diagonal.

Ex $\begin{bmatrix} 2 & 0 & \frac{1}{2} \\ 0 & 5 & 7 \\ 0 & 0 & 7 \end{bmatrix} = 2 \times 5 \times 7$

$= 70$

Properties

$\det(AB) = \det(A)\det(B)$

$\det(A^T) = \det(A)$

If B is A x K then

$\det(B) = KA$

Cramer's Rule

- A way to solve a system of linear equations

Example

Solve using Cramers Rule: $x+3y=5$, $2x+2y=6$

① Take coefficient matrix: $A = \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix}$

② Find determinant of the coefficient matrix:

$\det(A) = -4$

③ $A_1 = \begin{bmatrix} 5 & 3 \\ 6 & 2 \end{bmatrix}$ Find $\det(A_1)$

Notice how replaced by the augmented side

$= -8$

$A_2 = \begin{bmatrix} 1 & 5 \\ 2 & 6 \end{bmatrix}$ $\det(A_2) = -4$

④ $x = \frac{\det(A_1)}{\det(A)}$ $y = \frac{\det(A_2)}{\det(A)}$

$x = \frac{-8}{-4}$, $x = 2$ $y = \frac{-4}{-4}$, $y = 1$

∴ Cramer's Rule works no matter how many variables and how many unknowns. $x_i = \frac{|A_i|}{|A|}$

$|A| \neq 0$

Rank Nullity Theorem: rank + nullity $= n$

$\dim(\ker(T)) + \dim(\text{Im}(T)) = n$

$\ker(A^T) = \{\vec{0}\}$ implies A^T is injective $A^T\vec{x} = 0$, $x = 0$

Bijective implies invertibility, we can undo the transformation.

$\det(A^T) \neq 0$, implies A^T is invertible, which means only vector in $\ker(A^T)$ is the zero vector.

Process of finding basis of eigenspace

1) characteristic polynomial $(A - \lambda I)$

2) eigen values $\det(A - \lambda I) = 0$

3) Plug eigen values into $A - \lambda I$ matrix

4) RREF, parameterize to find eigenvectors / basis for the eigen space E_λ.

The Fundamental Theorem of Invertible Matrices: Version 3

Let A be an $n \times n$ matrix. The following statements are equivalent:

a. A is invertible.
b. $A\mathbf{x} = \mathbf{b}$ has a unique solution for every $\mathbf{b}$ in $\mathbb{R}^n$.
c. $A\mathbf{x} = \mathbf{0}$ has only the trivial solution.
d. The reduced row echelon form of A is I_n.
e. A is a product of elementary matrices.
f. $\text{rank}(A) = n$
g. $\text{nullity}(A) = 0$
h. The column vectors of A are linearly independent.
i. The column vectors of A span $\mathbb{R}^n$.
j. The column vectors of A form a basis for $\mathbb{R}^n$.
k. The row vectors of A are linearly independent.
l. The row vectors of A span $\mathbb{R}^n$.
m. The row vectors of A form a basis for $\mathbb{R}^n$.
n. $\det A \neq 0$
o. 0 is not an eigenvalue of A.

Recall:

Diagonal Matrix: $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

To Prove $A \sim B$ with a matrix P.

Show $AP = PB$.

Similiar Matrix:

$A \sim B$ if there is an invertible $n \times n$ matrix P such that $P^{-1}AP = B$

Properties

$A \sim A$

If $A \sim B$ then $B \sim A$

If $A \sim B$ and $B \sim C$ then $A \sim C$.

Diagonalizability

By definition: A matrix is diagonalizable if there exists a diagonal matrix D and an invertible matrix P such that $P^{-1}AP = D$

* Note: This is only possible if it has linearly indep eigenvectors and
* If the algebraic multiplicity of each eigenvalue of A is equal to its geometric.

Properties of Similiar Matrice

suppose $A \sim B$.

1) $\det(A) = \det(B)$
2) A is invertible if aof B is invertibl
3) A and B have same rank
4) Same eigen values

Orthogonal Matrices

Define: A is said to be orthogonal if

$$A^T \cdot A = I_n$$

Properties: ortho matrices preserve length and angles. (orthogonal unit vectors)

$$A^{-1} = A^T$$

Orthogonal matrices are often used in transformations: rotation, reflection.

$$\det(A) = \pm 1$$

$$|\pi| = 1$$

Orthogonal Vectors

- In order for a set of vectors to be orthogonal it means, the vectors must be non-zero and linearly indep.

- To check if vectors are orthogonal compute all dot products.

Let W be a subspace of finite-dimensional vector space V.

a) W is finite dimensional and $\dim W \leq \dim V$.

b) $\dim W = \dim V$ if and only if $W = V$.

Orthogonal Basis

set of vectors:

1. orthogonal
2. spans vector space

This implies lin indep.

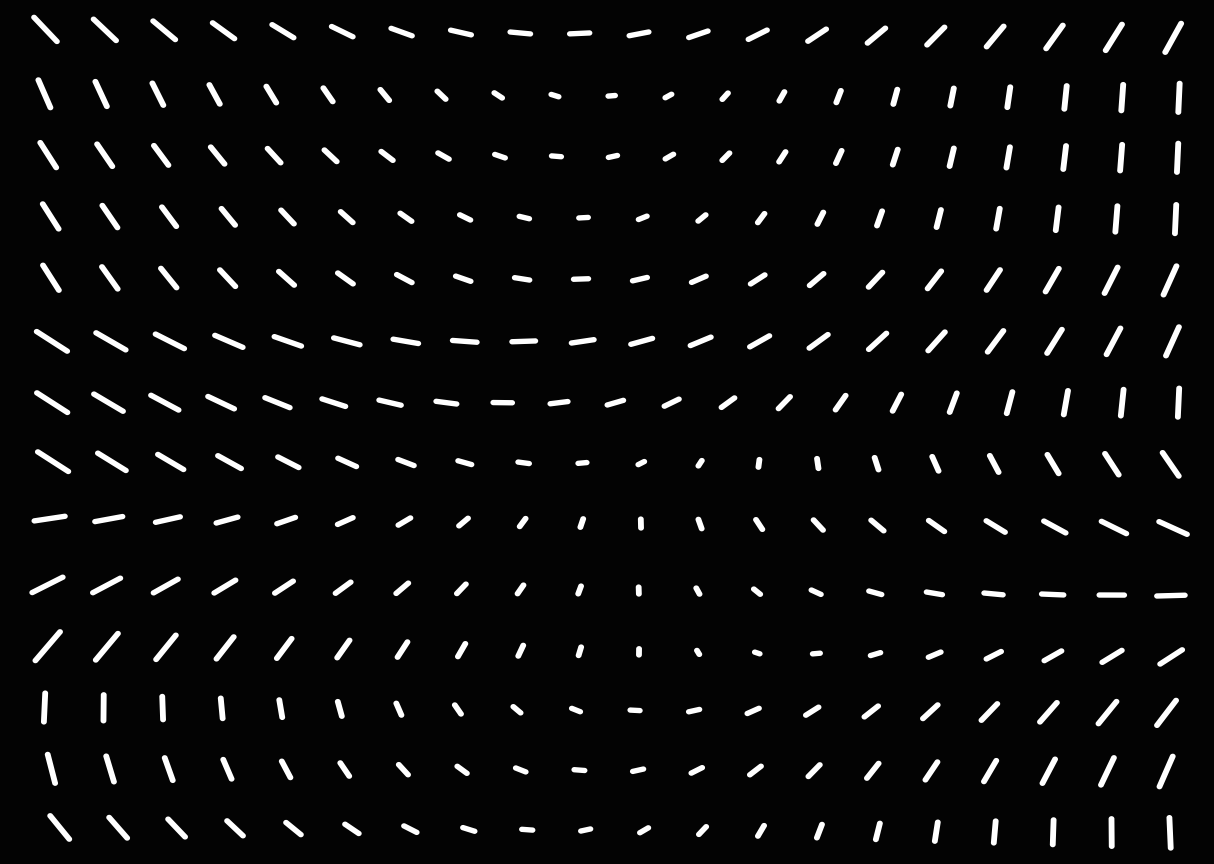
Usefull for projections

Orthonormal Basis

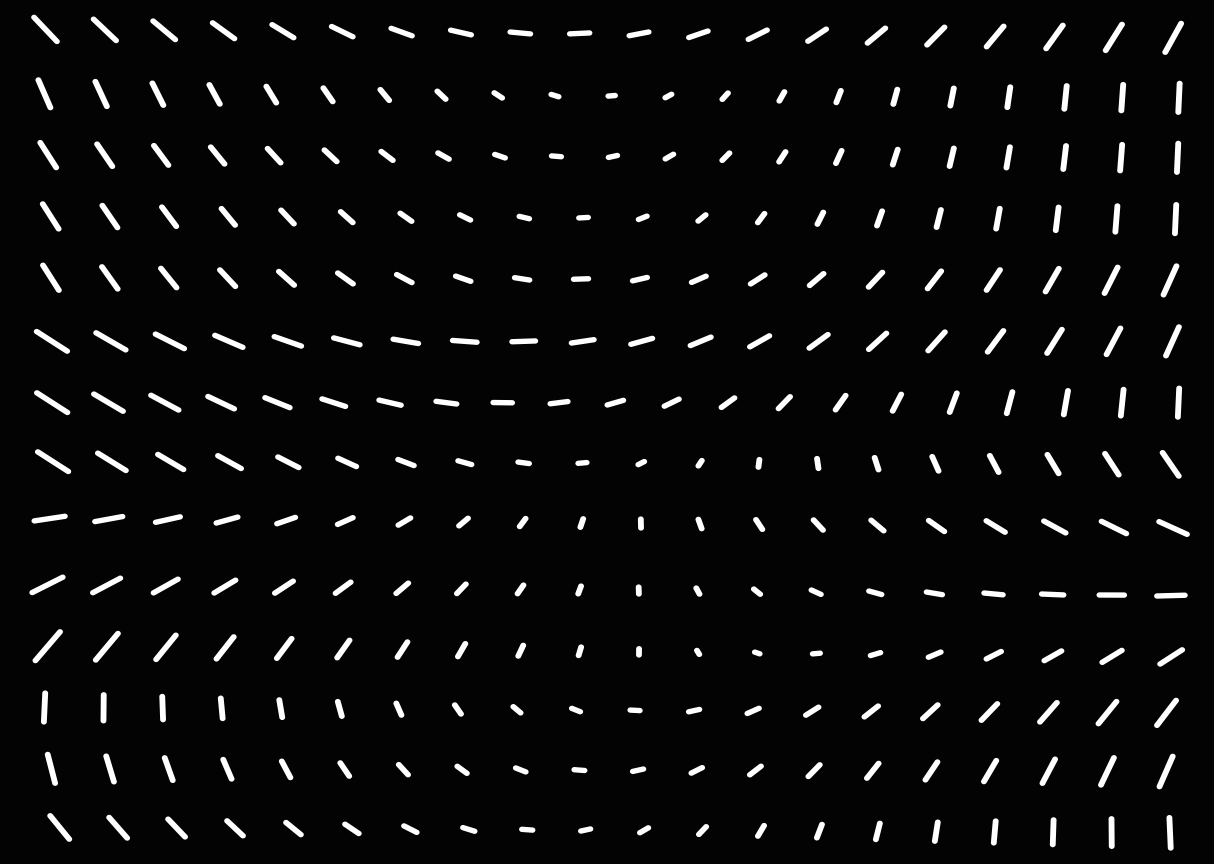
Find orthogonal vectors then divide by the magnitude. (in a root)

1. orthogonal
2. spans vector space
3. magnitude = 1

All Lecture notes



Unit 1



Lecture 1

Lecture 1 PDF Notes

1.) Matrices and Linear Transformations

- Matrix is set of ordered columns
- A linear transformation maps one vector to another through operation $y = Ax$, where
  - A is transformation
  - x is input vector
  - y is output vector

2.) Eigenvalues and Eigenvectors

- Eigenvector (v): A vector that only changes its magnitude when transformed by A

$Av = \lambda v$, λ : Eigenvalue, the scaling factor.

Properties:

- Eigenvectors keep direction
- Eigenvalues represent magnitude of scaling

3. Characteristic Polynomial and Equation

- To Find Eigenvalues
  1) use $(A - \lambda I)v = 0$
  2) $\det(A - \lambda I) = 0$, for non trivial solution

- Null Space is set of vectors v, such that $Av = 0$

- Span and Basis:
  - Span: All possible lin combo's of a set of vectors
  - Basis: Minimal set of vectors that span a space and that are linearly indep.

Eigen value / Eigen Vectors Review

$A\vec{x} = \lambda\vec{x}$

$= A\vec{v} - \lambda\vec{v} = \vec{0}$

$= A\vec{v} - \lambda I v = \vec{0}$

$= [A - \lambda I]\vec{v} = \vec{0}$

$A = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$

$\det(A - \lambda I) = 0$

$A - \lambda I = \begin{bmatrix} 4-\lambda & 3 \\ 3 & 4-\lambda \end{bmatrix}$

$(4-\lambda)^2 - 33 = 16 - 8\lambda + \lambda^2 - 9$

$= \lambda^2 - 8\lambda + 7 = 0$

$= (\lambda - 1)(\lambda - 7) = 0$

$\lambda = 1, 7$

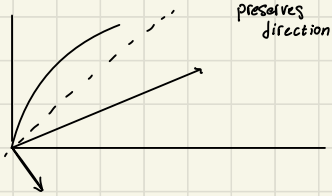
What is

$A\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$

$A\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$A\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \end{bmatrix}$

$A\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$



---

often

$\vec{a} = \frac{\vec{v}}{\|\vec{v}\|}$

$d_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ For $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\|\vec{v}\| = \sqrt{1^2+1^2} = \sqrt{2}$

$d_2 = \begin{bmatrix} 1/\sqrt{2} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ For $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ $\|\vec{v}\| = \sqrt{1^2+(-1)^2} = \sqrt{2}$

$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 5 \\ 0 & 0 & 6 \end{bmatrix}$ Triangular matrix

Eigenvalues are 1, 3, 6

$A = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$

$A\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$

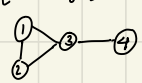
$A = \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$

$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

V = Vertices

E = edges



degree of vertex # is the number of vertices adjacent to #1

$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ $\vec{d} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 1 \end{bmatrix}$ Sum of each rows

Consider $A\vec{1} = A\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

A has ones $(\lambda, 1)$

Laplacian matrix is

$L = D - A$

$= \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$

Given: $\{(1\,3\,1)\ (1\,5\,1)\ (3\,5\,1)\ (6\,2\,1)\ (2\,4\,1)\}$

Find Laplacian:

$L = D - A$

$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

Node	
1	2
2	2
3	2
4	1
5	2
6	1

$D = \begin{bmatrix} 2 & & & & & \\ & 2 & & & & \\ & & 2 & & & \\ & & & 1 & & \\ & & & & 2 & \\ & & & & & 1 \end{bmatrix}$

$L = \begin{bmatrix} 2 & 0 & -1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$

Lecture 2

Lecture 2 - PDF Notes

1.) Graph G is defined by 2 sets:

- Vertex set V: Finite, non empty set of vertices
- Edge set E: Set of connections (edges) between vertices

EX

$G1$: $V = \{1,2,3,4,5\}$, $E = \{(1,2),(1,3),(2,3),(4,5)\}$

Adjacency Matrix

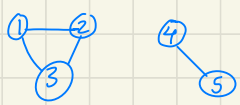
- The adjacency matrix A is a square matrix

where: $a_{ij} = \begin{cases} 1 & \text{if theres an edge between } v_i \text{ and } v_j \\ 0 & \text{otherwise} \end{cases}$

- Symmetric: $a_{ij} = a_{ji}$
- Diagonal entries all 0
- Real Eigenvalues

EX

G_1:



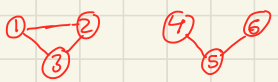
$A_1 =$

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

EX

Non-Bipartite Graph

G_3



$A_3 =$

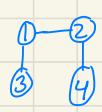
$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Degree Matrix:

- The Degree matrix D is a diagonal matrix

where each diagonal entry represents the degree (num of edges) of a vertex

EX



$D =$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Lecture 2 - PDF Notes

Laplacian Matrix

The laplacian matrix is defined by

$$L = D - A$$

- D : Degree Matrix
- A : Adjacency Matrix

Properties:

- Symmetric and Real: All eigenvalues are non negative
- Smallest eigen value - connectedness of the graph
- number of 0 eigenvalues equals number of graph components

Computing By Hand

1.) write adjacency matrix
2) Negate nonzero entries
3) replace diagonal aii with di

r Bipartite / Not Bipartite

Fiedler Vector

- The eigenvector corresponding to the second smallest eigenvalue of L is called Fiedler vector
- Cluster vertices based on positive or negative
- Helps split the data into two chuncks

Graphs: Adjacency Matrix and Laplacian Matrix

Main concepts

- Graph, vertices, edges are $G(V, E)$
- Adjacency matrix $A(G)$ is symmetric and binary

Graphs G

- Made up of vertices (V): A non-empty, finite set of points
- Edges (E): A set of connections between vertices with no direction or repitition

Ex

$G_1 = \{\{1,2,3,4,5\}, \{(12),(13),(23),(45)\}\}$

Definitions:

- Edges are incident when they share a vertex
- Vertices are adjacent when linked by edge
- Subgraph: is a smaller part of bigger graph which contains less edges/less nodes.
- Bipartite: graph that can be split up into two groups so that, every edge connects a vertex from one group to a vertex from the other group.
  And, there are no edges between vertices in same group
- Path: sequence of edges that goes from one vertex to another without repeating any edge/vertex.

Adjacency Matrix

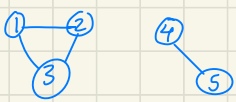
- The adjacency matrix A is a square matrix

where: $a_{ij} = \begin{cases} 1 & \text{if there's an edge between } v_i \text{ and } v_j \\ 0 & \text{otherwise} \end{cases}$

- Symmetric: $a_{ij} = a_{ji}$
- Diagonal entries all 0
- Real Eigenvalues

EX

G_1:



$A_1 =$

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

EX

Non-Bipartite Graph

G_3

$A_3 =$

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Main Concepts

- Degree matrix $D(G)$
- Laplacian matrix $L(G)$
- Fiedler vector provides binary clustering

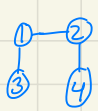
Degree Matrix

- Special matrix, tells us how many connection each node in a graph has

Properties:

- all non diagonal entries are 0
- The value in position $D[i][i]$ is the degree of node i

EX



$D =$

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Laplacian Matrix

The laplacian matrix is defined by

$$L = D - A$$

- D: Degree matrix
- A: Adjacency matrix

Properties:

- Symmetric and Real: All eigenvalues are non negative
- Smallest eigen value - connectedness of the graph
- number of 0 eigenvalues equals number of graph components

Computing By Hand

1.) write adjacency matrix

2) Negate non-zero entries

3) replace diagonal a_{ii} with d_i

For Bipartite/Not Bipartite

Laplacian Matrix Properties

- Symmetric, real
- Diagonally dominant (each diagonal entry is greater than or equal to sum of absolute values, off diagonal entries)
- Any dd is positive semi-definate
- $L\vec{i} = [D - A]\vec{i}_i = \vec{d} - \vec{d} = \vec{0}$
- $\lambda \geq 0$, at least one $\lambda = 0$

- Note: The number of 0 eigenvalues is the number of components in the graph

Fiedler Vector

- The eigenvector corresponding to the second smallest eigenvalue of L is called the Fiedler vector

- Cluster vertices based on positive or negative

Lecture 3

Lecture 3 PDF - Notes

Vector Spaces:

- Set of vectors that satisfies 8 axioms
  - Associativity, commutativity, indentity....
  - Distributive and associative properties for scalar multiplication

Size vs Dimension

- Size: Number of entries in a vector ($v \in \mathbb{R}^m$)
- Dimension: # of indep vectors needed to span space

Ex $v = \begin{bmatrix} x \\ 2x \end{bmatrix}$ is 1D vector in $\mathbb{R}^2$

Matrices and Vector Spaces

- Linear Transformation: A matrix A represents linear transformation $A: \mathbb{R}^n \to \mathbb{R}^m$, mapping vector in $\mathbb{R}^n$ to $\mathbb{R}^m$

- Coloumn Space: Set of all vectors that can be expressed as a lin combo of matrix cols

- Null Space: All vectors x, s.t $Ax = \vec{0}$
  To find null space, RREF, then parametrize

Subspaces and Linear Span

- Subspace: Subset of vector space
  - closed under addition and scalar multiplication

- Span:
  Set of all lin combo of a set of vectors

Ex $v1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ span $\mathbb{R}^2$.

Partition:

- A partition means dividing a set/matrix into smaller non overlapping parts
- Suppose S is a set, subsets are $P_1, P_2 \ldots$
  - each subset is non empty
  - union gives original set
  - no overlap

Block Partitioning

- Dividing matrices into submatrices
- coloumn partitioning of a matrix:
- A matrix $A \in \mathbb{R}^{m \times n}$ can be seen as

$$A = [a_1, a_2 \ldots a_n]$$

where each $a_j \in \mathbb{R}^m$ is coloumn vector

- Relation to linear combinations

$A = [a_1, a_2 \ldots a_n]$, lin combo: $Aw = C$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$ coloumn vector of weights

Note: $\vec{v} \cdot \vec{v} = \vec{v}^T \vec{v} = \vec{v}^T \vec{v}$

ussually have ecclidean norm

$$\|\vec{v}\| = \sqrt{\vec{v} \cdot \vec{v}} = \sqrt{\vec{v}^T \vec{v}}$$

$$= \sqrt{\sum_{j=1}^{n} v_j^2}$$

Comes from $\vec{v} \cdot \vec{v} = \vec{v} \cdot \vec{v}$

Subspace

- A subspace is a linear subset of a vector space V is a suset: $W \subseteq V$.
- Closed under addition and scalar multiplication

$$v, v \in W \Rightarrow av + bv \in W$$

- Must include $\vec{0}$
- The dimension is the min number of vectors needed to construct all other vectors in the space (basis)

$$\vec{v} = \begin{bmatrix} x \\ 2x \end{bmatrix}$$

1 basis vector $\rightarrow$ dim = 1

NULL SPACE

- The null space is the set of all vector x that make $A\vec{x} = \vec{0}$
- To Find Null space, reduce matrix to RREF
  - Pivots define constraints
  - Free variables determine null space basis

If A reduces to: $\begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$x_1 - \frac{1}{2}x_2$

$x_3 = t$

$$\begin{bmatrix} 1 \\ -\frac{1}{2} \\ 0 \end{bmatrix} + t\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Trivial Null Space ($\{0\}$):

- Happens if $Aw = b$ has one solution
- Non trivial null space:

  occurs when many free variables exist

Lecture 4

Lecture 4 - PDF Notes

Spanning Sets

- The span of a set of vectors is the set of all possible linear combinations of those vectors

$$\text{span}(\{v_1, v_2, \ldots v_n\}) = \{w_1v_1 + w_2v_2 + \ldots + w_nv_n \mid w_i \in \mathbb{R}\}.$$

- Describes $\mathbb{R}^m$
- Spanning set may contain lin dep vectors

Basis

- A basis is a minimal set of vectors that spans a subspace.
- Basis vectors are linearly independent

Properties:

- Basis for an n dimensional subspace has n vectors
- Each vector in subspace can be made from lin combo of basis vectors

Rank Nullity Theorem

- For a matrix $A \in \mathbb{R}^{m \times n}$, rank nullity theorem

$$\text{rank}(A) + \dim(\text{null}(A)) = n$$

Orthogonal Subspaces

- Orthogonal subspaces, two subspaces U and V are orthogonal, if every vector in U, is orthogonal to every vector in V:

$$U^T V = 0, \quad \forall u \in U, \ v \in V$$

Orthogonal Compliment

- The O.C of a subspace U, denoted as U^T, is the set of all vectors in $\mathbb{R}^n$, that is orthogonal to every other vector

Orthonormal Basis

Orthogonal Basis: $(i \neq j) \rightarrow (\vec{v_i} \cdot \vec{v_j} = 0)$

Orthonormal Basis:

- All basis vectors are mutually orthogonal
- Each basis vector has unit length 1.

- The inverse of an orthogonal Matrix is it's Transpose:

$$Q^T Q = I \Longrightarrow Q^{-1} = Q^T$$

Ex

$$u_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad u_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

- $u_1 \cdot u_2 = 0$
  $u_2 \cdot u_3 = 0$
  $u_1 \cdot u_3 = 0$

$\therefore$ The set of vectors is an orthogonal basis for $\mathbb{R}^3$.

Lecture 5

Lecture 5 PDF Notes

Diagonalizable Matrices

- A matrix A is diagonalizable, if it is a similar to a diagonal matrix:

$A = PDP^{-1}$ $\longrightarrow$ $P^{-1}DP = P^{-1}PA = A$

- P is an invertible matrix, whos columns are eigenvectors of A
- D is a diagonal matrix whos diagonal entries are eigenvalues of A.
- P^{-1} is the inverse of P.

Notes:

- A matrix is diagonalizable if its eigenvectors form a basis (lin indep)
- If all eigenvalues are distinct its guaranteed diagonalizable

Eigenvector Basis

- Any vector v can be expressed as a lin combo of eigenvectors: $v = \alpha_1 v_1 + \alpha_2 v_2 + \dots \alpha_n v_n$
- Applying A to v: $Av = \lambda_1 \alpha_1 v_1 + \lambda_2 \alpha_2 v_2 + \dots$
- Only exists if A is diagonalizable

Orthogonal Matrices

A matrix is orthogonal if:

$Q^T Q = I$ and $Q Q^T = I$

- rows/columns of Q form an orthonormal basis

Properties:

- Preserves vector lengths and angle transformations

$Q^{-1} = Q^T$

Symmetric Matrices

- A matrix is Symmetric if $B^T = B$
- All real eigenvalues, eigenvectors are orthogonal

Ex

$B = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$

- Find eigen vector/values
- Verify orthogonality of eigenvectors

Skew-Symmetric Matrices

- A matrix is skew symmetric if $S^T = -S$

Properties:

- Diagonal entries of S are 0
- Eigenvalues are imaginery or 0.

Non diagonalizable

- A matrix is non diagonalizible if it has fewer linearly indep eigenvectors than its size.

Matrix Powers

- For a diagonalizable matrix A:

$$A^K = PD^KP^{-1}$$

- Note: Small changes to a nearly singular matrix causes large changes in eigenvalues/vectors

$$A = PDP^{-1}$$

$$A^2 = A \cdot A = (PDP^{-1})(PDP^{-1})$$

$$A^2 = PD^2P^{-1}$$

$$A^K = PD^KP^{-1}$$

- Square root matrix:

$$A^{\frac{1}{2}} \cdot A^{\frac{1}{2}} = A$$

$$A^{\frac{1}{2}} = EA^{\frac{1}{2}}E^{-1}$$

$$A^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2} \ldots . \sqrt{\lambda_n})$$

Taking square root of each eigenvalue

Equivalent Matrices

$$P_1 A P_2 = C$$

Similiar Matrices

$$PAP^{-1} = C$$

Lecture 6

Lecture 6 PDF Notes

Spectral Theorem

Only applies to square, diagonalizable, symmetric matrices

- For a symmetric matrix $B \in \mathbb{R}^{n \times n}$;

$$B = Q \Lambda Q^T$$

- where Q is an orthogonal matrix with eigenvectors as columns ($Q^T Q = I$)
- Λ: Diagonal matrix with eigenvalues of B
- Symmetric matrices always have real eigenvalues and orthogonal eigenvectors

Q: matrix of orthonormal eigenvectors
Λ: diagonal matrix of eigenvalues
Q^T: transpose of Q, or inverse of Q

$$Q^T = Q^{-1}$$

$$B = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \qquad B = Q \Lambda Q^T$$

Positive Definite and Semi-Definite Matrices

- Positive Definite ($B \succ 0$):
  - All eigenvalues $\lambda_i > 0$
  - Quadratic form $U^T B U > 0$ for any non zero vector U.
- Positive Semi-Definite ($B \succeq 0$):
  - All eigenvalues $\lambda_i \geq 0$
  - Quadratic form $U^T B U \geq 0$
- Eigenvalues give you the important info

Quadratics Forms

- A form is $U^T B U$, where:
  - U is a vector
  - B is a symmetric matrix
- Analyze energy, variance and optimization problems
- If $U^T B U \geq 0$, B is positive definite
- The quadratic form $U^T B U$ is a practical way to check if B is pos/semi def

Covariance Matrix in Statistics

- A covariance matrix C measures the variance and relationships between data vectors $x_1, x_2 \ldots$.

$$C = \frac{1}{m-1} M^T M \rightarrow (\vec{d_1} \cdot \vec{d_2})$$

$$\text{Cov}(x_1, x_2) = \frac{1}{m-1} d_1 \cdot d_2$$

- where M contains zero-mean vectors as columns.
- C is symmetric and positive semi definite.
- m is number of data points
- d_1 and d_2 are deviations

Learn how to find means and how to find zero mean vectors:

Applications in Physical Systems

Spring Systems (Hookes Law)

- Models: Displacement vector x results in forces F governed by the Stiffness matrix K:

$$F = Kx$$

- Energy: $E = x^T K x$

K must be positive definite to ensure $E > 0$

Difference vector

$$\vec{d} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \end{bmatrix} = \vec{x} - \vec{x\bar{t}}$$

a has degrees of freedom

Variance: how data vary from mean

"Mean Squared Difference"

Rank(A) is the number of lin indep columns. (# of Basis vectors)

Nullity(A) is the set of $A\vec{x} = 0$

$\text{Rank}(A) = \text{Rank}(A^T)$

Rank Nullity Theorem

$\text{rank}(A) + \text{nullity}(A) = n$ → # of columns

rank(A): linearly indep columns

nullity(A): dim of null space

Note:

$CC^T = C^TC$

This is normal matrix

- $C^T = C^{-1}$ → orthogonal
- $C^T = C$ → symmetrical
- $C^T = -C$ → Skew Symmetrical

Consider n by m matrix

A matrix C is similar such that → similar matrix $A \sim B$ an invertible

$C = PAP^{-1}$

Most Important: Diagonal matrix

$D = PAP^{-1}$

then matrix D is diagonalizable

Recall $A\vec{v} = \lambda\vec{v}$

These are a basis, all linearly indep

Consider $B \in R^{n \times m}$ and $R = R^T$

$B = V\lambda V^{-1}$

$B^T = \left[V\lambda V^{-1}\right]^T$ $V = V^T$, V is orthogonal

$= [V]^T \cdot A^T \cdot V^T$

=

Spectral Decomposition

Eigen value $A \in R^{n \times n}$

$AV = VA$

V is invertible, because it has all linearly indep eigenvectors.

$= AI - VVA$

$= A - VAV^{-1}$

Consider $A \in \mathbb{R}^{n \times n}$

Consider $A^T = AA$

$= [V\lambda I\lambda V^{-1}$

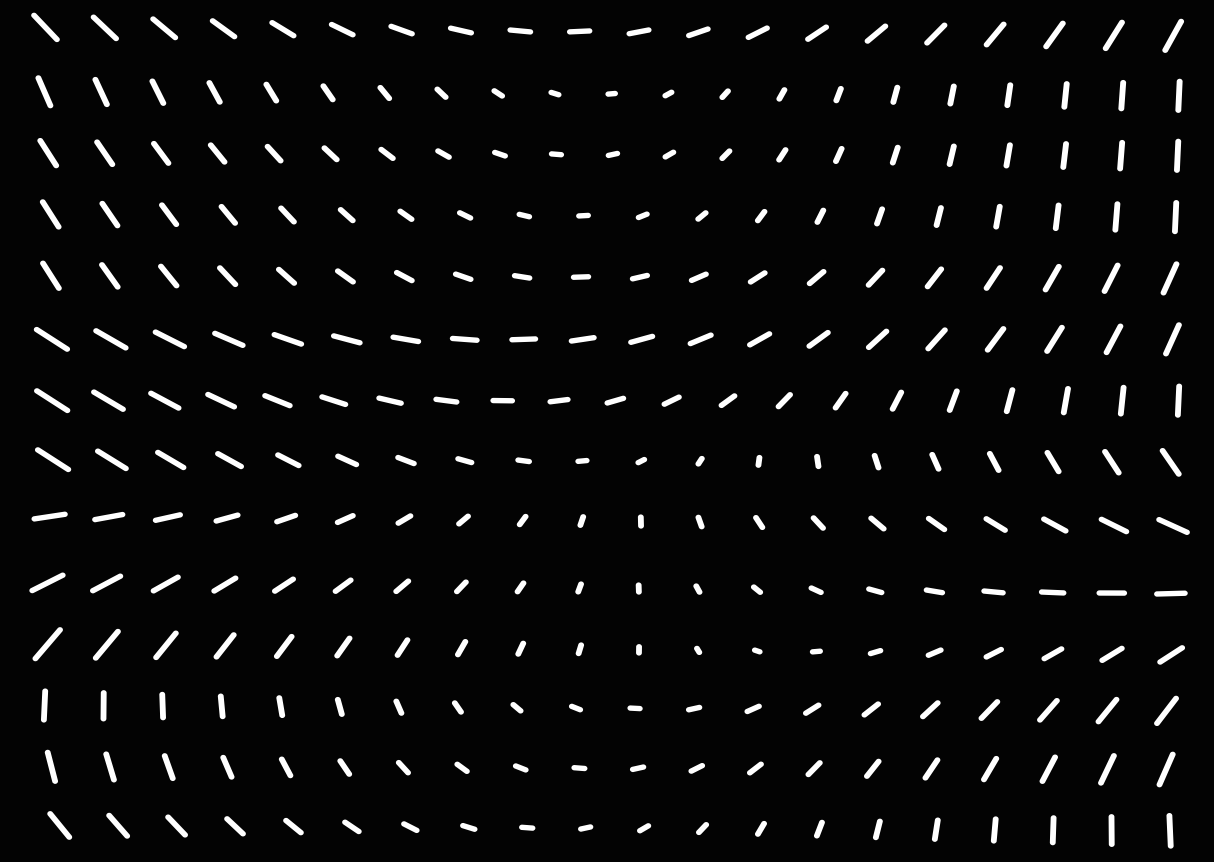
$= V\lambda^2 V^{-1}$

positive semi definite matrix all $\lambda \geq 0$.

Find sqrt of

$A = V \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \sqrt{\lambda_2} & \\ & & \sqrt{\lambda_3} \end{bmatrix}$

Unit 2



Lecture 7

Lecture 7 - PDF Notes

Design Matrix

- A matrix gathers multiple data vectors for analysis
- Independent variables → Columns
- Observations (data points) → Rows
- Example:
  Age, height, weight - indep var's
  A row represents individual's data

Note: Observations must be converted into numerical representation
Weighted Sum: $A \cdot w$

Matlab: A_standardized = (A - mean(A)) ./ std(A))

Zero Mean Data

- Subtract mean from each column to center the data around 0
- Relates observations to dependent variables

$$A\vec{w} \approx \vec{c}$$

1) zero mean
2) unit variance

- A is input data
- w is weight vector
- c is target output

Standardization

- Measures how far data point is from mean in terms of standard deviation

$$z = \frac{x - \mu}{\sigma}$$

- Data has mean = 0, variance = 1

Example

$a = [15, 17, 31, 19, 3]$

Mean: $\bar{a} = 17$

Subtract: $[-2, 0, 14, 2, -14]$

Variance $= \sigma^2 = 100$, $\sigma = 10$

divide by σ to get Standardized $[-2, 0, 1.4, 0.2, -1.4]$

Example

Standardize: $A = \begin{bmatrix} 5 & -3 \\ -1 & 0 \\ 3 & 2 \\ 7 & 6 \\ 6 & 5 \\ 4 & 2 \end{bmatrix}$

$z = \frac{x - \mu}{\sigma}$

1) zero mean: $\begin{bmatrix} 1 & -5 \\ -5 & -2 \\ -1 & 0 \\ 3 & 4 \\ 2 & 3 \end{bmatrix}$

$\sigma_1 = \sqrt{\frac{(1^2+5^2+1^2+3^2+2^2)}{4}}$

$\sigma_1 = \sqrt{10}$

$\sigma_2 = \sqrt{\frac{(5^2+2^2+0^2+4^2+3^2)}{4}}$

$= \sqrt{13.5}$

2) Find Std

3) Compute $z = \begin{bmatrix} 0.316 \\ -1.5811 \\ \end{bmatrix}$

Linear Regression & Standardization

Linear Model Representation

- Matrix form of linear regression

$$Aw \approx c$$

Residual $= Aw - c$

$\|Aw - c\|^2$

- A = independent variable matrix
- c = dependent variable vector
- w = unknown weight vector

Standardizing Design Matrix

- To convert to Design matrix from A,

Formula: $$X = \frac{A - \text{mean}(A)}{\text{std}(A)}$$

- The dependent variable c is standardized.

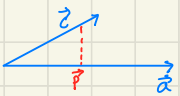
Lecture 8

Lecture 8 - Orthogonal Projections - PDF Notes

- Orthogonal projection is used to find closest vector in a subspace to a given vector.
- Useful in least squares regression, PCA

Projection Using Normal Vector

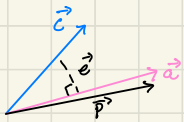
- The goal is to find the closest vector in a subspace to a given vector.



- Scalar weight for a single vector
  - Projecting onto 1D subspace
  - Given a vector C, find the closest vector P, that is a multiple of another vector a.

$$p = \frac{a^T c}{a^T a} \cdot a$$

  - This ensures C-p is perpendicular to a.
- Concept: use the error vector: $\vec{e} = \vec{c} - \vec{p}$
  Minimize the error



some w, makes $\vec{e}$ perp or minimizes error.

$\vec{a} \perp \vec{e}$

$\vec{a} \cdot \vec{e} = 0$

$a^T e = 0$

$a^T(\vec{c} - \vec{p}) = 0$

$a^T\vec{c} - a^T\vec{p} = 0$

$a^T p = a^T c$

$\vec{a}w\vec{a} = \vec{a}^T\vec{c}$

$= w(a^T \cdot a) = \vec{a}^T\vec{c}$

$w = \frac{\vec{a}^T \cdot \vec{c}}{a^T a}$

$p = \vec{w}\vec{a}$

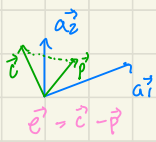
$= \frac{\vec{a}^T\vec{c}}{a^T a}\vec{a}$

Weight Vector For a Subspace (Higher Dim)

- Given multiple basis vectors, we find a projection within their span

Projection Formula: $p = Aw$

- where A is a matrix with basis vectors as columns
- w is weight vector that determines lin combo



$e \cdot a_1 = 0$

$e \cdot a_2 = 0$

$a_1^T\vec{e} = 0$

$a_2^T\vec{e} = 0$

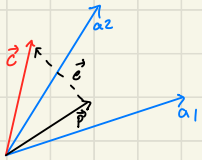
$\begin{bmatrix} a_1^T \\ a_2^T \end{bmatrix}\vec{e} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$A^T e = \vec{0}$

$\vec{e}$ is in null space of A^T

Vector Projection

- Given matrix A with indep vectors
- The column space of A is W
- The projection of $\vec{c}$ into W is $\vec{p}$
- Error vector is $\vec{e} = \vec{c} - \vec{p}$
- Weights $\vec{w}$ produce $\vec{p} = A\vec{w}$ → actual Projection Vector



$$p = \frac{a^T \cdot c}{a^T \cdot a} \cdot a$$

- Normal Equation:

$$A^T A \vec{w} = A^T \vec{c}$$

$$\vec{w} = \frac{A^T \vec{c}}{A^T A}$$

- Projection matrix: $P = A[A^T A]^{-1} A^T$
  - ↳ Matrix directly projects any vector $\vec{c}$ onto column space of A

- error vector $\vec{e}$ is always perpendicular to the column space

$$A = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \qquad c = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$$

Find error of projection

$\vec{e} = c - p$

$$p = \frac{A^T c}{A^T A} \cdot A$$

$$\vec{e} = \frac{\begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}}{\begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}}$$

$$= \frac{\begin{bmatrix} -28 \\ 14 \end{bmatrix}}{= \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}}$$

weight $= \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$\begin{bmatrix} -28 \\ 14 \end{bmatrix} \div \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}$

2 by 1 2 by 2

$$c = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix} - \begin{bmatrix} -4 \\ 2 \end{bmatrix} \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}$$

Normal Equation for Projection

- To compute weights w for projection:

$$A^T A w = A^T c$$

  - This is derived by ensuring the error vector is perpendicular to all basis vectors

Projection Matrix

- Projection matrix P transforms any vector c into its projection

$$P = A(A^T A)^{-1} A^T$$

  - Useful for least squares regression / reducing dimensionality

$$\vec{p} = A\vec{w}$$
$$= A\left[A^T A\right]^{-1} A^T \vec{c}$$
$$= P\vec{c}$$

Note: $A\vec{w} \approx \vec{c}$

only possible if c is in the column space of A.

$$\vec{e} = c - A\vec{w}$$

Lecture 8 - In class Note

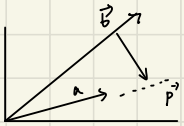
Independend data:

Trying to approximate $c_i = a_i$

To find scalar w, such that c_i is approx wa_i

Consider $Al = Cl$,

$C \approx w_i$. Cant divide vector by vector



Exact multiple

$p = \vec{w} \cdot \vec{a}$

Orthogonal $\vec{c} \perp \vec{a}$ $= a \mid c = w\bar{a} \mid = 0$

$a \cdot e = 0$

$a^T e = 0$ $\quad a^T c = w \cdot a^T a$

if a is not 0 vector then $\vec{a} \neq 0$

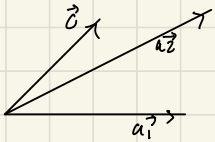
$$w = \frac{a^T \vec{o}}{a^T a}$$

$$w = \frac{[1\ 1]\begin{bmatrix} 2 \\ 3 \end{bmatrix}}{[1\ 1]\begin{bmatrix} 1 \\ 1 \end{bmatrix}} = \frac{5}{2}$$

Consider lin indep a_1, a_2 and a dependene $\vec{c}$.

Ex $a_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $a_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$ $\quad \vec{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Spans xy plane



$$\begin{bmatrix} \vec{a_1}, a_2 \end{bmatrix}\begin{bmatrix} \vec{w_1} \\ \vec{w_2} \end{bmatrix} = p$$

Error is perpendicular to projection vector

$\vec{e} \cdot \vec{p} = 0$

Error vector has to be orthogonal to basis

$a_1^T e = 0$

$$\begin{bmatrix} a_1^T \\ a_2^T \end{bmatrix} = \vec{a_1}\ \vec{a_2}$$

$= A^T e^{-1}$

$= A^T [C - P]$

$= A^T [\vec{C} - A\vec{w}] = \vec{0}$

$= A^T \vec{C} - A[A\vec{w}]\vec{0}$

$= A^T \vec{C} = [A^T A]\vec{w}$

$w = A \backslash C$

If A is full rank, A is positive definite, A is A^T invertible.

Solved general linear regression problem.

Lecture 10

Lecture 10: Linear Regression and Patterns

Key Definitions:

- independent variable (a): value that can be controlled or changed
- dependent variable (c): measured based on independent variable
- Residual error (e): the difference between actual value and predicted value from model
- Linear regression: model the relationship between variables by minimizing residual error
- Least squares: technique to find the best fitting line by minimizing sum of squared residuals. $\sum (e_i(\vec{w}))^2 = \|\vec{e}(\vec{w})\|^2$ want smallest norm

Residual Error and Regression Problems

- Define model function as: $c_i \approx F(w; a_i)$, w is weight param, a is indep var, c is depend var
- Residual error is: $e_i(w) = c_i - F(w; a_i)$, error depends on $\vec{w}$.
  - c_i: model; $F(w; a_i)$: actual

Linear Regression Model

- General form: $F(w; a_i) = a_{i1}w_1 + a_{i2}w_2 + \ldots + a_{in}w_n = a_i^T w$
- Error minimization leads to normal equation: $(A^TA)w = A^TC$ → normal eq
  - Linear regression is an orthogonal projection onto a vector space.
  - Solve as projection of $\vec{C}$ to V spanned by columns of A.

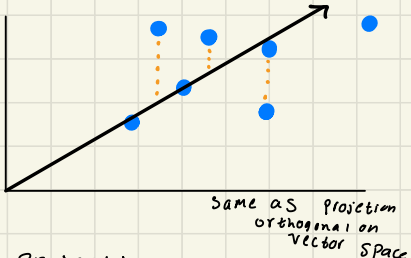
Hookes Law

- Simple mechanical spring follows Hooke's law:

$$C = wa$$

- Normal Equation solution:

$$w = \frac{a^T C}{a^T a}$$



- Shows least squares optimization is same as projection.

Linear Regression with Intercept

- If data follows linear trend but does not pass through the origin, we introduce intercept term: $c_i \approx w_1 a_i + w_2$

Matrix Equation: $A \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = C$

Normal Equation Still: $(A^T A) w = A^T C$

Standardizing

$$\tilde{x} = \frac{m}{\sigma_m}, \quad \tilde{y} = \frac{b}{\sigma_b}$$

Scaling Factor:

relationship between weights in original and standardized

$$U = \left(\frac{\sigma_b}{\sigma_m}\right)^2$$

affects whole model

Assessment of Linear Regression

- $RMS(A, \vec{c}, \vec{w}) = \sqrt{\frac{1}{m}\left(e_1^2 + e_2^2 + \dots + e_m^2\right)}$

root; mSE; Squared error

$$= \frac{\|\vec{e}(\vec{w})\|}{\sqrt{m}}$$

Lecture 11

Lecture 11 - Cross validating Linear Regression

Key Topics

- Cross-validation - A method for assessing a linear regression model - on unseen DATA
- Leave One out: simplest cross validation
- K Fold cross validation
- Training vs Testing
- Measuring Error

Data representation (matrix form)

$$A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix}$$

m observations
n features
each row is one data point

linear regression: $Av \approx c$

Before solving for v

$A \rightarrow X$ inputs

$c \rightarrow \tilde{y}$ outputs

$$X = \frac{A - \mu}{\sigma}$$

$$\tilde{y} = \frac{c - \mu}{\sigma}$$

Final standardized regression: $X w \approx \tilde{y}$

Now solve for w

Training vs Testing

- Training: Finding parameters (w) of regression model
- Testing: Evaluating model on unseen data.

Types of Cross-Validation

a) Leave one out C.V

- Each data point gets left out once, model gets trained on remaining data
- Helps detect outliers

b) Leave many out C.V

- Leave many data points out, then test

K Folds cross validation

- Data is split into K subsets
- Train on K-1 subsets and test on remaining
- Repeat K times, average results

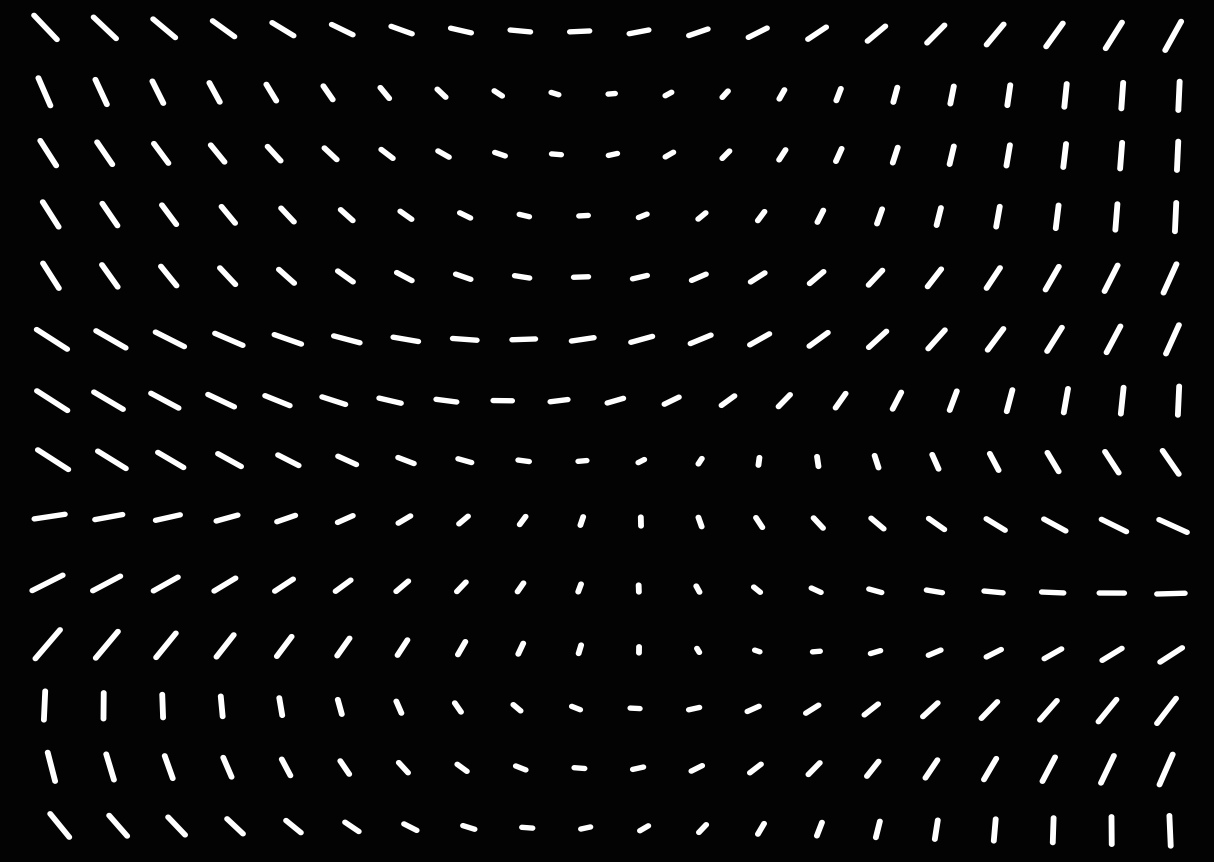
If 10 data points, K=5 split into 5 groups then train on 4 then test on 1.

$$RMS = \sqrt{\frac{(Xw-y)^T(Xw-y)}{m}}$$

Xw is predicted output
y is actual output
m is number of data points

If error, on new test data is great model is not a great fit.

Unit 3



Lecture 13

Lecture 13: Singular value Decomp (SVD)

- Generalizes eigenvalue decomp for non square matrices.

Key concepts

- Eigenvectors of AA^T → left Singular vectors → symmetric, positive semi definite
- Eigenvectors of A^TA → right singular vectors
- Non zero eigenvalues of AA^T and A^TA are the same
- Allows us to analyze eigenvalues of a non-square matrix

$A^TA = V\Lambda V^T$

Eigenvectors form an orthonormal basis.

A is full rank

- "Tall thin" $(m > n)$ → A^TA is square
- "Short wide" $(n > m)$ → AA^T is square

Left Transpose Product

- Eigenvectors of A^TA, if A is a tall thin then A^TA is symmetric square matrix and has orthonormal diagnalization: $A^TA = V\Lambda V^T$

V is eigenvectors, Λ is diagonal matrix of eigenvalues

Right Transpose Product

- Similiar to left, with short wide, AA^T can be diagonalized: $AA^T = U\Lambda U^T$
- The non zero eigenvalues of AA^T and A^TA are the same

Ex

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 1 & 1 \end{bmatrix}$$

$B_U = AA^T$

$$B_U = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

Eigenvalues 4, 2, 0

$\lambda_1 = 4$ eigen vector $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$

$\lambda_2 = 2$ $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

$\lambda_3 = 0$ $\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}$

- eigenvectors of AA^T form orthonormal basis

Singular Value Decomposition (SVD)

- For any real matrix $A \in \mathbb{R}^{m \times n}$

$$A = U \Sigma V^T$$

U is same as $B_U = AA^T$

V is same as $B_V = A^TA$

- where U is $m \times m$ orthogonal matrix (left singular vectors)
- V is $n \times n$ orthogonal matrix (right singular vectors)
- Σ is an $m \times n$ diagonal matrix with singular values $\sigma_1, \sigma_2 \ldots \sigma_r$

Properties of Σ:

- every σ_j is a real number
- for rank r, if $j \leq r$, then $\sigma_j > 0$
- values are ordered $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$

Relationship between SVD and Eigenvalues

$$AA^T = U \Sigma \Sigma^T U^T$$

$$A^TA = V \Sigma^T \Sigma V^T$$

- The singular values of A are the square roots of eigenvalues of AA^T and A^TA.
- Singular values tell us how a matrix stretches / transforms space

Lecture 14

Lecture 14: orthonormal Basis Vectors / SVD

Main Concepts

1. Left Singular vectors: Form orthonormal basis for Colum space (data vectors)
2. right Singular vectors: Form orthonormal basis fo row space (weight vectors)
3. Singular values: Positive real numbers, act as generalized eigenvalues of matrix

* How to find best basis

1. SVD of a Square Matrix

- For a sauare $m \times m$ matrix, the SVD decomp is:

$$A = U \Sigma V^T$$

- U: Orthogonal matrix whose columns are a basis of a data space
- Σ: Diagonal matrix with Singular values
- V: Orthogonal matrix whose columns are a basis for weight space

Ex 1

$$A = \begin{bmatrix} 3 & 4 \\ 0 & 3 \end{bmatrix} \qquad U_1 = \begin{bmatrix} 0.88 & -0.47 \\ 0.47 & 0.88 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 5.66 & 0 \\ 0 & 1.61 \end{bmatrix}, \quad V_1 = \begin{bmatrix} 0.47 & -0.88 \\ 0.88 & 0.47 \end{bmatrix}$$

- Eigenvals on diagonal
- Eigenvectors are not indep
- SVD still composis for non diagonalizable matrices

Ex 2

$$A_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

- Symmetric and rank defecient
- SVD only has one singular value
- Singular vector represents Colourmn space
- Second Singular vector is orthogonal complement

Non Sauare Matrix

For full rank non Sauare

$$A_3 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 0 \end{bmatrix}$$

- SVD gives best basis for vector space

If it is like

$$A_4 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & -2 \end{bmatrix}$$

lin combo of each other

second col represents null space for V_4

SVD for Approximate Basis Selection

- If σ_r (smallest singular value) is very small compared to σ_i it can be neglected
- approx the matrix, lower rank version, Keep most important singular values
- These are key for PCA

Important SVD Properties

- Decomposition: $A = U\Sigma V^T$ with orthogonal U, V and diagonal Σ.
- Rank, number of nonzero singular value in rank of A
- Basis Selection: 1) first r columns of U are basis for column space
  2) First r columns of V are basis for row space
  3) The last $(n-r)$ columns of V form a basis for row space
- SVD is a generalization of eigenvalue decomp
- Helps basis vector, PCA

Lecture 16

Lecture 16: PCA

What is PCA?

- Dimensionality reduction
- Used to find main patterns of variation in data
- new set of orthogonal axes that capture most of variance
- First principal component captures most variance
- Transforms data to a lower dimensional space

Zero-Mean Data Matrix

- Before PCA, Standardize data

1) Compute mean of each test
2) Subtract mean from each score

$$M = A - \bar{A}$$

M: matrix with mean = 0; A: original; $\bar{A}$: mean

Loading vector V consists of weights applied to each variable.

Z_i is the score of each weighting

$Z = MV$

Project onto top P components

PCA via SVD

- 1) compute the covariance matrix

$$B = \frac{1}{m-1} M^T M$$

B is covariance captures relationships

M is zero mean matrix

Covariance Matrix should tell you how much each coloumn varies.

$$B \approx \begin{bmatrix} 21 & 17.5 & 27.5 \\ 17.5 & 19 & 25.5 \\ 27.5 & 25.5 & 36.3 \end{bmatrix}$$

□ is variance of the col

Must be B positive semi definite

$B = E \Lambda E^T$

2) Compute eigenvalues and eigenvectors

- eigenvectors of B are the principal components
- eigenvalues tell us how much variance each component captures

$$\lambda = \begin{bmatrix} 75.60 \\ 4.82 \\ 0.17 \end{bmatrix}$$

→ first principal component captures 75.60%

$$= V \frac{\Sigma^T \Sigma}{m-1} V^T \rightarrow B = E \Lambda E^T$$

3) Compute the SVD

- $M = U \Sigma V^T$
  - U contains left singular vectors, Σ is diagonal matrix, V contains right singular vectors
  - right singular values V are same as eigenvectors of B: $E = V$
  - singular values σ_j relate to eigenvalues: $\lambda_j = \frac{\sigma_j^2}{m-1}$

Computing PCA Scores

- after finding principal components

PCA scores: $Z = MV$

- Z is score matrix
- Each row in Z represents data projected on new PCA basis

Ex

$$Z_1 = \begin{bmatrix} -0.67 & 1.75 \\ 12.80 & -1.84 \\ -0.5 & 2.76 \\ 0.06 & 0.40 \\ -11.64 & -2.26 \end{bmatrix}$$

→ First col reps scores along first principal component

→ Strong patterns in first component

- Euclidean norm: $\|A\|_2 = \sqrt{\pi \max(A^T A)}$
  - Largest Singular Value

- Eckart-Young Theorem

$$A \approx U_p \Sigma_p V_p^T$$

  - Best low rank approx, keeps first p components
  - if most variance is captured in first few, discard rest

Scores 0 mean data

$Z = MV$

First PCA component is the first eigen vector of B,

Matrix Norms: Axioms

- $\|A\| \geq 0$
- $\|A\| = 0$ iff $A = 0$

$\|\alpha A\| = |\alpha| \, \|A\|$

$\|A + C\| \leq \|A\| + \|C\|$

L2 norm (spectral norm)

- largest singular value (σ_1, square root of largest eigenvalue of $A^T A$). Represents max stretch of A when applied to a vector.

Frobenius norm: measures size of matrix as square root of sum of absolute elements

Lecture 17

Lecture 17 - Lecture Notes PDF

Scatter Matrix and Covariance Matrix

- Scatter matrix S: used in PCA instead of Covariance matrix
- $S = M^T M$

where M is the zero mean matrix

- Eigenvectors of S are the same as those of covariance matrix B

$$B = \frac{1}{m-1} S$$

- Scatter matrix is scaled version of covariance matrix

SVD and PCA

- PDA closely related to SVD, shown in decomp: $M = U \Sigma V^T$
- Scatter matrix can be written as $S = V \Sigma^2 V^T$
- right singular vectors V are eigenvectors of Scatter matrix S
- Singular values in Σ relate to eigenvalues of S

PCA Scores (Projection of Data)

- PCA scores represent data projected onto principal components:

$$z_j = M v_j = \sigma_j u_j$$

- First PCA score is most important σ_j
- Each score vector z_j captures a diff aspect
- form lower dimensional representation

PCA for Dimensionality Reduction

-3 ways to reduce dimensions

- compute first P PCA scores
- compute first P eigenvectors of Scatter matrix
- Compute SVD of o mean data matrix M

Scree Plot: Choosing number of Components

- Scree plot helps determine, the optimal number of Principal Components
- x-axis represents principal component index
- y-axis represents the variance explained by each component
- where plot elbows off is ideal # of components

Low rank approx and E. Young Theorem

- Best approximation of a matrix given by its truncating SVD $M \approx \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \dots \sigma_p u_p v_p^T$
- Keeping largest singular values, gives best lower dim aprox
- usefull for compression, clustering

Latent variables - eigen values of B (covariance matrix)

- First PCA score vector z_1, approximation $\sigma_1 \vec{v}_1$
- Approximate M as rank-p, M_p, use $Z_p = M_p V_p$

Measure $\|A - C\|$

Consider : $\text{rank}(C) = 1$

Build C from $\vec{z} \in \mathbb{R}^m$

$$C = [\alpha \vec{z}, \alpha_2 \vec{z} \dots \alpha_n \vec{z}]$$
$$= z \vec{\alpha}^T$$

Alternative use $\|\vec{w}\| = 1$

$$\alpha = p\vec{w}$$
$$C = z p \vec{w}^T$$

Optimal values $\vec{z}, \vec{p}, \vec{w}$

$$C = \vec{v}_1 \sigma_1 v_1^T$$

- elbow tells us good K value for approx of rank of A.

Eckart Young Theorem

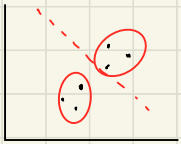
- optimal rank-K approx means using only top K Singular values

$$A = \begin{bmatrix} 3 & 4 & 5 \\ 2 & 4 & 6 \\ 0 & -2 & 3 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$A \simeq 9 \times$ first col $U \times$ first row V^T

Lecture 17 - In Person notes



- Can cluster many ways
- A cluster is a set of data vectors
- Clustered by a partition
- If K partitions, K clustered set

Consider 2 Scenarios

1) given 2 vectors, partition vs d_i
   Calculate distance from 2 points, take minimum
2) given two sets, compute then B_1 / B_2

This is K means clustering

While loop - can vary, as long as order of steps does not change

How do we find optimal K for given.

- no given methods
- depends on data

· minimum variance, orthogonal

	x_1	x_2	x_3	x_4
x_1	0			
x_2		0		
x_3			0	
x_4				0

Looks like weighted agency matrix

Consider:

what is classification of a new vector X.

$q_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ Hypothesis

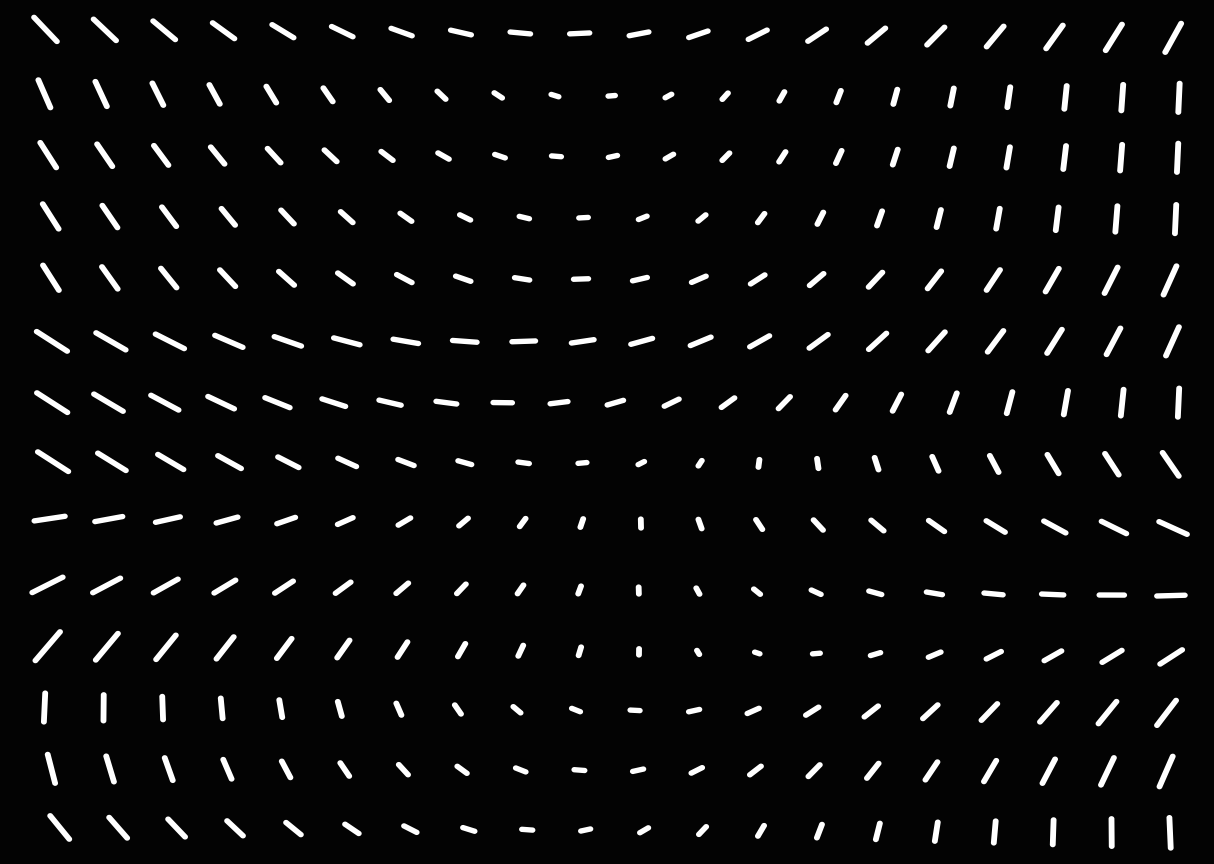
$m = g_1 - g_2$

$m \cdot x - P = 0$

$b = -0^T p^{\rightarrow}$

Equation holds for every point

Unit 4



Lecture 19

Lecture 19 - PDF Notes

Supervised vs Unsupervised Learning

- Supervised: Labelled data (classification/Regression)
- Unsupervised: Unlabelled data (Clustering)

Clustering

- grouping data points, so that points within same group are more similiar to each other than to points in other groups.
- Each group is a cluster and can be distinct or overlapping

Clustering Def

1) Cluster as Set: A cluster S_i is the subset of data X, $S_i \subset X$
   - Each data point belongs to exatcly one cluster

2) As a centroid: Cluster s_i has a centroid g_i such that:
   - Every data point belongs to cluster whos centroid is closest
   - Assigning each point to the nearest centroid

$$S_i = \{ v \in X : \|v - g_i\| < \|v - g_K\|, \forall g_K \neq g_i \}$$

K-means Clustering

- Finding K groups that best cluster data
  1) Choose K clusters
  2) Randomly initialize centroids
  3) Assign each data point to nearest centroid
  4) Update each centroid to be average of its assigned points
  5) Repeat until centroids don't change

Evaluating Performance

- Accuracy Check: Predicted labels vs True labels
- Misclassified points
- Human vs Algorithm

Lecture 20

Lecture 20 - PDF Notes

What is Binary classifications?

- Assigning a data vector to one of two classes
- Uses supervised learning (each data point has a label)
- Classes are non-overlapping
- Labels are often +1, -1

Linear Separability & Decision Boundaries

- a dataset is linearly seperable if a 2D straight line, 3D plane, or hyperplane can seperate the two classes
- If a data point is in class +1, should be closer to the respective centroid.
- Boundary between two classes is defined by: $\|t-g_1\|^2 = \|t-g_2\|^2$
  - The data vector t is on the decision boundary if it is equidistant from both centroids

Hyperplane Represesentation

- Hyperplane is a decision boundary that seperates data into classes
- Equation of hyperplane: $m^T X + b = 0$
- where m is normal vector
- X is a data point
- b is bias term

Properties:

- all points on the hyperplane satisfy $m \cdot t + b = 0$
- Points in +1 class $m \cdot t + b > 0$ Positive half space
- Points in class -1 $m \cdot t + b < 0$

Mathematical Derivation of Decision Boundaries

- Given centroids g_1, g_2:
  Vector Difference: $m = g_1 - g_2$
- Midpoint: $h = \frac{g_1 + g_2}{2}$
- hyperplane Eq: $m \cdot t = m \cdot h$
  $m^T t + b = 0$
  $b = -m \cdot h$

EX

$g_1 = (3,4), g_2 = (1,2)$

$m = (3,4) - (1,2) = (2,2)$

$h = \frac{(3,4)+(1,2)}{2} = (2,3)$

$b = -m \cdot h$

$= -(2,2) \cdot (2,3) = -4+6 = -10$

Eq $= 2x + 2y - 10 = 0$

Generalization More than 2 clusters

- more than 2 clusters, use multiple hyperplanes
- amt of clusters = amt of centroids
- Each pair of clusters has its own decision boundary (hyperplane)

Matrix notation:

$$M = \begin{bmatrix} m_{12}^T \\ m_{13}^T \end{bmatrix} \qquad b = \begin{bmatrix} b_{12} \\ b_{13} \end{bmatrix}$$

data point belongs to Class 1 if:

$$Mx + b \geq 0$$

Hyperplanes for multi class classification

- Compute multiple hyperplanes
- Assign point to closest class

Each class is defined:

$$x \in S_1 \iff (m_{12}^T x + b_{12} \geq 0) \wedge (m_{13}^T x + b_{13} \geq 0)$$

- x belongs to S_1 if it lies on the positive side of both seperating hyperplanes

Computing Distance to hyperplane

1) Convert normal vector to unit vector: $n = \frac{m}{\|m\|}$

2) Transform bias Term: $c = \frac{b}{\|m\|}$

3) Compute Distance: $d = \frac{|m \cdot x + b|}{\|m\|}$ or use $d = n^T x_0 + C$

Ex

hyperplane:

$$3x + 4y - 10 = 0$$

at a point $P = (1, 1)$, the distance is

$$d = \frac{|3(1) + 4(1) - 10|}{\sqrt{3^2 + 4^2}} = \frac{3}{5}$$

This is how far the point is from decision boundary.

Rough Notes

Consider "new" $\vec{v}$

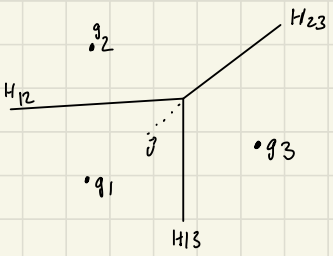
$\vec{v}$ is in S_1 iff its on positive side of hyperplane

$$\vec{w}^T \vec{v} + b \geq 0$$

Multiple Clusters: S_1, S_2, S_3

Centroids g_1, g_2, g_3

Hyperplanes: H_{12} H_{13} H_{23}



$$\vec{w}_{12}^T \vec{v} + b_{12} \geq 0$$
$$w_{13}^T \vec{v} + b_{13} \geq 0$$

$$b_1 = \begin{bmatrix} b_{12} \\ b_{13} \end{bmatrix}$$

$$w = \begin{bmatrix} w_{13}^T \\ w_{13}^T \end{bmatrix}$$

$$w_1 \vec{v} + \vec{b_1} \geq 0$$

DB index

- How good are these clusters.
- in S_1, have m_1
  in S_2, have m_2
  Centroid $\vec{g_1}$
  Centroid $\vec{g_2}$

Measure: $\|g_1 - g_2\|$

Larger DB index, means worse cluster

Smaller DB index, better cluster

Measure: mean distance within partition.

$$DB = \frac{d_1 + d_2}{\|\vec{g_1} - \vec{g_2}\|}$$

→ distance between centroids

Dispersion: $d_1 + d_2$

Dispersion in clustering is how spread out the points are within a cluster.

- Variance
- standard Deviation
- Range

$A = U \Sigma V^T$

$= U \Sigma V^T$

$= v r \Sigma v^T$

$= Q$

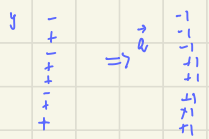
Take Transpose: $A = B' Q'$

Consider

How good is H_+?

Labels are + −

classes are ○ △



How many data points are correct?

Positive match 4 vectors

Negative match 4 2D vectors

How many are wrong?

True Positive (TP) Correct

True Negative (TN)

False Positive (FP) False Alarm

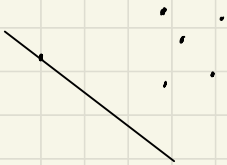
False Negative (FN) Miss

Labels \ Class	+	−
+1	TP	FN
−1	FP	TN

True Pos rate = $\frac{TP}{TP + FP}$ } Sensitivity

Accuracy: $\frac{TP + FN}{TP + FN + TN + FP}$

$[X \mid 1] \cdot \vec{w} = \text{Scores}$



$\vec{s}$	
−1	−
−2.5	+
−2	−
2.5	+
4.0	−
4.5	+
6	+

Confusion Matrix

Labels \ class	+1	−1
	TP	FN ↳ what it really is ↳ what we classified
	FP	TN

Relative matrix

has row sums equal to one

e.g. Confusion matrix

R.M has 2 degrees of freedom

row sums can be used to compute col sums

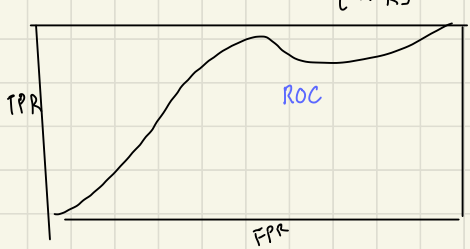
TP	FN
FP	TN

4 degrees of freedom

which is independent values

FPR → TPR

$\begin{bmatrix} FPR \\ TPR \end{bmatrix}$



4	1
1	2

⇓

0.8	0.33
0.33	0.66

Lecture 22

Lecture 22 - PDF Notes

- evaluating performance of classification models using confusion matrix
- compare actual labels to predict sensitivity, specificity, type I and type II errors.

Confusion Matrix

- 2x2 binary classification table

Positive (+1)	True Positive	False Positive
Negative (-1)	False Negative	True Negative

Positive instance - data points

Type I error - false pos

Type II error - false negative

right metric depends on application

Accuracy: $\frac{TP+TN}{P+N}$ - correctly identified classes

Sensitivity (recall): $\frac{TP}{P}$ - Identify positive instances

Specificity: $\frac{TN}{N}$ - Identify Negative instances

Type I error rate: $\frac{FP}{N}$ - Prob of rejecting H_0

Type II error rate: $\frac{FN}{P}$ - Prob of incorrectly classifying Pos as negative

Precision: $\frac{TP}{TP+FP}$ - Prob of predicted positives, that are truly positives

Relative Confusion Matrix

- represents proportions rather than raw counts

Predicted/Actual	Pos (+1)	Neg (-1)
Pos (+1)	TPR (Sensitivity)	FPR
Neg (-1)	FNR	TNR (Specificity)

Each row sums to 1 (TPR + FPR = 1)

FNR + TNR = 1

- has 2 Degrees of Freedom

Example 1: Medical Diagnosis

A classifier is used to detect a disease. The actual and predicted results are:

Actual / Predicted	Positive (Disease)	Negative (No Disease)
Positive (Disease)	50 (TP)	10 (FN)
Negative (No Disease)	5 (FP)	100 (TN)

Calculating Metrics:

- Accuracy = $\frac{50+100}{50+10+5+100} = 0.91$ (91%)
- Sensitivity (TPR) = $\frac{50}{50+10} = 0.83$ (83%)
- Specificity (TNR) = $\frac{100}{100+5} = 0.95$ (95%)
- Type I Error Rate (FPR) = $\frac{5}{5+100} = 0.05$ (5%)
- Type II Error Rate (FNR) = $\frac{10}{10+50} = 0.17$ (17%)

Since missing a disease is worse than a false alarm, **reducing FNR is crucial.**

Independent variables cannot be categorical, you can't scale category

label is dependent var in classification

Scoring function takes features and maps them to real number.

Lecture 23

Lecture 23

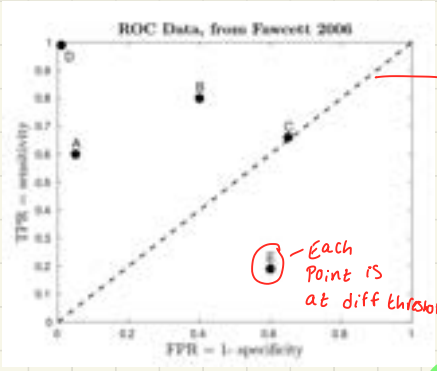
- Reciever Operator Characteristic (ROC) curves, vissually evaluates classifier's performance
- FPR, TPR, AUC (Area under curve)

Roc Curve

- method to evaluate Classification Performance
- used in ww2 for radar detection
- ROC curve plots TPR against FPR

$TPR = \frac{TP}{P}$, $FPR = \frac{FP}{N}$ or $1 - TPR$

- x axis of ROC curve = FPR
- y axis of ROC curve = TPR



Each circle is ratio of $\frac{TPR}{FPR}$

Each circle is its own confusion matrix

How is the ROC Curve Generated

1) Classifier assigns scores to data points
2) Threshold applied, if score > threshold, its Pos
   - if score < below, its neg
3) Changing threshold shifts TPR, FPR creating diff points

Area under Curve (AUC)

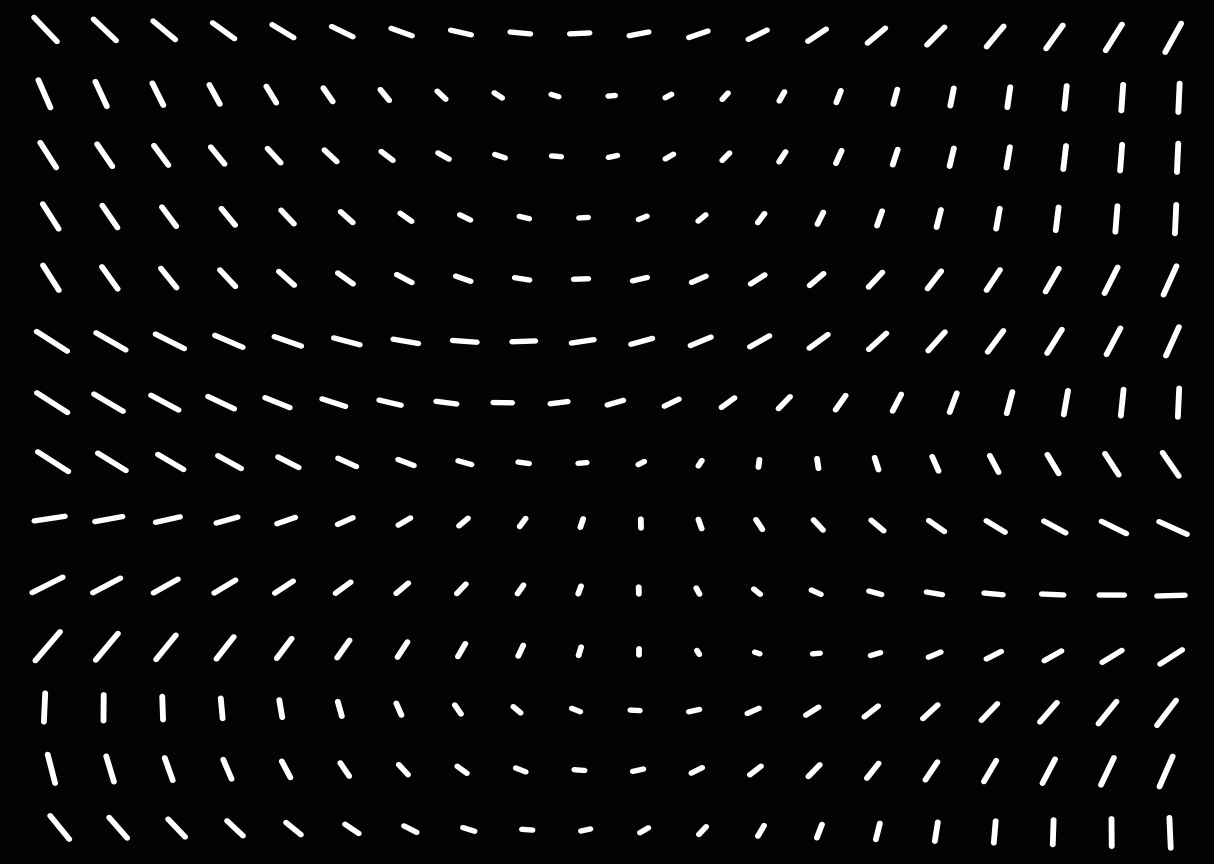
- AUC is a single number summary of ROC performance
- ranges from 0 to 1:

1.0 - Perfect classifier, TPR = 1, FPR = 0
0.5 - random guessing TPR ≈ FPR = 0.5
< 0.5 worse than random, (You better off guessing)

Closer to 1, the better

- Higher AUC indicates better discrimination between classes.

Unit 5



Lecture 25

Lecture 25 - PDF Notes

- relationship between odds, probability, and logistic regression.
- logistic function maps: Score → Probability

Odds and Probability

- odds represent ratio of likelihood of event occuring
- Probability is P, $S = \frac{P}{1-P}$, $P = \frac{S}{1+S}$
  S is odds

Odds of 9:1 means 9x more likely to occur $P = \frac{0.9}{0.1} = 0.9$

odds of 1 in 4 means $\frac{1}{4+1} = \frac{1}{5} = 0.2$

Logistic Function and Logistic Regression

- linear classification, we use hyperplane H to seperate two classes
- data point x is classified as posisitive (+1) if:

$$d(x) = w^T x + b \geq 0$$

  - $d(x)$ is signed distance, from x to hyperplane
  - w is normal vector to hyperplane
  - b is bias term
  - $d(x) \geq 0$ classify as +1
  - $d(x) < 0$, classify as -1

- instead binary decision, we want a probability $p(x)$ that point belongs to class 1
- logistic function, which transforms real number d into a probability

$$P(x) = \frac{1}{1 + e^{-d(x)}}$$ → when d is very large, function approaches 1

maps $d \in (-\infty, +\infty)$ to $P \in (0,1)$

How is log function derived?

1) Define odds as: $S = \frac{P}{1-P}$
2) log both sides $\ln(S) = \ln\left(\frac{P}{1-P}\right)$
3) log regression, assume: $d = \ln(S)$
   which leads to $d = \ln\left(\frac{P}{1-P}\right)$
4) Solving for P: $P = \frac{e^d}{1+e^d}$

$P = \frac{1}{1+e^{-d}}$

d is score (distance from decision boundary)

Probability its in a specific class

Notes: log regression predicts probabilities
- model learns weights w and bias b
  s.t $P(x) = \frac{1}{1+e^{-(w^T x+b)}}$
- useful when uncertainty matters
- maps Score to probability

Neural Networks

- logistic function (sigmoid activation) is used in neurons
- logistic function allows for: non linearity in deep learning models
  - back propagation

Properties of the logistic Function (maps score to probability)

Symmetry: $1 - P(d) = P(-d)$

$P(d)$ is class +1, $P(-d)$ is class -1

Slope: $P'(d) = P(d)(1 - P(d))$

Hyperplane Classification and Pseudo Distance

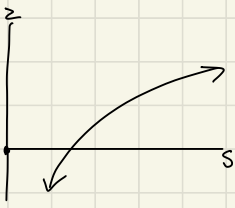
- a Hyperplane H is defined by w and b
  - instead of true distance d, can use $v(x) = w^T x + b$
  - Sign of $v(x)$ tells us if x is above or below hyperplane
  - Commonly used in SVM's

Logistic regression

- Decision function: $P(x) = \frac{1}{1 + e^{-(w^T x + b)}}$
- w (weights) and b (bias) are learned from training data
- Thresholds, if $P(x) > 0.5$, classify +1
  if $P(x) < 0.5$, classify -1

- Being able to give a probability that a point lies in class 1 or 2 is significant if the point is close to the hyperplane.

Mapping $S \rightarrow Z$



Choose

$z = \ln(s)$

$s = e^z$

$p = \frac{e^z}{1 + e^z}$

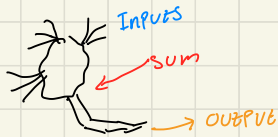
Relating score to probability

Lecture 26

Lecture 26 - Artificial Neurons for Classification

- Can model a simple neuron and use it for binary classification. (Choosing 0/1)

Biological Inspiration



- Inputs (electrical signals) come in through dendrites
- These are combined in the cell body
- If the sum crosses a threshold, the neuron fires
- Output travels through axon

Math Model of Neuron

- Weighted sum
- Each input a_j is multiplied by weight m_j. Then summed

$$u = \sum_{j=1}^{n} a_j m_j$$

- We include bias (threshold) B, so neuron fires when $u \geq B$

So, we get: $\sum a_j m_j - B \geq 0$

or

$\sum x_j w_j - B \geq 0$

Note: Neuron fires when $\sum_{j=1}^{n} x_j w_j + b \geq 0$

Augmented Vectors/ Data Matrix and Labels

- Data as augmented vector $\vec{x} = [\vec{a}, 1]$
- weights/bias $\vec{w} = [\vec{w}, b]$, where $b = -B$

$$U = x^T \vec{w}$$

(inputs, weights)

- all input vector $\vec{x_i}$, are rows in matrix X
- Each data point has a label $y_i \in \{0, 1\}$

weighted sum for observation: $u_i = x_i^T \vec{w} + b$

Neuron fires iff $u_i \geq 0$

$\vec{w}, b$ is (pseudo distance) — want to learn w, b

$v_i = x_i^T \vec{w} + b$

assign $v_i \geq 0$ to 1

Activation Function

- Function decides how neuron behaves
- takes input u, outputs score $z = \varphi(u)$
  - like sigmoid/step function

Quantization - process of mapping a large set of values to a smaller set

Simple classifier with Heavyside Function

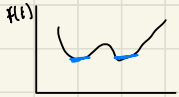
- Heaviside function converts linear sum to binary class

$$H(u) = \begin{cases} 0 & \text{if } u < 0 \\ 1 & \text{if } u \geq 0 \end{cases}$$

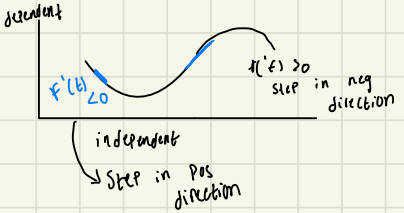
- Just determines if neuron fires

$$a_i(\vec{w}_i \vec{x_i}) = H(x_i^T \vec{w})$$

Lecture 26 - In Class notes



Recall Stationary Point
$f(t)=0$



Using $t_{k+1} = t_k + s$ is oscillation

- Negate Derivative, step

Consider Learning Rate: (eta)
at step "k", direction is dk
at $-f'(t_k)$

- eigenvalues / eigenvectors used to find curvature of surface

Consider $f(w_1, w_2)$
as derivativive

Instead Partial derivative
with respect

gradient is one form
$\nabla f_2(\vec{w}) = \left(\frac{df}{dw_1} \quad \frac{df}{dw_2}\right)$
$= [\quad , \quad]$

The direction of steepest ascent
Defines as $D = \eta \begin{bmatrix} 2w_1 \\ 4w_2 \end{bmatrix}$

Consider Hyperplane and Neuron
H/ specified with vector m and bias scalar B
take data matrix put 1 on it

Linear response is $V = A \cdot m + b$
Fact: ∇V = observation

Lecture 28

Lecture 28: Elementary Numerical Optimization

- Finding the best (min) value of a function using numerical optimization.

Scalar Optimization - 1D

- Stationary Point:
  - Point t^* is a Stationary Point, if the derivative is 0
    $f'(t^*) = 0$
  - This is where the function stops increasing or decreasing
    - could be min, max, flat spot

Fixed Stepsize Gradient Descent (1D):

- $F(t)$ - objective function (minimizes error)
- t_k - current guess
- η - stepsize (learning rate)
- $f'(t_k)$ - derivative at current guess
- Then $t_{k+1} = t_k - \eta f'(t_k)$
- move opposite to derivative (keep repeating until a min)

Note
- if η is too big, you overshoot
- if η is too small, super slow convergence

Vector Optimization (Multivariable)

- Functions with multiple inputs
  $f(\vec{w}) = \quad \vec{w} = [w_1, w_2, \dots w_n]$
- Partial derivatives
  - take derivative each variable seperately, gradient
    $\nabla f(\vec{w}) = \left[\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2} \dots \frac{\partial f}{\partial w_n} \right]$
  - tells you how steep function is in each direction

Steepest Descent (Multi-D)

- move in direction opposite gradient:
  $\vec{d} = -\nabla f(w)$
- Then update guess:
  $\vec{w}_{k+1} = \vec{w}_k + \eta \vec{d}$

Algorithm Summary
1) Start initial guess $\vec{w}_0$
2) Compute gradient $\nabla f(\vec{w})$
3) move opposite gradient:
   $\vec{w}_{k+1} = \vec{w}_k - \eta \nabla f(\vec{w}_k)$
4) Repeat until gradient is close to 0.

Lecture 29

Lecture 29 - Logistic Regression and Artificial Neuron

Artificial Neuron and Data Representation

- Data matrix: $A \in \mathbb{R}^{m \times n}$, labels $y_i \in \{0, 1\}$
- Augmented input: $x = [a^T \; 1]$
- weight vector: $w = [m \; b]^T$
- Linear response: $U(w; x) = x \cdot w$

Logistic Activation Function

- activation function: $\phi(v) = \frac{1}{1+e^{-v}}$
- Properties: Continous, differentiable, easy compute
- Neuron response: $\phi_i = \phi(w; x_i) = \frac{1}{1+e^{-x_i \cdot w}}$ → ?

Learning with Steepest Descent

- no closed form solution to learn weights, we use iterative optimization
- we want to minimized the squared error

$$f_i = \frac{1}{2}(y_i - \phi(v_i))^2$$

↳ label ↳ score

- Find gradient:

$$\nabla f_i = -(y_i - \phi(v_i)) \cdot \phi'(v_i) \cdot x_i$$

where $\phi'(v) = \phi(v)(1 - \phi(v))$

$$b_i = (y_i - \phi(v_i)) \cdot \phi'(v_i)$$

$$\nabla f_i = -b_i x_i$$

?

Single Input Logistic Regression Ex

input $x = [2, 1]$

$w = [0.5, -1]$

$y = [1]$

1) linear response: $v = x \cdot w = 2(0.5) + 1(-1) = 0$

2) activation: $\phi(v) = \frac{1}{1+e^0} = 0.5$

3) gradient factor: $b = (1 - 0.5)(0.5)(1 - 0.5) = 0.125$

4) gradient vector $g = bx = 0.125 \cdot [2, 1] = [0.25, 0.125]$

5) update weights w Learning rate $\eta = 0.1$ $w = w + \eta \cdot g = [0.5, -1] + 0.1 \cdot [0.25, 0.125] = [0.525, -0.9875]$

Linear Activation - Limitations

- Linear activation functions just pass input through unchanged
- Problem: only seperates linearly seperable data
  - ↳ real world: often non linear

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

↳ Semilinear range 0 - 1

- Perceptron - most basic type of artificial neuron

Residual Error and Functions

- logistic activation $z(\hat{w}; \hat{x}i)$
- Residual error is $ri = yi - z(w; xi)$
  - actual label
  - predicted label
- Logistic function is smooth approx of heavyside function (differentiable) (no closed form solutions)
- Plotting errors is important for finding min values

Other Notes

- Single neuron with linear activation, by adding linear neurons, you still get a linear function
- Errors are added for all m observations
  - ↳ total loss/error is computed by adding individual errors
- Sigmoid activation: error and weights are nonlinear
  - ↳ Rely on numerical techniques:
    - Squared error by steepest descent
    - Logistic regression by max likelihood

Lecture 31

Lecture 31 - Non linear Seperation - Embeddings and Gram matrix

- How do we work with nonlinearly seperable data in ML?

Core Concepts

- Embedding: Transforming data to a higher-dimensional space
- Kernel Functions: Computing dot products in this higher dimensional space without doing the embedding explicitly
- Gram matrix: A matrix of these dot products

Motivation

- data isn't always linearly seperable, like donut shaped data
- Solution: embed to higher dimension, where linear seperation is possible

Linear Seperation via Embedding

- given data points $[u_1, u_2]$

Ex1

$[u_1, u_2] \rightarrow [u_1, u_2, u_1^2 + u_2^2]$

adding squared norm, adds distance to origin

Ex 2

$[u_1, u_2] \rightarrow [u_1^2, u_2^2, \sqrt{2}\, u_1 u_2]$

- maps data using quadratic terms

key: don't need to compute embedding, if you can calculate dot product to higher space using Kernel.

Kernel Trick and Gram matrix

- When doing PCA, SVM in higher dimensions:
- need dot products between embedded vectors $\hat{a_i} \cdot \hat{a_j}$
- instead of computing those directly, we use Kernel function $K(a_i, a_j)$

$K(u, v) = f(u) \cdot f(v)$

This gives us Gram Matrix, where:

$K_{ij} = K(a_i, a_j)$

- Symmetric
- Positive Semi definite

- Kernel function takes 2 inputs, computes dot products of their embeddings, without computing embeddings.
- Kernel formula is like: $(u \cdot v + 1)^2$

Lecture 31 - In Class Notes

Consider

Circle is $x^2+y^2=r^2$

$(0,0)$ has radius 0

$(1,1)$ has radius 2

$(1,-1)$ has radius 1

Try $(x,y) \hookrightarrow (x, y, x^2+y^2)$

fish hook

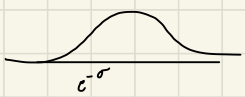
lower dimension to higher dimension

An embedding is a map from a low dimensional vector space to a high dimension vector space W

The symbol is $\hookrightarrow$

Deduce a hyperplane for b solutions.

Let B be a hyperplane with unit normal and bias scalar



How can we use a gaussian distribution for embedding?

$e^{\tau} \approx \sum_{i}^{\infty} \frac{\tau^N}{N!} = 1$, add on term for every dim

Some embeddings require a lot of memory and how do we interpret them.

Note: will treat row as a vector

Embedding: $\mathbb{R}^n \hookrightarrow \mathbb{R}^p$

augmentation: $u \hookrightarrow \hat{u}$

$[v_1, v_2] \hookrightarrow [v_1\ v_2\ \|u\|^2]$

polynomial:

$[v_1, v_2] \hookrightarrow [v_1^2\ v_2^2\ \sqrt{2}v_1 v_2]$

After Embedding

- original $A \in \mathbb{R}^{m \times n}$
- After $\hat{A} \in \mathbb{R}^{n \times p}$

Then, apply PCA on $\hat{A}$

, compute scatter matrix $\hat{S} \in \mathbb{R}^{p \times p}$

, get eigenvectors $\hat{V} \in \mathbb{R}^p$

Polynomial embeddings are powerful but number of terms grow fast

Examples of Kernel Functions

Linear Kernel $K(u,v) = u \cdot v^T$

Polynomial Kernel $K(u,v) = (u \cdot v^T + c)^2$

Gaussian (RBF) $K(u,v) = \exp(-\gamma \|u-v\|^2)$

Extra

Kernel function: any symmetric positive semidefinite function

Gram matrix: constructed from kernel function over all pairs of data vectors

- map zero-mean matrix $M \hookrightarrow \hat{M}$
- Scatter matrix variables is $\hat{S}_y = \hat{M}^T \hat{M}$

Lecture 32

Lecture 32 - Nonlinear Seperation - Kernel PCA

- Using PCA techniques on nonlinearly seperable dat - using Kernel trick

Core Concepts

- Avoid embedding : skip direct computing
- Kernel function: simulate dot prod
- perform PCA : on Gram matrix that holds those dot produces

Why Kernel PCA?

- PCA involves : centering data, finding direction with most variance, linear structure
- Sometimes data has curves, twists, spirals that PCA can't unwrap
  - ↳ Instead we use Kernel function to perform Kernel PCA

Steps in Kernel PCA

1) Define Kernel function:

$$K(u, v) = \exp\left(-\frac{\|u - v\|^2}{2\sigma^2}\right)$$ → gaussian example

2) Build Gram Matrix:

$$K_{ij} = K(a_i, a_j)$$

- Symmetric / positive semi definite, represents dot products in embedded space

3) Centre Gram matrix

- to remove bias (shift), centering matrix G_m:

$$\hat{K}_U = G_m K G_m$$

4) Do PCA on Centered Gram matrix

- Eigenvalues π_j
- Eigenvectors v_j
- Scores: $\hat{z} = \hat{v}\hat{\Sigma}$

∴ now each data point is represented by projection on new principal components but in Kernel space

Fishers IRIS Dataset

- normal PCA misses 2 points (missclassifies)
- Kernel PCA with gaussian kernel fixes it:
  - Choose $\sigma^2 = m$
  - Gram matrix → centre it → eigen decompose
  - top 2 components, cluster K means

Scree plots

- Plot eigenvalues of centered Gram matrix
- tells you how many principal components to keep
-

Scatter matrix observations

- data matrix $A \in \mathbb{R}^{m \times n}$

1) Center data:

$$\hat{A} = \frac{1}{m} \cdot 1^T A$$

- gives mean row vector of A
- 1 is vector of ones size m

Then:

$$M = A - 1 \cdot \hat{A}$$

which subtracts mean from each row A, giving mean cented data matrix M

$$M = \underbrace{(I - (1/m) \cdot 1 \cdot 1^T)}_{\text{centering matrix}} \cdot A$$

$$M = G_m \cdot A$$

PCA: Recap

- $S_V = M^T m$ (want to find principal components, analyze covariance matrix)
  - ↳ covariance in feature space
- SVD of M:

$$M = U \Sigma V^T$$

Then $S_V = V \Sigma^T U^T U \Sigma V^T = V \Lambda_V V^T$

$\Lambda = \Sigma^T \Sigma$ contains eigenvalues of S

- $Z_V = MV = U\Sigma$

Transformed data scores

- For PCA, we can use $\hat{S}_V$ or $\hat{S}_U$

$$\hat{S}_U = G_m [\hat{A}\hat{A}^T] G_m^T$$
$$= G_m \hat{W} G_m^T$$

Kernel PCA

$$\hat{S}_U = \hat{U} \hat{\Lambda} \hat{U}^T$$

Kernel Scores

$$\hat{Z} = \hat{U} \hat{\Sigma}$$

Slower than PCA

Lecture 34

Lecture 34 - PDF Notes

- Binary Classification using logistic regression

2 main loss functions:

1) Squared Error
2) Negative log-likelihood

Goal: Learn best way to seperate binary data
- ↳ Define objective function (error)
- ↳ Use optimization to minimize it.

Loss Functions/Error Models

1) Squared Error Loss
   - MSE $E_2(X) = \sum_{i=1}^{m} (y_i - \phi_i)^2$
   - $y_i \in \{0, 1\}$
   - ϕ_i = Prediction from logistic function

Simple, bounded — not ideal for probabilities

2) Negative Log-likelihood (Log loss) → Very high prob for correct class, Very low prob for incorrect class
   - better for probabilistic models

$$l_i = \begin{cases} -\ln(1 - \phi(v_i)), & \text{if } y_i = 0 \\ -\ln(\phi(v_i)), & \text{if } y_i = 1 \end{cases}$$

   - Can be unified as:

$$l_i = (1 - y_i)(-\ln(1 - \phi(v_i))) + y_i(-\ln(\phi(v_i)))$$

   - Total loss: $E_1(X) = \sum_{i=1}^{m} l_i$

Reflects confidence in classification
Can lead to large weights in optimization

Visual Comparison

- MSE: Errors are bounded
  - Error plateaus for wrong predictions
- Log loss
  - Errors are unbounded
  - Strong penaly for confident wrong predictions

Note:
- Squared error are used in simple neural networks
- log loss is standard in log regression

MSE: $[-30, +15]$
log loss: $[-200, +80]$

- Distance matrix
  - Entry $c_{ij} = \|v_i - v_j\|$

    Define adjacency using ε-neighborhood

    $a_{ij} = \begin{cases} 1 & \text{if } 0 < c_{ij} \leq \varepsilon \\ 0 & \text{otherwise} \end{cases}$
  - Compare K means

Laplacian: $K = D - A$

Eigen decomp: $K = Q \Lambda Q^T$

Eigenvectors: $a_n = \frac{1}{\sqrt{n}} \vec{1}$

$a_{n-1} \cdot a_n = 0$

- Spectral Clustering $K = 2$

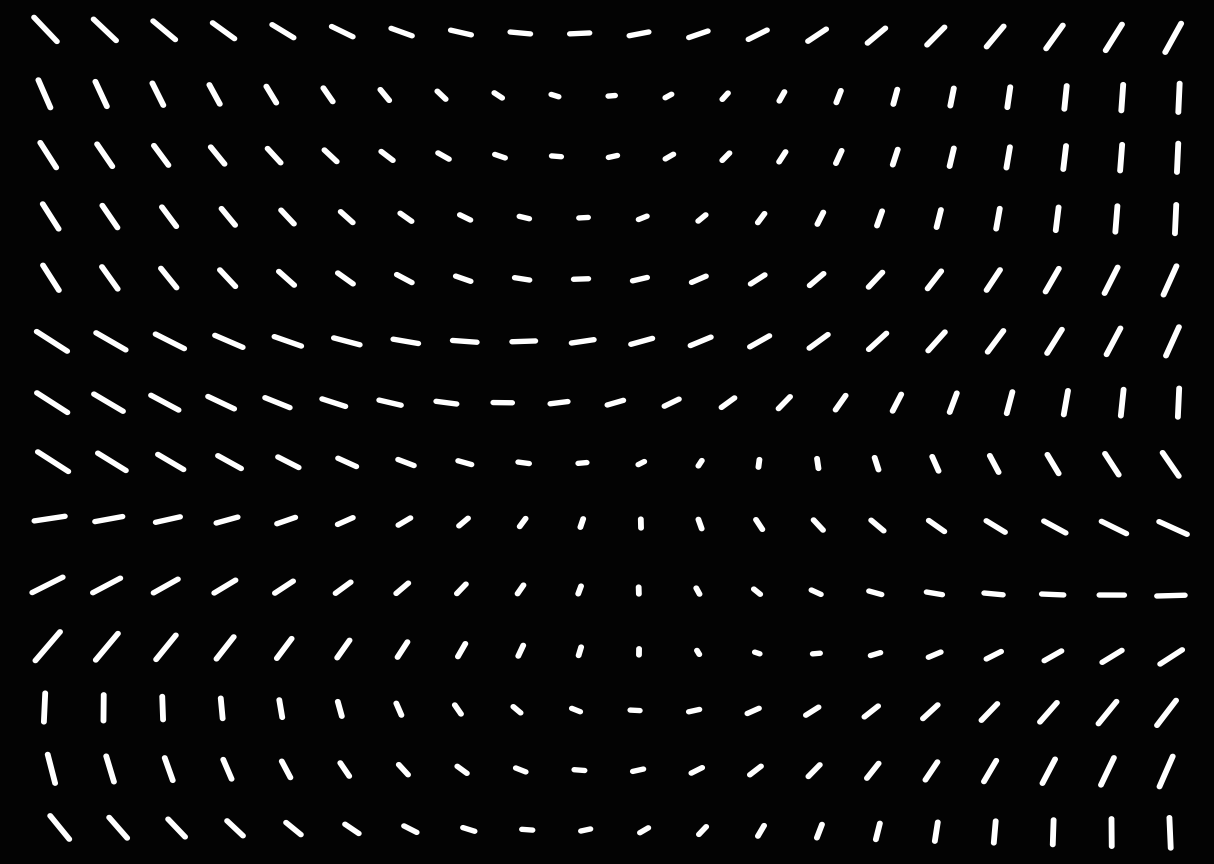
  $E_2 = [a_{n-1} \; a_n]$

  - apply K means with 2.

- weighted graph, gaussian Kernel
  - Kernel PCA and graphs are very closely related

All Quizzes and Practise Units 1-5

Unit 1



Homework Set 1

1.3, 1, 7, 12

1.4, 14, 16

3.1, 10, 16, 17

1.3, q1)

a) $A = \begin{bmatrix} 2 & 2 \\ 1 & 1 \\ 5 & 6 \end{bmatrix}$ Plane

3 by 2 matrix, lin indep in 2 columns, I think this represents a Plane

b) $A_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ 3D

has 3 eigen values with multiplicity 3, lin indep. Could be all of 3D

c) $A_3 = \begin{bmatrix} 1 & 5 \\ 2 & 10 \\ 1 & 5 \end{bmatrix}$ Line

This is a Line

2 vectors which are the same, just scaled by 5.

d) Point

$A_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

Just a point at 0.

Note: Point: Column Space has dimension 0
Line: Column Space has dimension 1
Plane: Column Space has dimension 2
3D: Column Space has dimension 3

7) a) $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}$, $A+B = \begin{bmatrix} 3 & 3 \\ 7 & 7 \end{bmatrix}$ → one indep column

b) $A = \begin{bmatrix} -1 & -3 \\ -2 & -4 \end{bmatrix}$ $B = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$, $A+B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ – NO indep col

c) $A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $A+B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

12) $Ax = b$ has a solution x, if b is in the Column Space of A. To be in the column space means to be in the set of all possible linear combinations, that are lin indep. B is in the column space of A, then b is a combination of columns of A and those numbers give a solution x.

1.4) - 14)

a) $\begin{bmatrix} 3 & 6 \\ 5 & 10 \end{bmatrix}$ b) $\begin{bmatrix} 6 & 7 \\ 7 & -6 \end{bmatrix}$

c) $\begin{bmatrix} 2 & 7 \\ 3 & 6 \end{bmatrix}$ d) $\begin{bmatrix} 3 & 4 \\ -2 & -3 \end{bmatrix}$

3.1) 10) a) $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ b) $\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$

Smallest Subset: $\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 1 & a \\ b & 1 \end{bmatrix}$

16) a) $\dim \text{Col}(A) = 1$

Column space of $A = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}$

all vectors line $(x, 0, 0)$

b) $\dim \text{col}(B) = 2$

Column Space of $B = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}\right\}$ xy plane

c) Column Space of $C = \text{Span}\left\{\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}\right\}$

Column Space is line of vectors $(x, 2x, 0)$

17) 1) Set up augmented

$\left[\begin{array}{ccc|c} 1 & 4 & 2 & b1 \\ 2 & 8 & 4 & b2 \\ -1 & -4 & -2 & b3 \end{array}\right]$ $R2 = 2R1 - R2$

$\left[\begin{array}{ccc|c} 1 & 4 & 2 & b1 \\ 0 & 0 & 0 & 2b1 - b2 \\ -1 & -4 & -2 & b3 \end{array}\right]$ $R2 \leftrightarrow R3$

$\left[\begin{array}{ccc|c} 1 & 4 & 2 & b1 \\ -1 & -4 & -2 & b3 \\ 0 & 0 & 0 & 2b1 - b2 \end{array}\right]$ $R2 = R1 + R2$

$\left[\begin{array}{ccc|c} 1 & 4 & 2 & b1 \\ 0 & 0 & 0 & b1 + b3 \\ 0 & 0 & 0 & 2b1 - b2 \end{array}\right]$

$0 = 2b1 - b2$

$0 = b1 + b3$

Solutions

$b1 = -b3$

$b2 = 2b1$

17b) $\left[\begin{array}{cc|c} 1 & 4 & b1 \\ 2 & 9 & b2 \\ -1 & -4 & b3 \end{array}\right]$ $R2 = 2R1 - R2$ $R3 = R1 + R3$

$\left[\begin{array}{cc|c} 1 & 4 & b1 \\ 0 & -1 & 2b1 - b2 \\ 0 & 0 & b1 + b3 \end{array}\right]$

$\left[\begin{array}{cc|c} 1 & 4 & b1 \\ 0 & -1 & 2b1 - b2 \\ -1 & -4 & b3 \end{array}\right]$

$R1 = 4R2 + R1$

$\left[\begin{array}{cc|c} 1 & 0 & 4b2 + b1 \\ 0 & 1 & -(2b1 - b2) \\ 0 & 0 & b1 + b3 \end{array}\right]$ = if $b1 = -b3$

1.4) 16) a) Prove column space of AB is contained within the column space of A.

For any x, ABx represents lin combo of the columns of AB.

$ABx = A(Bx)$. The vector in the column space of AB is also in the column space of A.

∴ The Column Space of A contains the Columns space of AB.

Study Guide quiz 1

Graph Matrices

- Adjacency Matrix (A):
  - $A_{ij} = 1$ if there is an edge between vertices i and j, 0 otherwise
- Degree Matrix (D):
  - Diagonal matrix where we see how many connections each vertex has.
- Laplacian Matrix (L):
  - $L = D - A$
  - Symmetric and Real, all eigenvalues are non negative.
  - The number of 0 eigenvalues is the number of components

Eigenvalues / Eigenvectors

- $Av = \lambda v$
- Trace is the sum of eigenvalues
- Det(A) is product of eigenvalues

Vector Spaces

- Column space: span of columns of A
- Null space: set of vectors s.t $A\vec{x} = 0$
- Basis:
  - Linearly indep vectors that span the space
  - Orthogonal basis: each vetor is perpendicular to each other
  - orthonormal basis: basis vectors are orthogonal and unit length

Diagonalization

$A = PDP^{-1}$: P is eigenvector matrix
D is diagonal matrix of eigenvalues

- A is diagonalizable if eigenvectors are lin indep
- Symmetric matrices, always diagonalizable

Symmetric / Skew

- Symmetric Matrix:
  - $B = B^T$
  - real eigenvalues and orthogonal eigenvectors
- Skew Symmetric:
  $S = -S^T$
  Eigenvalues are imaginary

Unit 1 Practise Questions

1) $A = \begin{bmatrix} -2 & 4 & 2 \\ 2 & -1 & 1 \\ -2 & 3 & 1 \end{bmatrix} \xrightarrow{\text{RREF}} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

2 l.in indep columns

rank = 2, basis vectors must be 2-D ∴ b) $\begin{bmatrix} -1 & 2 \\ 1 & 1 \\ -1 & 1 \end{bmatrix}$ is correct

2) $A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 0 \\ 2 & 1 & 3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Not in col of A, all lin indep

Which vector is not lin combo of others

Which vector is orthogonal to column space

$A^T C = 0$ b is multiple of col 3

d)

3) $A = A^T$

$\lambda_1 = 5, \quad \lambda_2 = 10$

$\vec{v_1} = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \quad \vec{v_2} = \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$

$A = V \Lambda V^T$

$A = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$

$A = \frac{1}{5} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 5 & 10 \\ -20 & 10 \end{bmatrix}$

$A = \frac{1}{5} \begin{bmatrix} 45 & -10 \\ -10 & 30 \end{bmatrix} = \begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$

4) $\lambda_1 = 1, \quad \lambda_2 = 2$, when $0 < \varepsilon < 1$

$A = \begin{bmatrix} 1+\varepsilon & 1+\varepsilon \\ 0 & 2 \end{bmatrix}$

$\det(A - \lambda I) = 0$

$(1+\varepsilon-\lambda)(2-\lambda) = 0$

$\lambda_1 = 1+\varepsilon \quad \lambda_2 = 2$

$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ Normally $\det = ad - bc = 2$

$\begin{bmatrix} 1-\lambda & 1 \\ 0 & 2-\lambda \end{bmatrix}$

$(1-\lambda)(2-\lambda) = 0$

$\lambda_1 = 1 \quad \lambda_2 = 2$

∴ eigenvalue 2 does not change, eigenvalue 1 increases a bit

5) $A^T B^T = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 4 & -3 \end{bmatrix}$

$= \begin{bmatrix} 5 & 0 \\ 8 & -6 \end{bmatrix}$

$\det(A - \lambda I) = 0$

$(5-\lambda)(-6-\lambda)$

$\lambda = -5 \quad \lambda = 6$

∴ C)

6) Keep in mind $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$

If $A = A^T$, $A \geq 0$

then $x^T A \vec{x} \geq 0$

Practise 1 - Matlab

Binary Clustering of Graph Vertices

1.) Transforming an edge list into an adjacency matrix

2.) Computing Laplacian matrix of a graph

3.) Using Fiedler vector to cluster vertices in 2 groups

Edge list:

- List of connections between nodes/vertices

EX
- 1 2 5 — 1 and 2 are connected, with weight 5

2 3 7

1 3 2

- Suppose, you have 4 eigenvalues, Fiedlers vector is the one associated with second smallest eigenvalue.

- Then seperating nodes into positive/negative clusters of those eigenvectors.

Key Ideas:

- Adjancency Matrix, represents connections/vertices
- weighted egdes
- Laplacian: $L = D - A$
- Fiedler vector: eigenvector, corresponding to second smallest eigenvalue of L partition.

  Neg Entries → Set 1

  Pos Entries → set 2

Helpfull MATLAB Commands

Creation

- Identity matrix: `eye(n)`
- Zeros matrix: `zeroes(m, n)`
- Ones matrix: `ones(m,n)`

Matrix operations

- Transpose: `A'`
- Inverse: `inv(A)`
- Determinant: `det(A)`
- Rank: `rank(A)`

Eigenvalues and Eigenvectors

`[V, L] = eig(A)` — Spectral Decomp

V: Eigenvectors

L: Diagonal matrix of eigenvalues

Null Space:

`null(A)`: orthonormal basis for null space

Graph Construction:

- `A = [011; 100; 100]`
- Degree matrix: `D = diag(sum(A, 2))`
- Laplacian matrix: `L = D-A`

Key Definitions and Concepts

Content
week 1-2
Pre Req, HW1, HW2
Practise 1
class notes 1-6.

Graph Matrices

- Graph $G(V, E, w)$: Set of vertices (V), edges (E) and weights (w)
- Adjacency matrix where $A_{ij} = 1$, if vertices i and j are connected
- Symmetric for undirected graphs
- Degree Matrix $D(G)$: diagonal matrix where each entry D_{ii} = degree of vertex i
- Laplacian Matrix:

  $L = D - A$, Null Spaced dimension = number of graph components

Eigenvalues and Eigenvectors

Note: Use calculator to find inverse matrix or determinant

Eigenvalue: $Av = \lambda v$

where v is an eigenvector, and λ is an eigenvalue

- The trace of a matrix is the sum of its eigenvalues
- The determinant is the product of eigenvalues
- If eigenvalues are distinct, all eigenvectors are lin indep
- If A is Symmetric, all eigenvalues are Real

Diagonalization

- A matrix is diagonalizable if $A = PDP^{-1}$

  where D is diagonal, P Contains eigenvectors.
- Symmetric matrices are always diagonalizable.
- Normal matrices ($AA^T = A^TA$) are diagonalizable
- If all eigenvalues are distinct, matrix is diagonalizable

Special Matrices

- Orthogonal matrices: $Q^{-1} = Q^T$
- Symmetric matrices: $A = A^T$
- Skew Symmetric: $A = -A^T$

2019 Practise Test

1) Given $A = \begin{bmatrix} -2 & 4 & 2 \\ 2 & -1 & 1 \\ -2 & 3 & 1 \end{bmatrix}$

Find a basis for coloumn Space of A.

To span Col(A), coloumns must be lin indep

$\begin{bmatrix} 1 & -2 & -1 \\ 2 & -1 & 1 \\ -2 & 3 & 1 \end{bmatrix}$ $R2 = 2R1 - R2$ $\begin{bmatrix} 1 & -2 & -1 \\ 0 & -3 & -3 \\ -2 & 3 & 1 \end{bmatrix}$

$R3 = 2R1 + R3$ $\begin{bmatrix} 1 & -2 & -1 \\ 0 & -3 & -3 \\ 0 & -1 & -1 \end{bmatrix}$

not lin indep

2 lin indep vectors :

∴ b) or e)

2) Given $A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 1 & 0 \\ 2 & 1 & 3 \end{bmatrix}$, coloumn space has dimension of 2. Find vector not in null space of A.

In RREF,

$A: \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ Need to find vector that does not satisfy : $c1 \cdot a1 + c2 \cdot a2 = \vec{v}$

$c1 \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} + c2 \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} v1 \\ v2 \\ v3 \end{bmatrix}$

To find if it is in the coloumn space of A, the vector can be expressed as a lin combo. Not be in the coloumn space means, it must be orthogonal to all other coloumns.

3) Given $A = A^T$. Eigenvalues of A are $\lambda_1 = 5$, $\lambda_2 = 10$, eigenvector

$\vec{x}_1 = \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$ $\vec{x}_2 = \begin{bmatrix} -2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$

$A = PDP^{-1} \rightarrow A = V\Lambda V^T \longrightarrow$ spectral decomp

$A = \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{-2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}$

Finding Matrix given eigenvalue eigenvectors from $A = A^T$

$A = \begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$ D)

4) Given $M = \begin{bmatrix} 1+c & 1+c \\ 0 & 2 \end{bmatrix}$

If $c=0$, $\pi_1 = 1$, $\pi_2 = 2$

when $c=0$ matrix simplifies

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

$$\det(A - \pi I) = 0$$

$$= \begin{bmatrix} (1+c) - \pi & (1+c) \\ 0 & 2 - \pi \end{bmatrix}$$

$$((1+c) - \pi)(2 - \pi)$$

$$= (1+c-\pi)(2-\pi)$$

$$1+c-\pi = 0$$

$$1+c = \pi_1$$

$$2 = \pi_2$$

Comparing eigenvalues: only lambda 1 will change. b)

5) Given $A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 4 \\ 3 & -3 \end{bmatrix}$

$\pi_1 = 1$, $\pi_2 = 2$

$\pi_1 = -5$, $\pi_2 = 3$

$$A^T = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

$$B^T = \begin{bmatrix} 1 & 3 \\ 4 & -3 \end{bmatrix}$$

$$A^T B^T = \begin{bmatrix} 5 & 0 \\ 8 & -6 \end{bmatrix}$$

$$\det(A - \pi I)$$

$$= \begin{bmatrix} 5 - \pi & 0 \\ 8 & -6 - \pi \end{bmatrix}$$

$$\det = (5-\pi)(-6-\pi)$$

$$= 5 - \pi$$

$\pi_1 = 5$, $\pi_2 = -6$

$\therefore$ answer a)

6) $A = A^T$ Positive semi definite, eigen values are 0 and above

Positive definite, eigenvalues are above 1.

Practise Quiz #2

$$C = \begin{bmatrix} 2 & 2 & 1 \\ 0 & ? & ? \\ 0 & 0 & -1 \end{bmatrix}$$

Find scalar d, that produces matrix C.

A matrix is singular if its determinant is zero.

$\det(C) = 0$ $\longrightarrow$ Not invertible.

Upper triangle matrix, determinant is the product of the diagonal

$$\det(C) = 2 \cdot x \cdot (-1) = 0$$

$$d = 4$$

2) $A = \begin{bmatrix} 2 & 2 & 0 \\ 2 & -2 & -2 \\ 0 & 2 & 1 \end{bmatrix}$, $\vec{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

$$R = EA = \begin{bmatrix} 2 & 2 & 0 \\ 0 & -4 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$E = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ -1 & 1 & 0 \\ -0.5 & 0.5 & 1.0 \end{bmatrix}$$

What to study

- Practise 2
- Class notes 7 - 13
  - Pre real problems, HW, HWs

crossval( ) function in matlab gives the mean squared error. Not the root mean squared error (RMS)

Practise 2 - Part 1

- Predicting fragility index using age group proportions
- onesVector = ones (m,1) → creates a column of ones (to include an intercept in regression)
- Initializing vectors to hold rms and slope values.
- Performing linear regression for each variable.

```
A = [ data Matrix (:, i), onesVector];   Add intercept
w = A \ fragilityVector;
rmsvars (i) = rms( fragilityVector - A * w);   compute Rms error
Slopes(i) = w(1) ;
```

- Storing slopes helps us determine positive/ negative correlations.

- Simple linear regression: $y = mx + b$
- including intercept term, makes sense when you know for sure, it does not go through the origin.
- If you include an intercept term, add a coloumn of ones to data matrix

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ \vdots & \vdots \end{bmatrix} \longrightarrow \text{output} \qquad w = [m; b]$$

m: Slope, b: intercept

Cross Validation

- FragilityVector (dependent variable)
- dataMatrix (independent variable)
- Select coloumn, append a coloumn of ones for the intercept
- Define K = 5, crossvalind( ) Assigns each row to groups
- Train on 4/5 data, test on 1/5
- w = X_train \ y_train

Zero mean transformation: $x' = x - \bar{x}$

$y = Xw + \epsilon$ → residual error

↓ actual matrix of input data

↳ weight vector (regression coefficient)

$w = (X^T X)^{-1} X^T y$

SVD: is a matrix factorization techniave that decomposes a matrix A into 3: $A = U \Sigma V^T$

w tells us how much each input affects the output

w_1 is slope of regression line

Linear regression for $A\vec{w} \approx \vec{c}$ is projection of c to column space of A

↳ goal: to find weight vector $\vec{w}$ such that we get closest to the actual target/output values

Root mean square error

$$RMS(\vec{w}, A, \vec{c}) = \frac{\|\vec{e}(w)\|}{\sqrt{m}}$$

After finding best w to find how far predictions $A\vec{w}$ are from c

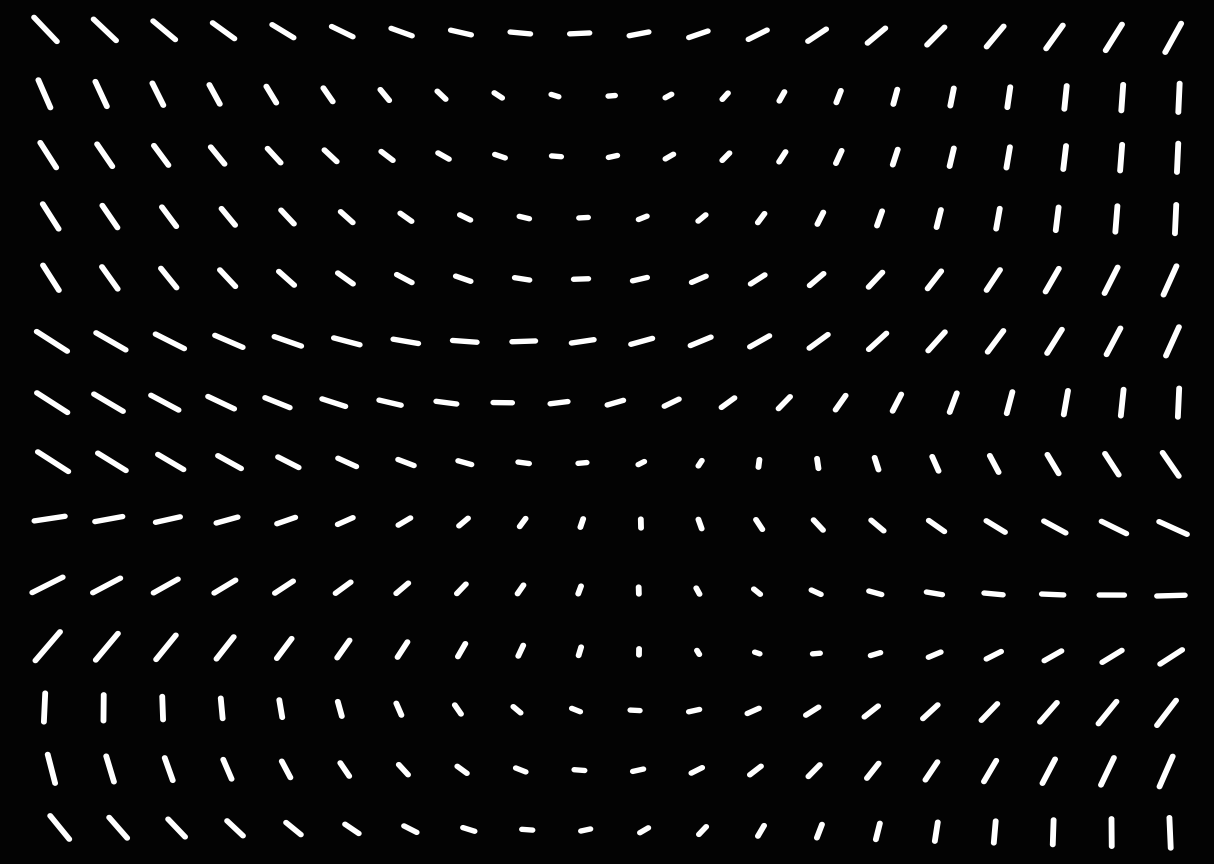
$\vec{e} = A\vec{w} - \vec{c}$ is error vector

$\sqrt{m}$, where m is number of data points

- Training and testing model on same data does not give good estimate of performance
  ↳ does not account for unseen data

K-Fold cross validation *

Unit 2



Quiz 2

- Error of projection: $e = C - P$
- Projection onto subspace: $P = Aw$
- Normal equation to find projection weights: $A^T Aw = A^T C$
- The projection matrix is , projects a vector onto the coloumn space of A.
- Root mean square

$$RMS(\vec{w}; A; \vec{c}) = \frac{\|\vec{e}\|}{\sqrt{m}}$$

- $A = U \Sigma V^T$

  U = orthogonal matrix (left singular vectors)

  Σ = diagonal matrix

  V^T = orthogonal matrix (right singular)
- The projection of b onto the space spanned by A.
- A matrix is singular if $\det(A) = 0$. A row/coloumn is linearly dependent
- The sum of eigenvalues = trace of A.
- Projection formula: $P = A(A^T A)^{-1} A^T b$

SVD - In Depth

- SVD is a process of breaking down any matrix into $A = U \Sigma V^T$

If $A \in \mathbb{R}^{m \times n}$

$A = U \Sigma V^T$

- U is $m \times m$ columns are left singular vectors, orthonormal
- Σ is $m \times n$ diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \sigma_r > 0$ (all positive)
- V is $n \times n$ columns are right singular vectors, orthonormal

---

Importance

$A = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}$

↑ ↑ ← Lin dep

Lin indep Lin indep

$U = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}$

↑ ↑ ↘ Orthogonal complement of col space

Left singular vectors of non zero singular values

Basis for col space

$\Sigma = \begin{bmatrix} \sigma_1 & x \\ x & \sigma_2 \end{bmatrix}$

$V^T = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}$

↘ Basis for null space

↙ Basis for row space of A

V Contains eigenvectors of A^TA

U Contains eigenvectors of $A \& A^T$

Matrices A^TA and AA^T are both Symmetric and positive semi-definite

↳ always have

To find eigenvalues of $A^TA = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 9 & 3 \\ 3 & 5 \end{bmatrix}$

$\begin{bmatrix} 9 & 3 \\ 3 & 5 \end{bmatrix}$ eigen values of this matrix are square roots of the singular values from Σ

Also Share same process for AA^T

Note: A^TA and AA^T have different dimensions.

Eigendecomp: $A = Q \Lambda Q^{-1}$

(only square, diagonalizable matrices)

↳ If Symmetric:

$A = Q \Lambda Q^T$ → Transpose of Q or Inverse

Same ← matrix whose cols are eigenvectors of A

↓ diagonal matrices of eigenvalues

Spectral Theorem:

- eigen vectors are orthogonal and orthonormal

HW-4

4.2) 1, 8, 9, 12, 20

1) $P = A\vec{w}$

a) $\vec{w} = \frac{A^T C}{A^T A}$

$b = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$, $a = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

$\vec{w} = \frac{[1\,1\,1]\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}}{[1\,1\,1]\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}$

$\vec{w} = \frac{5}{3}$

$P = \frac{5}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

$P = \begin{bmatrix} 5/3 \\ 5/3 \\ 5/3 \end{bmatrix}$

$\vec{e} = C - P$

$\vec{e} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 5/3 \\ 5/3 \\ 5/3 \end{bmatrix}$

$\vec{e} = \begin{bmatrix} -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$

8) $\vec{b} = (1, 1)$

$a_1 = (1, 0)$

$a_2 = (1, 2)$

$P_2 = \begin{bmatrix} 3/5 \\ 6/5 \end{bmatrix}$

$P_1 + P_2 \neq b$

$P_1 = \frac{a^T C}{a^T a}$

$P_1 = [1\ 0]$

9) $P = A(A^T A)^{-1} A^T$

$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, Find Projection matrix P

Projects vector $\vec{b}$ onto plane spanned by $\vec{a_1}$ and $\vec{a_2}$

$\text{Proj}_{\text{col}(A)}\,\vec{b} = P\vec{b}$

$P = I$ If A is invertible, it spans all of $\mathbb{R}^2$.

12) Project b onto column space of A

$A^T A x = A^T b$

$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} x = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$

$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} x = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$

$x = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$

$P = A\hat{x} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$

$\vec{e} = b - P$

$\vec{e} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$

$\vec{e} = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}$

$A^T\vec{e} = 0$ because b is in col space of A.

20) Projection matrix: $x - y - 2z = 0$

Two vectors in plane, plane will be column space of A

Plug in $\vec{v_1} = (1, 1, 0)$

$\vec{v_2} = (2, 0, 1)$

$A = \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$

$A^T A = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$

$= \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$

Inverse $\frac{1}{10 - 4} = \frac{1}{6}$

$P = \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \frac{1}{6}\begin{bmatrix} 5 & -2 \\ -2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix}$

$= \begin{bmatrix} 5/6 & 1/6 & 1/3 \\ 1/6 & 5/6 & -1/3 \\ 1/3 & -1/3 & 1/3 \end{bmatrix}$

For any basis vectors in plane $x - y - 2z = 0$

The matrix P projects any vector in $\mathbb{R}^3$

Test 2 - Prep

practise test - winter 2019

1.) Vector space V, spanned by basis vectors

What it means to project b into V.

- The goal of projecting a vector p is to find best approximation of b within V.

Finding p, such that $\|b-p\|$ euclidean norm is minimized. Equivalent to least squares solution.

2) Vector space V that is spanned by $A = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$

find error projection of $\vec{b} = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$ into vector space

$[A^TA]\,\vec{w} = A^TB$

$A\vec{w} = \vec{C}$

$A\vec{w} = p$

$B - p$

$\begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \vec{w} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$

$\begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix} \vec{w} = \begin{bmatrix} -28 \\ 14 \end{bmatrix}$

$p = Aw = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$C = b - p$

$C = \begin{bmatrix} -5 \\ 3 \\ 0 \\ -1 \end{bmatrix}$

3) $f(x) \approx c1\,x + c2$

for all points: $Aw = f$

- A is a matrix
- w is a vector with c1, c2

$A = \begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}$, $f = \begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix}$, $w = \begin{bmatrix} c1 \\ c2 \end{bmatrix}$

$\begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} c1 \\ c2 \end{bmatrix} = \begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix}$

$(A^TA)w = A^Tf$

$= \begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix}$, $A^Tf = \begin{bmatrix} 17.5 \\ 9.0 \end{bmatrix}$

$\begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} c1 \\ c2 \end{bmatrix} = \begin{bmatrix} 17.5 \\ 9.0 \end{bmatrix}$

$c1 = \frac{17.5}{10} = 1.75$

$c2 = \frac{9.0}{4} = 2.25$

6) $y \approx c1\,x + c2$

$M = \begin{bmatrix} x1 & y1 & 1 \\ x2 & y2 & 1 \\ x3 & y3 & 1 \end{bmatrix}$

$y = c1x_i + c2$

2) Normal Eq: $[A^TA]\vec{w} = A^TC$ 2 by 4 · 4 by 1 = 2 by 1

$A = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$ $\begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}\vec{w} = \begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix}\begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$ $\begin{matrix} -111 \\ -14 \\ -2 \\ -1 \end{matrix}$

$B = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$ $\begin{bmatrix} 70 \\ 07 \end{bmatrix}\vec{w} = \begin{bmatrix} -28 \\ 14 \end{bmatrix}$

$P = A\vec{w}$ 4 by 2 · 2 by 1 $\vec{w} = \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$\begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$P = \begin{bmatrix} -6 \\ -10 \\ -2 \\ 0 \end{bmatrix}$

$e = P - b$

$e = \begin{bmatrix} -6 \\ -10 \\ -2 \\ 0 \end{bmatrix} - \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$

$e = \begin{bmatrix} -5 \\ 3 \\ 0 \\ -1 \end{bmatrix}$

3) $f_i = C_1 x_i + C_2$

$Aw = f$

$A = \begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}$, $f = \begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix}$, $w = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$

Using normal Eq: $(A^TA)w = A^Tf$

$\begin{bmatrix} -2 & -1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} w = \begin{bmatrix} -2 & -1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix}$

2 by 4 · 4 by 2

$\begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix} w = \begin{bmatrix} 17.5 \\ 9.0 \end{bmatrix}$

$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} =$ $C_1 = 1.75$, $C_2 = 2.25$

4.2 Exercise 1

0) $P^2 = P$ is a projection matrix

$P^T = P$, is an orthogonal projection

The column space of $I - P$ has vectors projected onto compliment

1) Projection of b onto line through a,

$P = \frac{a^Tb}{a^Ta}a$

$c = b - P$

$P = \left(\frac{5}{3}\right)\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

$P = \begin{bmatrix} 5/3 \\ 5/3 \\ 5/3 \end{bmatrix}$

$e = b - P$

$e = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} - \begin{bmatrix} 5/3 \\ 5/3 \\ 5/3 \end{bmatrix}$

$c = 1$

8) $\vec{b} = (1,1)$

$a_1 = (1,0)$

$a_2 = (1,2)$

$P = \frac{ab^T}{a^Ta}a$ $P_2 = \frac{a_2b^T}{a_2^Ta_2}a_2$

$P_1 = \frac{b^Ta_1}{a_1^Ta_1}a_1$ $P_2 = \left(\frac{3}{5}, \frac{6}{5}\right)$

$P_1 = (1,0)$ $P_1 + P_2 = \left(\frac{8}{5}, \frac{6}{5}\right) \neq (1,1)$

9) $P = A(A^TA)^{-1}A^T$

$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$

$A^TA = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 5 \end{bmatrix}$

Find $(A^TA)^{-1}$

$\det(A^TA) = -4$

Inverse for 2 by 2

$= \frac{1}{-4}\begin{bmatrix} 5 & -3 \\ -3 & 1 \end{bmatrix}$

$= \begin{bmatrix} -5/4 & 3/4 \\ 3/4 & -1/4 \end{bmatrix}$

$A^T \cdot (A^TA)^{-1} = P = I$

12) $P = A(A^TA)^{-1}A^Tb$

$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $b = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$

$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$

2 by 3 · 3 by 2

$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

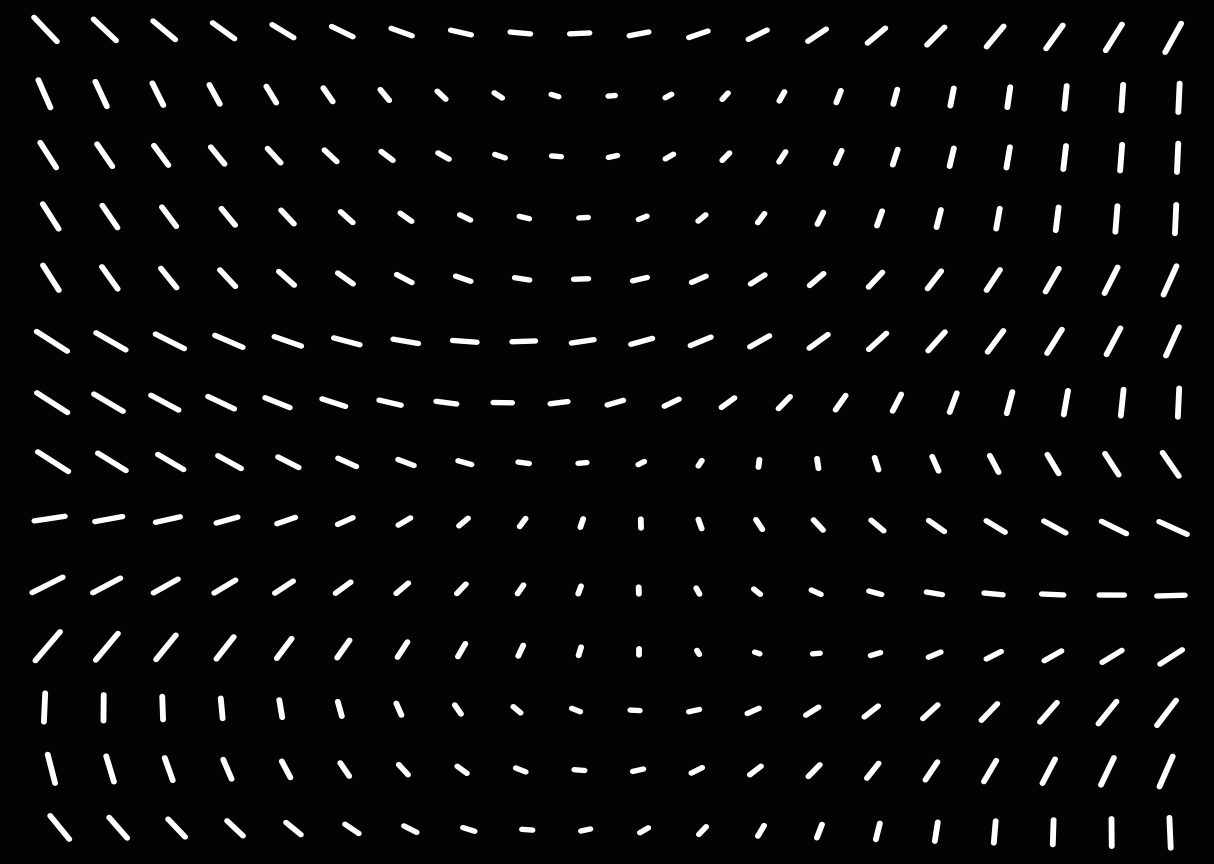
$-1 - 1 = -2$

$\frac{1}{-2}\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

$= \frac{1}{-2}\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$

$= \begin{bmatrix} -1 & \frac{1}{2} \\ \frac{1}{2} & -1 \end{bmatrix}$

Unit 3



Practise 3

- PCA Score: after finding principal components (eigenvector of covariance matrix) each data point is projected onto these new axes to obtain new coordinates.
- Eigen values of covariance matrix, represent the amt of variance captured by each principal component
- For PCA scores, the amt of dimensions your reducing to is the amt of coloumns
- DB index is used to assess effectiveness

---

1) Find best pair of features, for 2D rep
2) Apply PCA to reduce 13D wine set to 2D and score effectiveness
3) Standardize the data, apply PCA again

Proccess of Code

- read data, first row skipped, X mat contains features, yvec contains class labels.
- computes DB index, which helps evaluate quality of clustering.
- Lower Index = Better Clustering
- For loop iterates all possible pairs of data features.
- Data Plotting, according to two features
- Raw Data PCA: Function subtracts the mean from xmat.
  Performs PCA using SVD → zero mean data
- Standardized Data PCA: uses z score to normalize, performs PCA.
  ↓
  zero mean + standard deviations.

Analysis of Scores

- Smallest DB index - 0.7875 - Still good Cluster
- DB index of raw PCA scores - 1.5 - less good Cluster
- DB index of Standardized PCA - 0.6392 - best cluster

Quiz 3 Review

- SVD is used for matrix approximation

$$A = U \Sigma V^T$$

- U and V are orthogonal matrices
- Σ is a diagonal matrix with singular values
  - gives info abt each col of V and U

Matrix Approx

- $C_j = \sigma_j U_j V_j^T$
  - Each C_j is a rank 1 matrix
  - got by outer product of singular vector pairs
- $\sum_{j=1}^{p} C_j$ approximates A
  - left singular vectors are eigenvectors of AA^T → orthogonal directions row space
  - right singular vectors are eigenvectors of A^TA → orthogonal directions col space

$$C_1 = \sigma_1 U_1 V_1^T$$

$$A \approx C1 + C2$$

$$\therefore A \approx \begin{bmatrix} 3.74 & 4.32 \\ 4.32 & 6.26 \end{bmatrix}$$

→ approximates original matrix

- Matrix U-P provides basis for space that captures most variance in the data. If p is small compared to original vars. Up reduces dimensionality

- Scores in PCA, represent transformed coordinates of your data in the space of your principal components. $Zp = [\sigma_1 \vec{v_1}, \sigma_2 \vec{v_2} \dots]$ First few have most info and approx of data.

In Person Quiz Review

- given the SVD, explain properties
- SVD, PCA, DR

SVD

- Process that decomposes matrix A into 3 seperate matrices with usefun properties
- $A = U \Sigma V^T$
- Matrix U, orthogonal matrix (all coloumns dot product with each other is 0, and unit length)
- Matrix Σ, Diagonal matrix (singular values arranged in descending order, gives us weights of U)
- Matrix V^T, Transpose of orthogonal matrix.

$A = \begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix}$

$$\begin{bmatrix} 3 & 4 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{bmatrix}$$

↑ importance of coloumns

- U, V give you directions / orientations

Properties

- first r columns of U, span the column space of A
- first r columns of V, span row space of A
  or r rows of V^T
- Coloumn of U, that doesnt span column space spans othogonal compliment of the coloumn space.
- The coloumns of V that dont span the row space, span the null space.

Matrix Approx

- u_j, v_j are col vector, row vector respectively
- Each $C_j = \sigma_j u_j v_j^T$, where σ_j is a singular value
- rank 1 matrix, outer product, capturing patterns
- Original A: can be reconstructing by summing $A = C1 + C2 + C3 \dots Cp$
- Dimensiality Reduction of the data focuses on most important features

Example

$A = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix}$, $U = \begin{bmatrix} -0.924 & -0.383 \\ -0.383 & 0.924 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 5.398 & 0 \\ 0 & 0.926 \end{bmatrix}$, $V^T = \begin{bmatrix} -0.655 & -0.755 \\ 0.755 & -0.655 \end{bmatrix}$

$C1 = \sigma_1 u_1 v_1^T$

$C1 = 5.398 \begin{bmatrix} -0.924 \\ -0.383 \end{bmatrix} \begin{bmatrix} -0.655 & -0.755 \end{bmatrix}$

$C1 = \begin{bmatrix} 3.268 & 3.768 \\ 1.354 & 1.561 \end{bmatrix}$ $C1 + C2 = A$

$C2 = \begin{bmatrix} -0.268 & 0.232 \\ 0.646 & -0.561 \end{bmatrix}$

Example: 4 by 4 matrix, with singular values 5, 3, 1, 0.
We can approx $\hat{A} = 5u_1v_1^T + 3u_2v_2^T$

PRE PCA STUFF

- Zero mean Matrix M, Each entry is subtracted by mean of its col
  $(M = A - \underbrace{\frac{1}{m}\vec{1}\vec{1}^T}_{\text{mean of og coloum}})$

Using SVD: $M = U\Sigma V^T$

Up forms orthonormal basis for Principal Components

$Up = [u_1, u_2]$. This is used to project og data into new space defined by principal components

EX

$A = \begin{bmatrix} 1 & 4 \\ 3 & 8 \\ 5 & 2 \end{bmatrix}$

mean matrix: $\frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 4 \\ 3 & 8 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 14/3 \\ 3 & 14/3 \\ 3 & 14/3 \end{bmatrix}$

A - mean matrix = Zero mean matrix

$= \begin{bmatrix} -2 & 2/3 \\ 0 & 10/3 \\ 2 & -12/3 \end{bmatrix}$

Mean matrix: $\frac{1}{m}\vec{1}\vec{1}^T A$

PCA

What is PCA?

- Dimensiality reduction
- Used to find main patterns of variation in data
- New set of orthogonal axes that capture most of variance
- First principal component captures most variance
- Transforms data to a lower dimensional space

M is the zero mean matrix

$B = \frac{M^T M}{m-1}$, covariance matrix

PCA via SVD

- 1) Compute the covariance matrix → Covariance Matrix should tell you how much each column varies.

$B = \frac{1}{m-1} M^T M$ — B is covariance captures relationships

M is zero mean matrix

$B \approx \begin{bmatrix} 21 & 17.5 & 27.6 \\ 17.5 & 19 & 25.5 \\ 27.5 & 25.5 & 36.3 \end{bmatrix}$ □ is variance of the co

Must be B positive semi definite

$B = E \Lambda E^T$

2) Compute eigenvalues and eigenvectors

- eigenvectors of B are the principal components
- eigenvalues tell us how much variance each component captures

$\lambda = \begin{bmatrix} 75.60 \\ 4.82 \\ 0.17 \end{bmatrix}$ → first principal component captures 75.60%

3) Compute the SVD

- $M = U \Sigma V^T$
  - U contains left singular vectors, Σ is diagonal matrix, V contains right singular vectors
  - right singular values V are same as eigenvectors of B: $E = V$
  - singular values σ_j relate to eigenvalues: $\lambda_j = \frac{\sigma_j^2}{m-1}$

$B = V \Lambda V^T$ → eigenvectors transposed

- B: covariance matrix
- V: eigen vectors of B
- Λ: eigenvalues diagonal

Computing PCA Scores

If Symmetric matrix, Singular values are the eigenvalues

- after finding principal components

PCA scores: $Z = MV = U\Sigma$

- Z is score matrix
- Each row in Z represents data projected on new PCA basis

Ex

$$Z_1 = \begin{bmatrix} -0.67 & 1.75 \\ 12.80 & -1.84 \\ -0.5 & 2.76 \\ 0.06 & 0.40 \\ -11.64 & -2.26 \end{bmatrix}$$

→ First col reps scores along first principal component

→ Strong patterns in first component

- Z represents projection of each observation on direction of max variance. Each number is how much along the direction it is

---

Homework 7

$$A = \begin{bmatrix} 0 & -2 & 2 \\ -3 & 0 & 0 \\ 3 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

i) Eigenvalues of A^TA

$$A^TA = \begin{bmatrix} 0 & -3 & 3 & 0 \\ -2 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -2 & 2 \\ -3 & 0 & 0 \\ 3 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

3x4 by 4 by 3

$$= \begin{bmatrix} 18 & 0 & 0 \\ 0 & 5 & -5 \\ 0 & -5 & 5 \end{bmatrix}$$

$$\lambda = [18 \quad 8 \quad 2]$$

ii) Eigenvalues of AA^T

$$AA^T = \begin{bmatrix} 0 & -2 & 2 \\ -3 & 0 & 0 \\ 3 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & -3 & 3 & 0 \\ -2 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 \end{bmatrix}$$

4 by 3 x 3 x 4

$$AA^T = \begin{bmatrix} 8 & 0 & 0 & 4 \\ 0 & 9 & -9 & 0 \\ 0 & -9 & 9 & 0 \\ 4 & 0 & 0 & 2 \end{bmatrix}$$

$$\lambda = [18 \quad 8 \quad 2 \quad 0]$$

iii) $\vec{\sigma} = [\sqrt{18} \quad \sqrt{8} \quad \sqrt{2}]$

Singular values are square roots of eigen values of A^TA and AA^T

- If symmetric and positive semi definite

iv) $A = U\Sigma V^T$

$$U = \begin{bmatrix} 0 & -1 & 0 & -0 \\ 0.7 & -0 & 0 & 0.70 \\ -0.7 & 0 & 0 & 0.70 \\ 0 & 0 & 1.0 & 0 \end{bmatrix}$$

→ columns pace of thogonal complement

$$V^T = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0.7071 & 0.7071 \\ 0 & -0.7 & 0.7071 \end{bmatrix}$$

→ Row span of A

→ null space is 0? $n = \{\}$

$$S = \begin{bmatrix} 4.24 & 0 & 0 \\ 0 & 2.82 & 0 \\ 0 & 0 & 1.41 \\ 0 & 0 & 0 \end{bmatrix}$$

Col space of A because 3 Singular values Positive definite

Question 2)

$$A = \begin{bmatrix} 10 & 3 & 8 \\ 6 & 10 & 2 \\ 10 & 1 & 9 \\ 8 & 3 & 10 \\ 7 & 5 & 2 \end{bmatrix} \quad \begin{matrix} 8.2 & 4.4 & 6.2 \end{matrix}$$

$M = A - \bar{A} = U \Sigma V^T$

σ_j corresponding covariance

eigen value is $\pi_j = \frac{\sigma_j^2}{(m-1)}$

Explained variance is like $\begin{bmatrix} 25.2 & & \\ & 16.3 & \\ & & 7.1 \end{bmatrix}$

$$\frac{25.2}{16.3 + 7.1 + 25.2} = 74\% \text{ or } 0.74$$

viii) To find Singular values, convert to zero mean form, then SVD

$$M = \begin{bmatrix} 1.8 & -1.4 & -\frac{21}{5} \\ -11.5 & 6.6 & \frac{14}{5} \\ \frac{9}{5} & -3.4 & \frac{9}{5} \\ -0.2 & -\frac{7}{5} & \frac{9}{5} \\ -1.2 & -0.6 & -\frac{21}{5} \end{bmatrix} \quad \begin{matrix} \frac{9}{5} \end{matrix} \qquad SVD(M)$$

σ_j is $= \begin{bmatrix} 10.329 & 3.254 & 1.578 \end{bmatrix}$

viii) $B = \frac{M^T M}{m-1}$

$$B = \frac{\begin{bmatrix} 1.8 & -2.2 & 1.8 & -0.2 & -1.2 \\ -1.4 & 5.6 & -3.4 & -1.4 & 0.6 \\ 1.8 & -4.2 & 2.8 & 3.8 & -4.2 \end{bmatrix} \begin{bmatrix} 1.8 & -1.4 & 1.8 \\ -2.2 & 5.6 & -4.2 \\ 1.8 & -3.4 & 2.8 \\ -0.2 & -1.4 & 3.8 \\ -1.2 & 0.6 & -4.2 \end{bmatrix}}{(3-1)}$$

$$B = \begin{bmatrix} 3.2 & -5.35 & 5.45 \\ -5.35 & 11.8 & -10.85 \\ 5.45 & -10.85 & 15.2 \end{bmatrix}$$

x) Explained variance of each PC is determined by eigen values of covariance matrix

Eigen Values of Co variances: 26.93, 2.64, 0.63

$\det(A - \pi I) = 0$

$\pi_j = \begin{bmatrix} 0.663 & 2.64 & 26.93 \end{bmatrix}$

Explained Variance: $\frac{26.93}{(26.93 + 2.64 + 0.63)} = 0.89$

$\vec{c} = \begin{bmatrix} 0.892 & 0.979 & 1.000 \end{bmatrix}$

For $\theta = 85\%$, $P = 1$

For $\theta = 95\%$, $P = 2$

xiii) $Z = MV = U\Sigma$

$M = U \Sigma V^T$

$$Z = \begin{bmatrix} -2.718 \\ 7.187 \\ -4.68 \\ -3.5 \\ 3.76 \end{bmatrix}$$

1) $X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$, $B = \frac{M^T M}{m-1}$

$\begin{matrix} 3 & 4 \end{matrix}$

$$M = \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 2 & 2 \end{bmatrix} \qquad B = \frac{\begin{bmatrix} -2 & 0 & 2 \\ -2 & 0 & 2 \end{bmatrix} \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}}{2}$$

$$B = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

HW # 7 - Again!

$$A_3 = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & -1 & 3 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.707 & 0.707 \\ 0 & 0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.7 & 0.7 \\ 0 & 0.7 & 0.7 \end{bmatrix}$$

Part 1:

1) Eigenvalues of A^TA are square of singular values

$\pi_1 = 36, \pi_2 = 16, \pi_3 = 4$

2) Eigenvalues of AA^T are also

$\pi_1 = 36, \pi_2 = 16, \pi_3 = 4$

3) $\sigma_1 = 6, \sigma_2 = 4, \sigma_3 = 2$

4) Column Span: $V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.707 & 0.707 \\ 0 & 0.707 & 0.707 \end{bmatrix}$

5) Column orthogonal span is none

6) Row span is: $\begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.7 & 0.7 \\ 0 & 0.7 & 0.7 \end{bmatrix}$

7) Null space: None

Part 2:

$$A_6 = \begin{bmatrix} 6 & 5 & 1 \\ 1 & 10 & 7 \\ 4 & 8 & 2 \\ 1 & 9 & 8 \\ 10 & 2 & 5 \\ 4.4 & 6.8 & 4.6 \end{bmatrix} \quad M = \begin{bmatrix} 1.6 & -1.8 & -3.6 \\ -3.4 & 3.2 & 2.4 \\ -0.4 & 1.2 & -2.6 \\ -3.4 & 2.2 & 3.6 \\ 5.6 & -4.8 & 0.4 \end{bmatrix}$$

$$B = \frac{M^TM}{m-1} \longrightarrow B = \begin{bmatrix} 14.3 & -12.1 & -5.5 \\ -12.1 & 10.7 & 4.1 \\ -5.5 & 4.1 & 9.3 \end{bmatrix}$$

$$\pi_1 = \frac{(10.477)^2}{4} \qquad \pi_2 = \frac{(5.174)^2}{4}$$

Eigenvalues of covariance is $\frac{\sigma^2}{m-1}$

$$\pi = \begin{bmatrix} 27.42 \\ 6.69 \\ 0.195 \end{bmatrix} \longrightarrow \text{Explained Variance} \qquad [0.799, 0.995, 1]$$

$p = 2$

$Z = MV$

In covariance matrix, you already have eigenval

In SVD, you need to square explained var

Part 3:

$\text{rank}(A) \leq \min(m, n)$

$\text{rank}(A) = r$, theres r singular values, rest are 0

$\text{rank}(A) = r$, r non zero eigenvalues

A^TA and AA^T have same eigenvalues

SVD - Sign ambiguity and can be rotated

Homework 7 / Practise

- The rank of matrix A is always less than or equal to the min number of rows. $\text{rank}(A) \leq \min(m, n)$

- If $r = \text{rank}(A)$, the singular values corresponding to rank of A are positive.

  - If $m > n$ λ_i of $A^T A$ and AA^T
    - There are n eigenvalues

    matrices $A^T A$ and AA^T share the same non zero eigenvalues

  - If $m > n$, $r = \text{rank}(A) < n$, for certain we can say about λ_i. There are r nonzero eigenvalues λ of $A^T A$ and AA^T

Unit 3 - Exam Practise Questions

5) $\frac{33.61^2}{1259.57} = 0.89$ → aq: $\therefore 2$

1) $A = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 2 \end{bmatrix}$, find error projections of the vector $\vec{b} = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$

$P = A\vec{w}$

$\vec{e} = \vec{c} - \vec{p}$

$P = \frac{a^T c}{a^T a}$

$P = \begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$ / $\begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$

$P = \begin{bmatrix} -28 \\ 14 \end{bmatrix} / \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}$ $\quad \vec{w} = \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$P = \underset{4\text{ by }2}{\begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}} \underset{2\text{ by }1}{\begin{bmatrix} -4 \\ 2 \end{bmatrix}}$

$P = \begin{bmatrix} -6 \\ -10 \\ -2 \\ 0 \end{bmatrix}$

$\vec{e} = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix} - \begin{bmatrix} -6 \\ -10 \\ -2 \\ 0 \end{bmatrix}$

$\vec{e} = \begin{bmatrix} -5 \\ 3 \\ 6 \\ -1 \end{bmatrix}$

2) $f(x) = c_1 x + c_2$

$\begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix} \simeq c_1 \begin{bmatrix} -2 \\ -1 \\ 1 \\ 2 \end{bmatrix} + c_2$

f is outputs, x is inputs

$f \approx x \cdot c$

$X = \begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}$, $F = \begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix}$

$A^T A \vec{w} = A^T C$

$x^T x C = x^T f$

$x^T x = \begin{bmatrix} -2 & -1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -2 & 1 \\ -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix}$

$x^T f = \begin{bmatrix} -2 & -1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1.6 \\ 1.3 \\ 3.0 \\ 6.3 \end{bmatrix} = \begin{bmatrix} 17.5 \\ 9.0 \end{bmatrix}$

$C = \begin{bmatrix} 17.5 \\ 9.0 \end{bmatrix} / \begin{bmatrix} 10 & 0 \\ 0 & 4 \end{bmatrix}$

$C = \begin{bmatrix} 1.75 \\ 2.25 \end{bmatrix}$

1) Find Singular values of A and Transpose Products find eigenvalues:

$A = U \Sigma V^T$

Square roots of eigen values of $A^T A$.

$\begin{bmatrix} 5-\lambda & 4 \\ 4 & 5-\lambda \end{bmatrix}$

$(5-\lambda)(5-\lambda) - 16$

$25 - 10\lambda + \lambda^2 - 16$

$\lambda^2 - 10\lambda - 9$

$(\lambda - 9)(\lambda - 1)$

$\lambda = 9$, $\lambda = 1$

Square roots are 3, 1

2) Given $A = U \Sigma V^T$

$A = \begin{bmatrix} 2 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 0 \end{bmatrix}$ rank (A) = 2

$A^T A = [\quad]$

$A A^T = [\quad]$

Columns of U_2 form orthonormal basis column space

make A unit magnitude

$A^T A$ - Span row space

4) $Z = MV$

$Z = \underset{4\text{ by }2}{\begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}} \underset{2\text{ by }2}{\begin{bmatrix} -0.71 & 0.71 \\ -0.71 & -0.71 \end{bmatrix}}$

$Z = \begin{bmatrix} 0 \\ 0 \\ -2.84 \\ 2.84 \end{bmatrix}$

HW - 5

7.2 - 3
4

3) given $A = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 2 & 4 & 2 & 2 \\ 1 & 2 & 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix}$

both are rank 2, asked to express as sum of 2 rank 1 matrices

$$A = u_1 v_1^T + u_2 v_2^T$$

$$A = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & -2 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

$$BB^T = \underset{2 \text{ by } 3}{\begin{bmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \end{bmatrix}} \underset{3 \text{ by } 2}{\begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 2 & 3 \end{bmatrix}}$$

$$BB^T = \begin{bmatrix} 9 & 13 \\ 13 & 19 \end{bmatrix}$$

Trace: $9 + 19 = 28$

Det: $9 \cdot 19 - 13^2 = 2$

$\sigma_1^2 = 28$
$\sigma_2^2 = 2$

$\therefore$ Yes it is compressible
first component gives
good approx.

Singular values: $\sigma_1 = \sqrt{\lambda_1}$

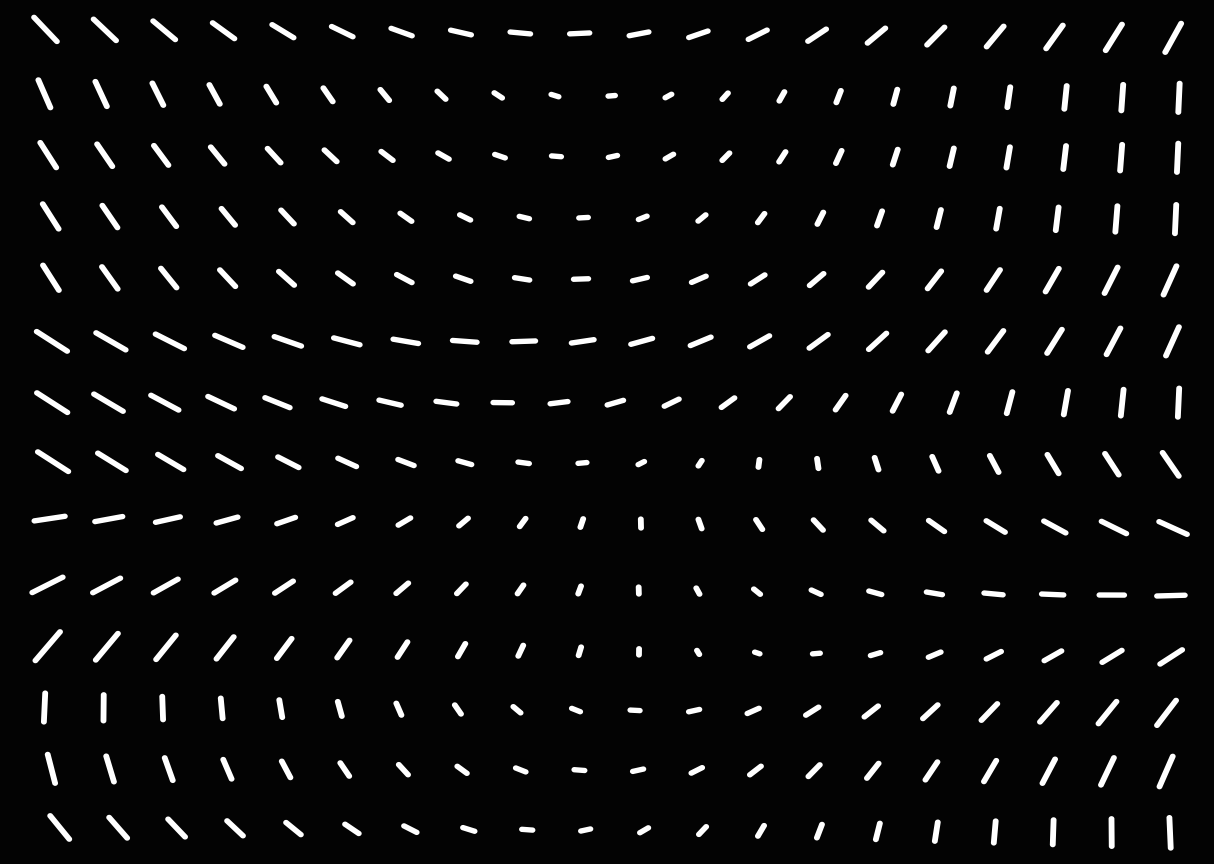
λ_1 are eigenvalues of B^TB or BB^T

$$\sigma_1^2 + \sigma_2^2 = tr(B^TB)$$
$$\sigma_1^2 \cdot \sigma_2^2 = \det(B^TB)$$

Polar Decomp:

$$A = QS$$
$$Q = UV^T$$
$$S = V \Sigma V^T$$
$$A = U \Sigma V^T = (UV^T)(V \Sigma V^T) = QS$$

Unit 4



Quiz 4 Review

K-Means Clustering

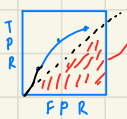
- Clusters seperated by hyperplane
- normal vector $\vec{w} = \vec{g_1} - \vec{g_2}$, Point is $\vec{p} = \frac{\vec{g_1} + \vec{g_2}}{2}$
- bias scalar is $b = -w^T \vec{p}$
- prefer unit normal $\vec{n}$ and bias scalar c, $\vec{n} = \frac{\vec{w}}{\|\vec{w}\|}$, $c = \frac{b}{\|\vec{w}\|}$

Classification

- given labels ± 1, scores $z \in R$
- To give a class, you get a z score and threshold θ

$$\begin{bmatrix} TPR & FNR \\ FPR & TNR \end{bmatrix}$$ → FNR = 1 - TPR

- 2 Degrees of freedom gives an ROC vector (FPR, TPR)

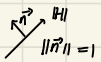


50% performance

AUC is under area ROC curve
- Bigger is better

odds of Occurrence

- Probability p



- odds of occurrence is $s = \frac{p}{1-p}$
- for hyperplane |H|, unit normal $\vec{n}$, Scalar c:
  Signed distance of $\vec{x}$ to |H| is $d = n^T \vec{x} + c$
  - If distance is positive, its on +1 side
  - In distance is negative, its on -1 side
- Probability x is Class 1 is $P = \frac{1}{1+e^{-d}}$

- Should be able to:
  - make hyperplane for binary data
  - Compute Scores of data from hyperplane
  - implement computation of ROC curve (TPR vs FPR)

Given Partitions, Compute Centroids
With Centroids, Can Compute distance to each point

1) Initialize Centroids
2) Create partitions based on distance to centroids
3) Update Centroids based on points in partition
4) With new centroids, repeat process
5) Converge, if nothing changes

Practise 4

Part 1

- general form of linear classifier is: $w_1 x_1 + w_2 x_2 + b = 0$
- where $w = (w_1, w_2)$ is weight vector
- b is bias term, which shifts the hyperplane
- $w = \mu_1 - \mu_0$, then transpose for w^T
- bias term (b): $b = -\frac{1}{2}(\mu_1 + \mu_0) \cdot w$

  after normalization: $\vec{n} = \frac{\vec{w}}{\|w\|}$, $b = \frac{b}{\|w\|}$

$w = [w; b]$

Part 2

- Compute score z for each sample: $z = X_{aug} \cdot w_{aug}$

  X_{aug} is data matrix with bias term

  w_{aug} is weight vector with bias

- ROC curve: TPR vs FPR
- perfcurve( ) generates ROC curve automatically

Part 3

- Finding best threshold for maximum accuracy
  - loop through all possible thresholds t from roc$T$

    $q = (z \geq T)$

$$\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$$

  - Compute accuracy $= \frac{TP + TN}{\text{All samples}}$

- binary classifier $f(x) = w^T x + C \rightarrow w_1 x_1 + w_2 x_2 + b = 0$
- $P(y=1|x) = \frac{1}{1 + e^{-(w^T x + C)}}$
- w determines direction of hyperplane
- distance of a point x from the hyperplane

$$d = \frac{|w^T x + C|}{\|w\|}$$

- $w^T x + C > 0$, point is on one side
- $w^T x + C < 0$, point is on other side

Quiz 4 - In person Test

K-means

- Each cluster has a centroid
- iterates between adjusting points to nearest centroids
- updating centroid as mean of points
- Binary case $K=2$
- normal vector: $w = g_1 - g_2$
- Midpoint: $p = \frac{g_1 + g_2}{2}$
- bias: $b = -w^T p$

Hyperplane: $w^T x + b = 0$

- w is normal vector to hyperplane
- b is bias scalar

normal vector is $w = g_1 - g_2$

Classification

- Accuracy = $\frac{TP + TN}{Total}$
- Signed distance: $n^T x + c$
- Label $y_i = \pm 1$, Each data point is $+1, -1$ class
- Scores z to each data point
- threshold θ. Decision boundary applied to score data

$$\begin{bmatrix} TPR & FNR \\ FPR & TNR \end{bmatrix}$$

- ROC curve plots TPR vs FPR at different theta θ values
- varying θ generates diff confusion matrices
- AUC — 1.0 perfect, 0.5 guessing.
- threshold θ, value determines how classifier assigns labels based on score
- TPR - (Recall)
- TNR - (specificity)

Odds/Probabilities

- odds $S = \frac{p}{1-p}$

if $S = 1$, equally likely odds
if $S > 1$, favourable
if $S < 1$, less favourable

- $d = n^T x + c$
- n is unit normal vector
- farther d is from 0, the more confidently we classify x.

$p = \frac{S}{1+S}$

$S = \frac{p}{1-p}$

$$p = \frac{1}{1+e^{-d}}$$

↓ Probability of being positive

Unit 4- Exam Prep

1) $x_1=\begin{bmatrix}1\\4\end{bmatrix}$, $x_2=\begin{bmatrix}4\\4\end{bmatrix}$, $x_3=\begin{bmatrix}3\\2\end{bmatrix}$, $x_4=\begin{bmatrix}0\\5\end{bmatrix}$

Find binary clusters, using euclidean distance

x_1 to x_2 = 3 ∴ $\{x_1, x_4\}$, $\{x_2, x_3\}$

x_1 to x_3 = $2\sqrt{2}$

x_1 to x_4 = $\sqrt{2}$

2) given $g_1=\begin{bmatrix}1\\2\\1\end{bmatrix}$, $g_2=\begin{bmatrix}3\\1\\3\end{bmatrix}$, $\vec{n}=\begin{bmatrix}0.667\\-0.333\\0.667\end{bmatrix}$

Find c2 = ?

$b=-\begin{bmatrix}4\\3\\4\end{bmatrix}\begin{bmatrix}-2 & 1 & -2\end{bmatrix}$

$h=\begin{bmatrix}2\\1.5\\2\end{bmatrix}$ $b=-\begin{bmatrix}2\\1.5\\2\end{bmatrix}\begin{bmatrix}0.667 & -0.333 & 0.667\end{bmatrix}$

$c=-2.1685$

3) Consider H, normal vector

$\vec{n}=\begin{bmatrix}-0.667\\0.333\\0.667\end{bmatrix}$, $c=1$

Classify data vectors $x_1=\begin{bmatrix}0\\2\\5\end{bmatrix}$, $x_2=\begin{bmatrix}4\\0\\1\end{bmatrix}$

$d_1=\begin{bmatrix}0\\2\\5\end{bmatrix}\begin{bmatrix}-0.667 & 0.333 & 0.667\end{bmatrix}+1$

$d_1=5.001$

$d_2=\begin{bmatrix}4\\0\\1\end{bmatrix}\begin{bmatrix}-0.667 & 0.33 & 0.667\end{bmatrix}+1$

$d_2=-2.0+1$

$d_2=-1$

x_1 is pos, x_2 is neg

4) given $m=\begin{bmatrix}-4\\3\end{bmatrix}$, $b=7$

Find the probability $x=\begin{bmatrix}2\\2\end{bmatrix}$ is Pos?

$\vec{n}=\begin{bmatrix}-4/5\\3/5\end{bmatrix}$

Also need to normalize bias scalar

$d=\begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}-4/5 & 3/5\end{bmatrix}+\frac{7}{5}$

$d=1$

$=\frac{1}{1+e^{-1}}=0.73$

HW #9

Part 1:

1) $g_2=[8\ 4]$ $P_2:\{(7,3),(8,4),(6,4)\}$

$g_1=[2\ 8]$ $P_1:\{(4,7)(2,8),(5,8)\}$

$P_1=\begin{bmatrix}4&7\\2&8\\5&8\end{bmatrix}$, $P_2=\begin{bmatrix}8&4\\6&4\\7&3\end{bmatrix}$

$g_1=(5\ 8)$ $g_2=(6\ 4)$

3) $\vec{w}=(5,8)-(6,4)$

a) $\vec{w}=(-1,4)$

midpoint = $(5.5, 6)$

$b=\begin{bmatrix}-1\\4\end{bmatrix}[5.5,6]$

$b=18.5$

b) $w^Tx+b=0$

$H=\begin{bmatrix}1\\-4\\18.5\end{bmatrix}$

c) Normalized: $\begin{bmatrix}-0.24\\0.97\\4.486\end{bmatrix}$

Part 2:

$\hat{w}=\begin{bmatrix}\vec{w}\\b\end{bmatrix}=\begin{bmatrix}2\\-2\\6\end{bmatrix}$

$A_2=\begin{bmatrix}0&4&1\\3&3&1\\1&3&1\\-1&2&1\\3&5&-1\\3&7&-1\\4&4&-1\\0&8&-1\end{bmatrix}$ y_i

1) $w^Tx+b=0$

$q_i=\text{sign}(w\cdot x_i+b)$

$2x_1+2x_2+6$

$q_1=\begin{bmatrix}2&-2\end{bmatrix}\begin{bmatrix}0\\4\end{bmatrix}$

$=-1$

$q_2=[2\ -2]\begin{bmatrix}3\\3\end{bmatrix}+6$

$=+1$

$q_3=[2\ -2]\begin{bmatrix}1\\3\end{bmatrix}+6$

$=+1$

$q_6=[2\ -2]\begin{bmatrix}3\\7\end{bmatrix}+6$

$=-1$

$q_4=[2\ -2]\begin{bmatrix}-1\\2\end{bmatrix}+6$

$=0$, on hyperplane +1

$q_5=[2\ -2]\begin{bmatrix}3\\5\end{bmatrix}+6$

$=+1$

$q_7=[2\ -2]\begin{bmatrix}4\\4\end{bmatrix}$

$=+1$

$q_8=[2\ -2]\begin{bmatrix}0\\8\end{bmatrix}+6$

$=-1$

$q=\begin{bmatrix}-1\\1\\1\\1\\-1\\1\\-1\end{bmatrix}$ $y=\begin{bmatrix}1\\1\\1\\1\\-1\\-1\\-1\\-1\end{bmatrix}$

TP	FN
FP	TN

3	2
1	2

2) $w=\begin{bmatrix}2\\-2\end{bmatrix}$, $b=6$

$d_1=\begin{bmatrix}2\\-2\end{bmatrix}[0\ 4]+6$

$=-2$

$d_5=[2\ -2]\begin{bmatrix}3\\5\end{bmatrix}+6$

$=\frac{2}{2\sqrt{2}}=0.71$

$\theta=-0.25$

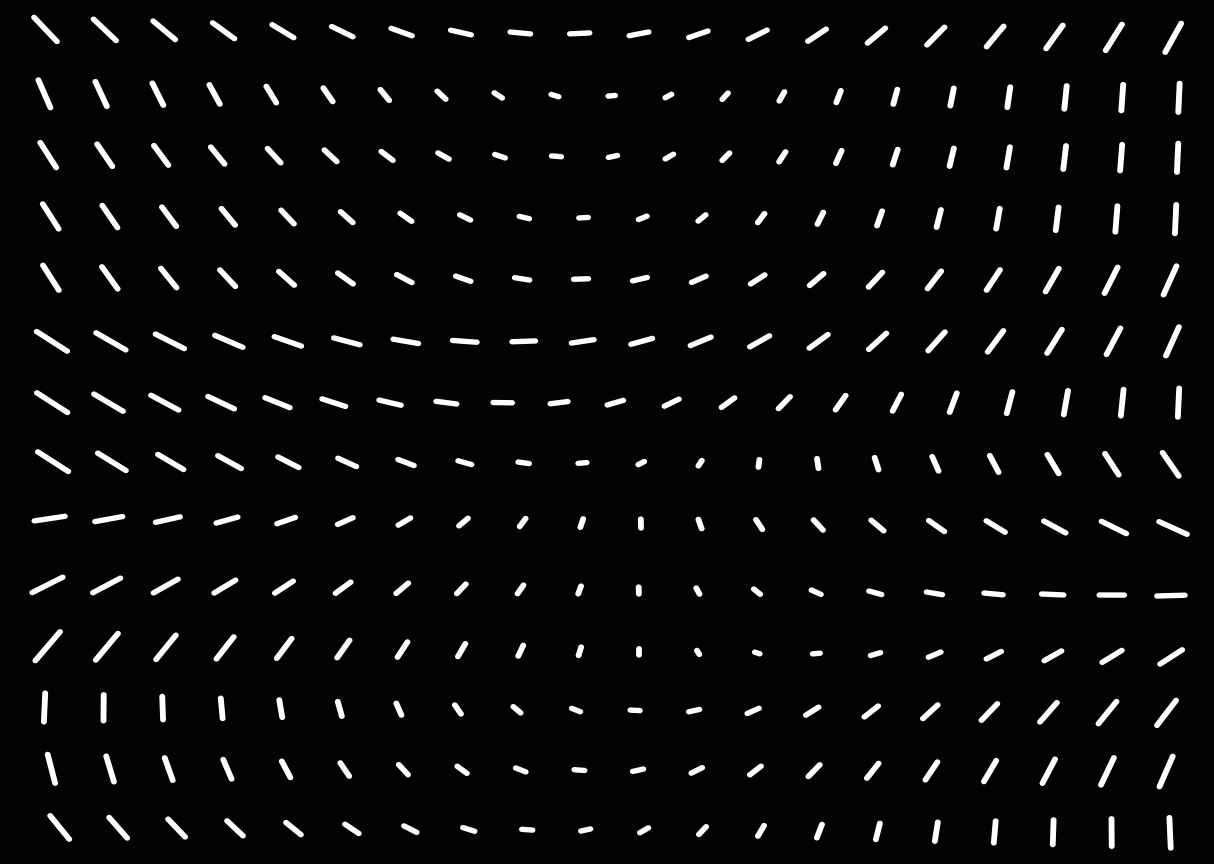
y: -1 -1 1 1 -1 1 -1 1 1 1

q: -1 -1 -1 -1 -1 -1 1 1 1 1

TP	FN
FP	TN

5	1
3	3

Unit 5



Unit 5 - In Person Quiz Prep

- Lectures 25-32 inclusive
- Practise 5/ Hw 11

Logistic Functions

- maps hyperplane distance to a value between 0 and 1, which is probability it's class 1 (+)

$$\phi(v) = \frac{1}{1+e^{-v}}$$

- where v is the result of lin combo
- For single input log regression $v = x \cdot w$
- For multi dimensional input, $v = w^T x$
- If data point is far from seperating hyperplane, function gives value near 0 or 1.
- Can be used as an activation function
- Helps us find best hyper plane → $\psi(v) = \phi(v)(1-\phi(v))$, used in gradient desc
- $v = w^T x + b$, raw score or distance to hyperplane.
  - If data has $\vec{1}$ vector at the end, use $w^T x$, if not use $w^T x + b$

Artificial Neurons

- inputs multiplied by weights, then summed, then check if greater than threshold, if yes shoot
- Linear response: $v = xw$ (raw score before activation)
- Logistic activation/score $z_i = \phi(v_i)$ (gets probability)
- Residual Error: $r = y_i - z_i$ (Prediction − Score) (guides weight updates)
- Squared error objective function: $f(\vec{r}) = \frac{1}{2} r^T r$ (cost function, how bad predictions are)

- Back propogation: $b_i = r_i \cdot \phi'(v_i)$ (captures how sensitive output is)
  - ↳ $\phi'(v_i) = \phi(v)(1-\phi(v_i))$, → Long version: $(y_i - \phi(v))\,\phi(v_i)(1-\phi(v_i))$
- Gradient vector $\nabla f_i = -b_i \vec{x_i}$, opposite direction of increase (gives a vector)
- Direction where we update weights: $\vec{d_i} = -[\nabla f_i]^T = b_i \vec{x_i}^T$
  - Descent Term
  - ↳ use this in $\vec{w}_{k+1} = w_k + \eta \vec{J}$ → descent term

- Descent vector: $\vec{d} = X^T \vec{b}$

$$\vec{J} = \sum_{i=1}^{m} \vec{d_i}$$

For single sample: $d_i = b_i x_i^T$

For all samples: $\vec{J} = X^T b$

Unit 6 - In Person Quiz Prep

Learning Steepest Descent

- Design matrix, each input is row x_i
- Vector of all y labels
- Initial weights $\vec{w}_0$ are random
- Learning rate $\eta > 0$, controls how fast you reach minimum / step size
  - ↳ gets multiplied by descent direction
- Update: $w_{K+1} = \vec{w_K} + \eta \vec{d}$ (moving toward better weights)
- repeat till max iteration.

Embedding and Kernel PCA

- deals with data thats not linearly seperable
- Map to higher dimension, do PCA there, then come back down
- Compares observations, not variables
- Kernel Functions:

  Quadratic: $K(\vec{u}, \vec{v}) = (u^T v + c)^2$

  Gaussian: $K(\vec{u}, \vec{v}) = e^{-r\|\vec{u} - \vec{v}\|^2}$

  ↳ points close togeher near 1, high similarity

- Pairwise dot products is gram matrix, Symmetric, Positive semi def
- Classic PCA, on matrix A, Using covariance matrix M.

Kernel PCA - Compare observations, not variables

Regular PCA - Compare variables, not observations

Practise 6

- 2 main tasks: the percepatron rule using steepest descent and Kernel PCA with Clustering

Part 1: Percepatron using Steepest Descent

- train neuron to classify colleges as public/private
- Perceptron is super simple artificial neurons
  - takes some input vectors
  - multiply by weights
  - adds them up
  - applies threshold
  - spits out predictions

$y = \text{sign}(w \cdot x + b)$

Steepest Desent, gradient descent, minimize error

$w = w - eta * gradient$

Bigger learning rate, bigger steps, risk of passing over minimum

lower learning rate, takes too long.

Overview flow

- calls PSa1
- calls Sepbinary (Perceptron training)

1) Augment data `xaug = [xmat ones(size(xmat,1),1)];`

2) set parameters Call Sepbinary

```
eta = 0.01
imax = 50000;
[w_ann, ix] = Sepbinary(xaug, yvec, eta, imax);
```

3) Compare with log regression

4) Score the data

6) ROC Curve

6) Final threshold and Accuracy

Linear response: $\vec{x} \cdot \vec{w}$ (input · weight)

residual error: e - actuals

gradient = − back-propogation * x

Part 2: Kernel PCA

- Using Kernel PCA, to reduce dataset, run K-means Clustering
- How well can we classify IRIS dataset species
- Goal: 1) Compute gram matrix, from Kernel function
  2) Finding Spectral decomp
  3) use eigenvectors to project data

Gram Matrix - Similarity matrix tells you how close each data point is to every other datapoint

Good Gram Matrix: 1) Pairwise squared distance
2) Apply Kernel
3) center gram matrix

Process

1) Build gaussian Kernel gram matrix
2) Centre the matrix
3) Run PCA on it to reduce dimensions to 2D
4) K-means Clustering on 2D
5) Vizualization

Actual code

1) center gram matrix
2) Dimensionality reduction Setup

   Imax, Igram, Sigma2 = 2 * m

3) Create Gram Matrix

   D = pdist2(Xmat, Xmat, 'euclidean').^2

   Kmat = Gmat(n) - Kmat * Gmat(m)

Practise - MC Questions

1) $\phi(v) = \frac{1}{1+e^{v}}$

$\phi(v) = \frac{1}{1+e^{-(w^T x + b)}}$

$= \frac{1}{1+e^{-(2)}} \rightarrow = 0.88$

3) $\vec{w}_{k+1} = \vec{w_k} + \eta \vec{d}$

$= [1, 2] + 0.1[-0.5, 1]$

$= [1, 2] + [-0.05, 0.1]$

$= [0.95, 2.1]$ A

5) $K(u,v) = e^{-\gamma \|u-v\|^2}$

$= e^{-1 \| 5 \|}$

$= e^{-5}$

A

2) $b_i = r_i \phi'(v_i)$

$\phi'(v_i) = 0.4(1-0.4)$

$= (1-0.4)(0.4(1-0.4))$

$= 0.144$

4) Which is not used in computing descent direction $\vec{d}$ in log reg?

B) weight vector $\vec{w}$

6) Logistic function can not be directly used for multi-class classification

D

8) $= \frac{1}{1+e^{-(w^T x + b)}}$

$= \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix}$

$= -\frac{1}{2} + 1 = \frac{1}{2} \rightarrow \frac{1}{1+e^{-\frac{1}{2}}} = 0.62$

B

7) $\phi(v) = 0.8$

$\phi'(v) = 0.8(1-0.8)$

$= 0.16$

A

9) Objective/cost function is minimized when $r = 0$.

10) Two points: $\vec{x_1} = [1, 0]$, $b_1 = 2$

$\vec{x_2} = [0, 1]$, $b_2 = -1$

$d_1 = b_1 \cdot \vec{x_1} = 2 \cdot [1, 0] = [2, 0]$

$d_2 = b_2 \cdot \vec{x_2} = -1 \cdot [0, 1] = [0, -1]$

$d = d_1 + d_2$

$d = [2, -1]$

11) $\vec{w}_{k+1} = w_k + \eta \vec{d}$

$= [0.4, 0.8] + [0.3, -0.1]$

$= [0.7, 0.7]$ A

$x = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 2 & 1 \\ 2 & 3 & 1 \\ 2 & 4 & 1 \end{bmatrix}$

$v_3 = x_3 \cdot w_3$

$v_3 = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -5.5 \end{bmatrix}$

$v_3 = -\frac{1}{2}$ ✓

$z_3 = 0.3775$ ✓

$r_3 = 1 - 0.3775$

$= 0.6225$ ✓

$f(w, x, y) = \frac{1}{2}(0.6225)(0.6225)$

$= 0.19375$ ✓

deriv term $= (0.3775)(1-0.3775)$

$= 0.235$ ✓

$b_3 = (0.6225)(0.235)$

$= 0.1463$ ✓

$\nabla F_3 = -b_3 x_3 = -0.1463 \begin{bmatrix} 2 & 3 & 1 \end{bmatrix}$

$= [-0.2926 \quad -0.439 \quad -0.1463]$ ✓

$\vec{d_3} = b_3 x^T$

$= 0.1463 \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$

$= 0.8778$

$d_{full} = X^T b$

$\begin{bmatrix} 1 & 3 & 2 & 2 \\ 2 & 2 & 3 & 4 \\ 1 & 1 & 1 & 1 \end{bmatrix} 0.1463$

$=$

$w_{k+1} = w_k + \eta d$

$= 0g +$

$\vec{d_2} = [0.2926 \quad 0.439 \quad 0.1463]$

Unit 5- Exam Practise

1) $A = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}$ $\vec{w} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$

$X = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 2 & 1 \end{bmatrix}$

$\vec{v} = X \vec{w}$

$\vec{v} = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$

$\vec{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

2) $a = [-1 \quad 3]$, $\vec{w} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$X = [-1 \quad 3 \quad 1]$,

$v = X w$

$v = [-1 \quad 3 \quad 1] \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

$v = -3$

$\phi(v) = \frac{1}{1+e^3}$

$= 0.0474$

3) $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$ $y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $w = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$

$V = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$

$v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$\phi(v_1) = \frac{1}{1+e^{-1}} = 0.731$ $\quad \frac{1}{1+e^{-2}} = 0.88$

$r_1 = 1 - 0.731$ $\quad r_1 = 0.269$

$r_2 = 0 - 0.88$ $\quad r_2 = -0.88$

4) $a = [1 \quad 1 \quad 1]$ $y = 1$ $\vec{w} = \begin{bmatrix} 0.3 \\ -0.2 \\ 0.1 \end{bmatrix}$ $\eta = 1$

$b = 0.1114$

$\vec{w}_{k+1} = w_0 + \eta \vec{d}$ $\quad \vec{d} = x_i b_i$

$\vec{d} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} 0.114$

$\vec{d} = \begin{bmatrix} 0.114 \\ 0.114 \\ 0.114 \end{bmatrix}$

$w_{k+1} = \begin{bmatrix} 0.3 \\ -0.2 \\ 0.1 \end{bmatrix} + 1 \begin{bmatrix} 0.114 \\ 0.114 \\ 0.114 \end{bmatrix}$

$= \begin{bmatrix} 0.414 \\ -0.086 \\ 0.214 \end{bmatrix}$

5) $a_i a_j^T$

$A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 5 \end{bmatrix}$ $\quad [1 \quad 1] \begin{bmatrix} 4 \\ 1 \end{bmatrix}$

$K = \begin{bmatrix} 2 & 5 & 9 \\ 5 & 17 & 21 \\ 9 & 21 & 41 \end{bmatrix}$

HW #11

Part 1

$v_3 = [2 \quad 3 \quad 1] \begin{bmatrix} 1 \\ 1 \\ -5.5 \end{bmatrix}$

$v_3 = -0.5$

$z_3 = \frac{1}{1+e^{0.5}}$

$z_3 = 0.3775$

$r_3 = 1 - 0.3775$

$r_3 = 0.6225$

$f_3 = 0.19375$

Part 2

Derivative: 0.235

$b_3 = 0.1462$

$\nabla f = -b_i x_i$

$\nabla f_3 = (0.1462) \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$

$\nabla f_3 = \begin{bmatrix} -0.2924 \\ -0.4386 \\ -0.1462 \end{bmatrix}$

Part 3

$w_{k+1} = w_0 + \eta d$

$= \begin{bmatrix} 1 \\ 1 \\ -5.5 \end{bmatrix} + 0.5 \begin{bmatrix} 0.1985 \\ 0.6057 \\ 0.1910 \end{bmatrix}$

$= \begin{bmatrix} 1.0992 \\ 1.30285 \\ -5.43 \end{bmatrix}$

Part 4

$K(a_i a_j) = a_i a_j^T$

$[1 \quad 2]$

$\begin{bmatrix} 5 & 7 & 8 & 10 \\ 7 & 13 & 12 & 14 \\ 8 & 12 & 13 & 16 \\ 10 & 14 & 16 & 20 \end{bmatrix}$

$K(u,v) = e^{-\|u-v\|^2}$

$\begin{bmatrix} 1 & & & \\ 0.0183 & 1 & & \\ 0.135 & & 1 & \\ 0.0067 & & & 1 \end{bmatrix}$

c^{-4}

Homework 11 - By hand

Q1) Linear response: $v_i = x \cdot w$

$v_1 = [1\ 2\ 1] \begin{bmatrix} 1 \\ 1 \\ -5.5 \end{bmatrix}$

$v_1 = -2.5$

$v_2 = [3\ 2\ 1] \begin{bmatrix} 1 \\ 1 \\ -5.5 \end{bmatrix}$

$v_2 = -0.5$

Score/log activation

$z_1 = \phi(v_1)$

$z_1 = \phi(-2.5) = \frac{1}{1+e^{2.5}}$

$z_1 = 0.0759$

$z_2 = \phi(v_2) = \frac{1}{1+e^{v}}$

$= \phi(-0.5) = \frac{1}{1+e^{0.5}}$

$z_2 = 0.3775$

Residuals

$r_i = y_i - z_i$

$r_1 = 0 - 0.0759 = -0.0759$

$r_2 = 0 - 0.3775 = -0.3775$

Overall objective function

$f(\vec{w}, x, y) = f_1 + f_2 + f_3 + f_4$

$= 0.3392$

Sum all

Squared objective function → always positive

$f(\vec{w}, \vec{x}, \vec{y}) = \frac{1}{2} r^T r$

$f_1 = \frac{1}{2} [-0.0759] [0.0759]$

$f_1 = 0.0029$

$f_2 = \frac{1}{2} (0.3775)^2 \simeq 0.0713$

Q2)

Derivative of logistic function

$\phi'(v_1) = (\phi(v))(1 - \phi(v_1))$

$\phi'(v_1) = 0.0759(1 - 0.0759)$

$= 0.07$

$\phi'(v_2) = (\phi(v_2))(1 - \phi(v_2))$

$= 0.3775(1 - 0.3775)$

$= 0.235$

gradient ∇f_i

$\nabla f_1 = -b_1 \cdot x_i$

$f_1 = (0.005313)[1\ \ 2\ \ 1]$

$f_2 = 0.088\ [3\ \ 2\ \ 1]$

$\nabla f = \begin{bmatrix} 0.0053 & 0.00106 & 0.0053 \\ 0.2662 & 0.1774 & 0.0887 \\ X & X & X \\ X & X & X \end{bmatrix}$

Back Propogation

$b_1 = r_1 \cdot \phi(v_1)$

$= -0.0759 \cdot 0.07$

$= -0.005313$

$b_2 = r_2 \cdot \phi'(v_2)$

$b_2 = -0.3775 \cdot 0.235$

$b_2 = -0.088$

Descent vector

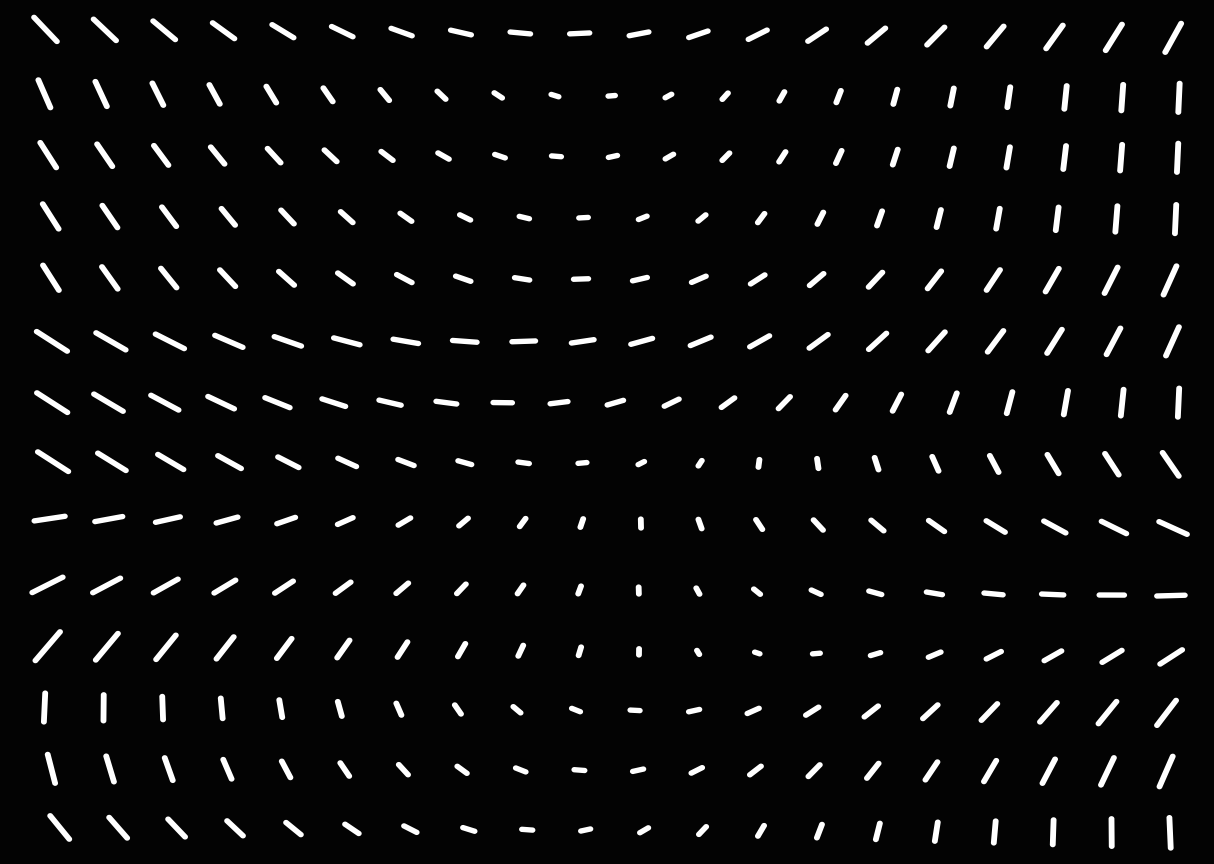
$d_i = -\nabla f_i^T$

$df = \begin{bmatrix} -0.0053 \\ -0.00106 \\ -0.0053 \end{bmatrix}$ $dz = \begin{bmatrix} -0.2682 \\ -0.1774 \\ -0.00887 \end{bmatrix}$

$\vec{d} = \vec{d1} + \vec{dz} + \vec{d3} + d4$

$\vec{d} = \begin{bmatrix} 0.19 \\ 0.6 \\ 0.14 \end{bmatrix}$

Exam Review



Exam Review All Topics Breakdown

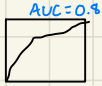
- Mostly Units 2-4. Few Unit 5 Questions
- No graphs, graph methods
- Need to know eigenvectors/eigenvalues very well
- Vector spaces, Matrix, column space, row space, Null space ($A\vec{x}=0$)

- Projection, (vector to vector) (vector to vector space) ⟷ • Normal Equation
- Linear Regression, (Projection of dependent data on indep)
- Validation: Root mean Squared Error; MSE
- Cross validation;
- Data Standardization, zero mean data, unit variance: $\sigma^2=1$
  - ↳ intercept term (yes/no)
- PCA and SVD $A = u_1 \Sigma_1 v_1^T + u_2 \Sigma_2 v_2^T + u_3 \Sigma_3 v_3^T$ (lin dependent)
  - ↳ $A = U \Sigma V^T$
  - ortho complement row space
  - Square root of eigenvalues
  - row space transpose Null space
  - PCA, is eigenvectors of covariance, which gives you Principal Components
  - For $r_j = \frac{\sum x_i}{\sum x_j}$
- Scree plot: Shows you explained variance and most important Principal components at the elbow. Based on covariance matrix $\frac{1}{n} M^T M$
- Unsupervised learning; K means clustering
  - ↳ centroid is in a set, from centroid find partitions, from partitions find centroids
- Supervised learning, data observation. a_i has label y_i.
  - ↳ Compute Hyperplane and Scores z_i
- Assessment: ϕ threshold assign quantization q_i. Check score $\geq \phi$ or $\leq \phi$
  - ↳ labels are ± 1, Confusion matrix

TP	FP
FN	TN

→ convert to rate

- AUC/ROC curve → FPR vs TPR



- Probability, odds (activation function)
  - ↳ for Hyperplane $|H| = \vec{m}, b$ $m = \frac{\vec{m}}{\|m\|}$ $b = \frac{b}{|\vec{m}|}$
- logistic function $\phi(\vec{i})$
- Artificial neuron: 0 or 1, take data reading
  - $z_i = \phi(v)$ → residual error: label − Score
- Descent Vector is residual times derivative of linear response, data input transpose $\vec{i}$
- Learning rate: Continually getting better

Polar Decomp

For $A \in \mathbb{R}^{n \times n}$

$A = U \Sigma V^T$

$A = UV^T V \Sigma V^T$

$A = Q B$

where $QQ^T = I$

$B = B^T$

Exam Review - All Test Questions

- Questions mostly from Test 2 - Test 4
- Few questions from Test 1 and Test 5

Test 1

1) Finding Laplacian matrix, given graph and weighted edge list.

$L = D - A$

D is diagonal matrix, where each diagonal entry is the sum of weights of all edges connected to i

A is adjacency matrix, diagonal of 0's A_{ij} = weight of edge between i, j

$\mathcal{E} = \{ (1\ 3\ 1) \quad (2\ 4\ 1) \quad \dots \}$

Vertices 1 and 3 are connected with weight 1

Sum weights for each node: $1 \to 2$, $2 \to 2$, $3 \to 2$, $4 \to 1$, $5 \to 2$, $6 \to 1$

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$L = D - A = \begin{bmatrix} 2 & 0 & -1 & 0 & -1 & 0 \\ 0 & 2 & 0 & -1 & 0 & -1 \\ -1 & 0 & 2 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

2) $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 3 & 2 \\ 3 & 1 & -2 \end{bmatrix}$ Which columns form a basis for the column space?

- column space of A is span of column vectors → check lin independence
- basis for col space is lin indep col

∴ want to find columns that are lin indep and capture all possible combinations

RREF = rank 2, dimension of col space is 2. ∴ any 2 lin indep columns form basis

3) Given: $A_3 = I - \frac{1}{9}\vec{w}\vec{w}^T = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}$ $w = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}$

Find $\vec{w}^T\vec{w} = \begin{bmatrix} 1 & 2 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \\ 3 \end{bmatrix}$

Need to find eigenvector corresponding to $\lambda = -1$

$A\vec{v} = \lambda\vec{v}$

$A\vec{v} = I\vec{v} - \frac{1}{9}\vec{w}\vec{w}^T\vec{v}$

$A\vec{v} = \vec{v} - \frac{\vec{w}^T\vec{v}}{9}\vec{w}$

$= \left(\frac{w \cdot w}{9}\right)\vec{w}$

$A\vec{w} = (I - \alpha\vec{w}\vec{w}^T)\vec{w}$

$= w - \alpha\vec{w}(w^Tw)$

$A\vec{w} = (1 - \alpha \cdot w^Tw)\vec{w}$

$\lambda = 1 - \alpha(w^Tw)$

∴ w is an eigenvector

4) Given $M = \begin{bmatrix} 1+\varepsilon & 1+\varepsilon \\ 0 & 2 \end{bmatrix}$, $0 < \varepsilon < 1$, how do eigenvalues of M change?

Typically: $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ → using M → $\det\left(\begin{bmatrix} 1+\varepsilon-\lambda & 1+\varepsilon \\ 0 & 2-\lambda \end{bmatrix}\right) = (1+\varepsilon-\lambda)(2-\lambda)$

$\det(A - \lambda I) = 0$

$(1+\varepsilon-\lambda)(2-\lambda) = 0$

When $\varepsilon = 0$ we get eigenvalues 1 and 2

the eigen value $1+\varepsilon$ increases a bit

the eigen value of 2 stays the same not dependent on ε.

Test 1

5) Given matrix $A_5 = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}$ with $tr(A_5) = 11$ $\quad det(A_5) = 30$

$\lambda_1 = 2 \quad \lambda_2 = 3$

Find $\lambda_3, \lambda_4 = ?$

Trace is sum of eigenvalues, check if options add up to 11

$\lambda_3 + \lambda_4 = 6$

Determinant = product of eigenvalues $\quad \lambda_1 \cdot \lambda_2 \cdot \lambda_3 \cdot \lambda_4 = 30$

$6 \cdot \lambda_3 \cdot \lambda_4 = 30$

$\lambda_3 \cdot \lambda_4 = 5$

$\lambda_3 + \lambda_4 = 6$
$\lambda_3 \cdot \lambda_4 = 5$

→ Check options: $x+y=6$, $xy=5$

6) $A_6 = \begin{bmatrix} 33 & -6 & -6 \\ -6 & 5 & 7 \\ -6 & 7 & 5 \end{bmatrix}$ $\quad \lambda_1 = 36$, $\lambda_2 = 9$, $\lambda_3 = -2$

Diagonal matrix similiar $\quad A_6 = V_6 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix} V_6^T$

whats the relationship of columns of V_6.

$A_6 = V_6 D_6 V_6^T$

V_6 is made of eigenvectors

$V_6^T = V_6^{-1}$

means orthonormal

Analysis:
- Because A_6 is diagonalizable, its Symmetric
- Eigenvectors of V_6 are orthogonal

Test 2

1) $A_1 = \begin{bmatrix} -12 & -7 \\ 3 & -1 \\ 4 & -1 \\ 10 & 2 \\ 10 & 17 \end{bmatrix}$ $\quad 3 \quad 2$

find standardized z values:

$z = \frac{x - \mu}{\sigma}$

$N_1 = 3$
$N_2 = 2$

→ Find Squared deviations

$(-12-3)^2 = 225$
$(3-3)^2 = 0$
$(4-3)^2 = 1$
$\dots = 49$
$= 49$

$\sigma_1 = \sqrt{\frac{225+0+1+49+49}{5}}$

$\sigma_1 = 8.05$

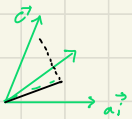
$z_{11} = \frac{-12-3}{8.05}$

$z_{11} = -1.667$

$z_{21} = \frac{3-3}{8.05} = 0$

$z_{31} = \frac{10-3}{8.05} = 0.778$

2) Given:
- Vector space $V \subseteq \mathbb{R}^m$
- spanned by vectors $\{\vec{a1}, \vec{a2}, \dots \vec{an}\}$
- data vector $c \in \mathbb{R}^m$



what does it mean to project $\vec{c}$ into V

- Finding projection vector p that is closest euclidean distance to $\vec{c}$.
- want to minimize distance
- Finding best possible vector in span of $\vec{a_i}$

3) For vector space V spanned by $A_3 = \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$, Find error Projection of $\vec{c_3} = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix}$ into V?

$\vec{e} = \vec{c_3} - Proj_V(\vec{c_3})$

$Proj_V(\vec{c}) = A_3 \vec{w}$

$A^T \vec{e} = \vec{c}$

Key: $A^T A \vec{w} = A^T \vec{c}$

$A^T A = \begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$

$= \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}$

$A^T A = 7[I]$

$A^T c = \begin{bmatrix} 1 & 2 & 1 & 1 \\ -1 & -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix} = \begin{bmatrix} -28 \\ 14 \end{bmatrix}$

weights $= \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$(-28/7)$

$\vec{e} = \vec{c_3} - A_3 \vec{w}$

$e = \begin{bmatrix} -11 \\ -7 \\ -2 \\ -1 \end{bmatrix} - \begin{bmatrix} 1 & -1 \\ 2 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -4 \\ 2 \end{bmatrix}$

$e = \begin{bmatrix} -5 \\ 3 \\ 0 \\ 1 \end{bmatrix}$

4) Given a linear model Find K, s.t $c_i = Ka_i$

i	0	1	2	3
a_i	-1	2	3	4
c_i	-3.1	3.5	5	8.5

$K = \frac{a^T c}{a^T a}$

$K = \frac{59.1}{30}$

$K = 1.97$

This regression goes through origin

Want to minimized Square error

Projecting c onto a
Find scalar to make it close as possible

Ka is projection of $\vec{c}$ onto $\vec{a}$.

5) Given $A = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}$ $\vec{c_5} = \begin{bmatrix} 7 \\ 12 \\ 15 \end{bmatrix}$ find RMS error of 3 fold CV of linear regression.

3 fold CV means train on 2, test on 1 repeat 3 times, averages error

$\hat{y} = w \cdot x$

$\vec{w_1} \approx \begin{bmatrix} 12 \\ 15 \end{bmatrix}$ → Fold 1: $\frac{12 \cdot 5 + 15 \cdot 7}{5^2 + 7^2} = \frac{165}{74} \approx 2.23$

$\vec{w_2} \approx \begin{bmatrix} 7 \\ 15 \end{bmatrix}$ → Fold 2: $= 7 - 3\vec{w_1}$ $= \underbrace{7 - 3(2.23)}_{e}$ (actual) $= 0.31$ – error 1

$\frac{3 \cdot 7 + 7 \cdot 15}{3^2 + 7^2} = 2.17$ $= \underbrace{12}_{\text{actual}} - 5(2.17) = 1.14$ – error 2

$\vec{w_3} \approx \begin{bmatrix} 7 \\ 12 \end{bmatrix}$ → Fold 3: $\frac{3 \cdot 7 + 5 \cdot 12}{3^2 + 5^2} \approx 2.38$ $= \underbrace{15}_{\text{actual}} - 7(2.38) = -1.68$ – error 3

$RMS = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2}{3}}$

$= 1.18$

In general linear regression:
Slope or weight is

$w = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}$

→ then you get Predicted y value $\hat{y} = w \cdot x$

→ error is $e = y_{actual} - \hat{y}$ for each fold

→ $RMS = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2}{3}}$

6) Given matrix $A_6 = \begin{bmatrix} 2 & 11 & 13 & 16 \\ 5 & 17 & 22 & 19 \\ 3 & 13 & 16 & 17 \\ 7 & 19 & 26 & 17 \end{bmatrix}$ with null space $null(A_6) = \begin{bmatrix} -1 & 3 \\ -1 & -2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$ SVD is $A_6 = U_6 \Sigma_6 V_6^T$

where $V_6 = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}$

Span row space Span null space

Select columns of V6 that are basis for null(A)

We know null space is 2 dim

2 non zero singular values 2 zero ones

Rank(A) = 2

null space basis = last n - rank(A) cols of V

Summary / Reminder:

Σ = diagonals σ_1, σ_2 positive numbers
Square roots of eigenvalues of $A^T A$

V = right singular vectors of $A^T A$
↳ form orthonormal basis
↳ last few col of V

U = left singular vectors of AA^T
↳ form orthonormal basis for output space

$\nabla p_i = -b_i x_i$

$\vec{d} = [\nabla p_i]^T$

$\sigma = \sqrt{\lambda_i}$

Test 3

1) Given $A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 2 \end{bmatrix}$, $A_1^T A_1 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$, $A_1 A_1^T = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 2 & 2 & 8 \end{bmatrix}$.

Find singular values

Square roots of eigenvalues of $A^T A$

$\det(A - \lambda I)$

$\det\begin{pmatrix} 5-\lambda & 4 \\ 4 & 5-\lambda \end{pmatrix}$

$= (5-\lambda)^2 - 16 = 0$

$(5-\lambda)^2 = 16$

$\lambda_1 = 1, \lambda_2 = 9$ — Singular values

square roots

$\{3, 1\}$

$(5-\lambda)(5-\lambda)$

$25 - 5\lambda - 5\lambda + \lambda^2$

$\lambda^2 - 10\lambda + 9 \quad (\lambda - 9)(\lambda - 1)$

$\lambda = 9, \lambda = 1$

*

2) Given matrix $A_2 = U_2 \Sigma_2 V_2^T$

$A_2 = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}$ $A^T A = \begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}$, $A A^T = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}$

Find U and V that describe basis for Colourm space and row space

U and V must be 2 ranked,

$\sigma = \sqrt{\lambda_i}$

U: basis for Column space

V: basis for row Space

3) Given a matrix $A_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$ and 3 properties

Which ones describe A_3 properly

Analysis: 3 by 4 matrix

1) $A^T A$ is a 4 by 4 matrix

AA^T is a 3 by 3 matrix

So by default they can't be identical

They can't share the eigenValues

2) Non zero singular values would be 1, 2, 3

3) Removing col # 1, does not change σ_j

4) Given zero mean $M = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$ $\begin{bmatrix} \sigma_1 = 4 \\ \sigma_2 = 2 \end{bmatrix}$ $U_4 = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}$ $V_4 = \begin{bmatrix} -0.71 & 0.71 \\ -0.71 & -0.71 \end{bmatrix}$

Find First PCA Score:

$Z = U\Sigma$

$Z = M \vec{v_1}$

- Each column of Z is a Principle component score vector

$Z = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix} \begin{bmatrix} -0.71 & 0.71 \\ -0.71 & -0.71 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -2.84 \\ 2.84 \end{bmatrix}$

4 by 2 · 2 by 2

5) For a matrix $M = \begin{bmatrix} 12 & -1 & -12 \\ 8 & 3 & -8 \\ -12 & -5 & 24 \\ -8 & 3 & -4 \end{bmatrix}$

$\sigma_1 = 33.61$, $\sigma_2 = 11.04$ $\sigma_3 = 2.84$, How many Principal components capture 96% of explained variance in data.

- Eigenvalues of covariance are explained Variance

$\sigma_1^2 = 1129.63$ → 89.6% of variance

$\sigma_2^2 = 121.88$ → 99% of variance

$\sigma_3^2 = 8.07$ → 100% of variance

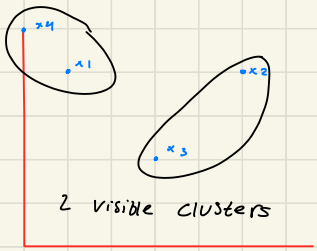
1259.58

Therefor 2 principal components are needed

Test - 4

1) Given 4 data vectors $x_1 = \begin{bmatrix}1\\4\end{bmatrix}$ $x_2 = \begin{bmatrix}4\\4\end{bmatrix}$ $x_3 = \begin{bmatrix}3\\2\end{bmatrix}$ $x_4 = \begin{bmatrix}0\\5\end{bmatrix}$
Find 2 best sets of data

Compute euclidean distance or plot



2) For two clusters, given g_1 g_2 and Hyperplanes unit normal
Find associated bias Scalar c_2.

$h = \frac{g_1 + g_2}{2}$ $g_1 = \begin{bmatrix}1\\2\\1\end{bmatrix}$ $\begin{bmatrix}3\\1\\3\end{bmatrix}$ $\vec{n_2} = \begin{bmatrix}0.667\\-0.333\\0.667\end{bmatrix}$

$b = -h\vec{n}$ $h = \begin{bmatrix}2\\1.5\\2\end{bmatrix}$

$b = -\begin{bmatrix}2\\1.5\\2\end{bmatrix}\begin{bmatrix}0.667\\-0.333\\0.667\end{bmatrix} \rightarrow b = -2.1685$

Other way → $\vec{w} = \vec{g_1} - \vec{g_2}$ $\|\vec{w}\| = 3$

$\vec{p} = \frac{g_1 + g_2}{2}$ $b = -w^T\vec{p}$

$b = [2\ -1\ 2]\begin{bmatrix}2\\1.5\\2\end{bmatrix}$

$b = 6.5$

$c = \frac{b}{\|w\|} = 2.167$

3) Given Hyperplanes normal vector and bias scaler.

$n = \begin{bmatrix}-2/3\\1/3\\2/3\end{bmatrix}$ $c_3 = -1$ $x_1 = \begin{bmatrix}0\\2\\5\end{bmatrix}$ $x_2 = \begin{bmatrix}4\\0\\1\end{bmatrix}$

Classifying two data vectors as pos or neg

$d = x_1\vec{n} - 1$

$d = \begin{bmatrix}0\\2\\5\end{bmatrix}\begin{bmatrix}-2/3 & 1/3 & 2/3\end{bmatrix} - 1$

$d = 3$

$d_2 = x_2\vec{n} - 1$

$d_2 = \begin{bmatrix}4\\0\\1\end{bmatrix}\begin{bmatrix}-2/3 & 1/3 & 2/3\end{bmatrix}$

$d = -3$

∴ x_1 is positive, x_2 is negative

4) Hyperplane of m and bias scalar
Find prob x is positive

$m_4 = \begin{bmatrix}-4\\3\end{bmatrix}$

$b_4 = 7$

$x_4 = \begin{bmatrix}2\\2\end{bmatrix}$

$n_4 = \begin{bmatrix}-4/5\\3/5\end{bmatrix}$

$c_4 = 7/5$

$d = nx + c \rightarrow d = 1$

$p = \frac{1}{1+e^{-d}}$ $p = \frac{1}{1+e^{-1}}$ $p = 0.73$

5)

z_i	-3	-2.5	-2	-1	0.5	1	1.5	3	3.5	5.0	5.5	6
Pred: y_i =	-1	1	-1	-1	-1	1	-1	1	-1	1	1	1
Actual:	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1
	TN	FP	TN	TN	TN	FP	TN	TP	FN	TP	TP	TP

TP	FP
FN	TN

4	2
1	5

Test 6

1) Given $A = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}$ weight vector $\vec{w} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

Find linear response

$U = x \cdot w$

$U = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \end{bmatrix}$

2 by 3 · 3 by 1

2) given $a = [-1 \quad 3]$, weight vector $\vec{w} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

$z = \frac{1}{1+e^{-v}}$

$V = [-1 \quad 3 \quad 1] \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

$V = -3$

$z = \frac{1}{1+e^{3}}$

$z = 0.0474$

3) Find residual error given the following: $A = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}$ $\vec{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\vec{w} = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$

$r_1 = y_1 - \phi(u_1)$

$u_1 = x \cdot w$

$u_1 = [1 \; 2 \; 1] \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$

$v_1 = 1$

$z_1 = \frac{1}{1+e^{-1}}$ $z_1 = 0.731$

$r_1 = 1 - 0.731$

$r_1 = 0.269$

$u_2 = x_2 \cdot w$

$u_2 = [1 \; 3 \; 1] \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}$

$u_2 = 2$

$z_2 = \frac{1}{1+e^{-2}}$ $z_2 = 0.88$

$r_2 = 0 - 0.88$

$r_2 = -0.88$

$r = \begin{bmatrix} 0.269 \\ -0.88 \end{bmatrix}$

4) given: $a = [1 \quad 1]$, $y = 1$ $\vec{w_0} = \begin{bmatrix} 0.3 \\ -0.2 \\ 0.1 \end{bmatrix}$ $\eta = 1$

$b = 0.1114$

$w_{k+1} = w_k + \eta \vec{d}$

$\vec{w_1} = \begin{bmatrix} 0.3 \\ -0.2 \\ 0.1 \end{bmatrix} + 1 \begin{bmatrix} 0.1114 \\ 0.1114 \\ 0.1114 \end{bmatrix}$

$\vec{w_1} = \begin{bmatrix} 0.4114 \\ -0.0886 \\ 0.2114 \end{bmatrix}$

$d = 0.1114 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$d = \begin{bmatrix} 0.1114 \\ 0.1114 \\ 0.1114 \end{bmatrix}$

5) Gram matrix K derived from linear kernel $K_{ij} = a_i a_j^T$

For data matrix $A = \begin{bmatrix} 1 & 1 \\ 4 & 1 \\ 4 & 5 \end{bmatrix}$, Find K

$\begin{bmatrix} 2 & 5 & 9 \\ 5 & 17 & 21 \\ 9 & 21 & 41 \end{bmatrix}$

$[1 \; 1] \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$[4 \; 1] \begin{bmatrix} 4 \\ 1 \end{bmatrix}$

$[4 \; 5] \begin{bmatrix} 4 \\ 5 \end{bmatrix}$

End