MLF

CLASSTIME | Page No.
Date / /

Week-1

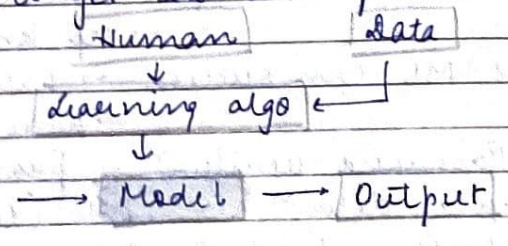
L1.1 What is ML?

- ML is the study of comp. algo that improves automatically through experience and by-the use of data.
- ML tasks — Weather predic", Face detec"
  - Task Hierarchy
  - Human → Tool design ← data
  - Input → Tool → Output
- Why & when ML?
  1. Human fails → scale / speed / cost
     → inability to express
     → no knowledge of algo
  2. ML can succeed → have lots of common data
     → have some struc. idea
- Password verifica" — programming
- Face detec" — ML (lots of image available)
- Weather predic" - ML
- ML in inbox — spam classifier
- ML in shopping cart - frequently recommended system — Netflix, YT
- Smart Assistant - Alexa, Siri, etc.
- ML in Robot AIs, Games (Chess, Checkers, Go), Marketing (budgeting)

L1.2 Data, Models and ML Tasks

- data - it is a collec" of vectors
- metadata - it is info. on the data
- model - it is a mathematical simplification of reality
- In ML — Predictive Model → Regression
  Probabilistic → Classifica

- Regression model → model the price of a house based on metro dist. & area
  (real valued output) price = 0.5 × area − distance
- Classificaⁿ → model whether a house is closer than 2 kms to a metro based on price and area
  (discrete values)
- probabilistic model → evaluate how likely an event could occur
- Learning Algo : data → Model
  ↳ choose from a collecⁿ of models, with same struc. but diff. parameters
  ↳ use data to get the 'best' parameters
- ML



L1.3 Supervised Learning : Regression

- supervised learning is curve-fitting
- Given $\{(x^1, y^1), (x^2, y^2) \dots (x^n, y^n)\}$
  ↳ find a model f such that $f(x^i)$ is ≈ to y^i
- Regression − eg. predict house price from room, area, distance

$f : R^d \to R$

$$\text{Loss}(f) = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(f(x^i) - y^i\right)^2$$

$$f(x) = w^T x + b = \sum_{j=1}^{d} w_j x_j + b$$
$$(w_1 x_1 + w_2 x_2 \dots)$$

eg.

rooms	A	dist.	price
⋮	⋮	⋮	⋮

$f = 2 \times \text{rooms} - 0.5 \times d$
$g = \text{rooms} + \dots$

→ calculate f & g, & check for the lower loss

L1.4 SL : Classificaⁿ

Eg. Predict if rooms > 3 from area & price

$x^i \in R^d$, $y^i \in \{+1, -1\}$

$f : R^d \to \{+1, -1\}$ Loss = $\frac{1}{n} \sum_{i=1}^{n} (1(f(x^i) \neq y^i))$ (frac^n for misclassified instances)

$$f(x) = \text{sign}(w^T x + b)$$
↳ linear expression

- Learning algo uses training data to get model f. But evaluaⁿ should n't be on the training data. Use test data which is not in the training data.
- Learning algo finds the best model in collecⁿ of models. This is c/d model selecⁿ and it is done by using another dataset c/d validaⁿ data that is distinct from train & test data.

L1.5 Unsupervised Learning : Dimensionality Reducⁿ

- UL — understanding data. data : $\{x^1, x^2, \dots x^n\}$
  $x^i \in R^d$
- it build models that compress, explain and group data
- Eg. cake tweet eg. → groups of ID.
- Dimensionality reducⁿ − compression and simplificaⁿ

Eg. represent 1 M gene expression of a million people

data : $\{x^1, x^2, \dots x^n\}$ $x^i \in R^d$

Encoder $f: R^d \to R^{d'}$ decoder $g: R^{d'} \to R^d$
$d' < d$

Goal: $g(f(x^i)) \approx x^i$

Loss $= \frac{1}{n} \sum_{i=1}^{n} \| g(f(x^i)) - x^i \|^2$

L1.6 UL: Density Estimation
- a probabilistic model
- create a robot accnt. which generates more such tweets

Eg. wisdom of chopra.com
- DE model takes in several samples from a random source, and outputs a model that assigns a probability score to every possible sentence.
- data: $\{x^1, x^2 \dots x^n\}$ $\quad - x^i \in R^d$
- Probability mapping $P: R^d \to R_+$ that 'sums' to 1.
- Goal: $P(x)$ is large if $x \in$ data, and low otherwise
- Loss: $\frac{1}{n} \sum_{i=1}^{n} -\log(P(x^i))$

-ve log likelihood

T1.1 To ML or not to ML
- data / samples
- do we need to have a ML approach to solve a given ques.? → If ML –
- find the metadata about the data.
- a, b, c are the parameters or wts. of the model. The best values for the parameters will be learning from data.

T1.2 Illustration with a real-world dataset
- breast cancer classification
  - → x_j^i $\quad 1 \le j \le 30$ ⇒ columns (features)
  - → x^i $\quad 1 \le i \le 569$ ⇒ rows (cases)
  - → $y^i = f(x^i)$
  - → loss $= \frac{1}{n} \sum \mathbb{1}(y^i \neq y^i)$
  - → which func^n is suitable?
- Train Set: 80% of Total
- Valida^n Set: 20% of Train
- Test Set: 20% of Total

T1.3 UL with applications
- how can we compress 30 features to 3 features w/o losing much information
  - ↳ dimensionality Reduc^n
  - (only data with no labels)

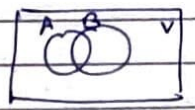
$u = f(x) = Wx$ → encoder
$x' = g(u) = W^T u$ ⇒ decoder

- Density estima^n → all samples are assumed to be from some probability distribu^n. 569 samples are insufficient.
  - ↓
  - Generate more samples by estimating the PDF for each features.
- new face – generated by deep learning model StyleGAN (Generative Adversial networks)
  - ↳ density estima^n

Week-2

L2.1 Sets & Functns

- $\mathbb{R}$ (set of real nos.)
  $\mathbb{R}_+$ (set of +ve real nos including 0)
  $\mathbb{Z}$ (set of integers)
  $\mathbb{Z}_+$ (set of +ve integers with 0)
  $[a,b]$ $a \le x \le b$
  (a,b) $a < x < b$
  $\mathbb{R}^d$ set of d-dimensional vectors
- $\mathbb{R}^d$: $D(x,y) = \|x-y\| = \sqrt{(x_d - y_d)^2}$
- (open) $B(x,\epsilon) = \{y \in \mathbb{R}^d : D(x,y) < \epsilon\}$
  (closed) $\bar{B}(x,\epsilon) = \{y \in \mathbb{R}^d : D(x,y) \le \epsilon\}$
- V is the universe



$A \cup B$, $A \cap B$, A^c ($V \setminus A$), B^c

$(A \cup B)^c = A^c \cap B^c$
$(A \cap B)^c = A^c \cup B^c$ } De-morgan's laws

- $\forall$ for all $\exists$ there exist $\Rightarrow$ implies
  $\Leftrightarrow$ equivalent to
- $x_1, x_2, \ldots$ where $x_i \in \mathbb{R}^d$
  $\lim_{i \to \infty} x_i = x^*$
  $\Updownarrow$
  $\forall \epsilon > 0$, $\exists N$
  $x_n \in B(x^*, \epsilon)$ $\forall n \ge N$
  } convergent sequence
- (i) $x_i \in \mathbb{R}$ $x_n = 1 + n$
  (ii) $x_i \in \mathbb{R}^2$ $x_n = \left(\frac{1}{2^n}\cos\left(\frac{\pi n}{2}\right), \frac{1}{2^n}\sin\left(\frac{\pi n}{2}\right)\right)$
- If V is a vector space, $u \in V$, $v \in V$
  $\boxed{\alpha u + \beta v \in V}$ (most crucial prop.)
  → $\mathbb{R}^d$ is a vector space
  → $x \cdot y = x^T y = \sum_{i=1}^{d} x_i y_i$ (dot prod.)
  → $\|x\|^2 = x \cdot x = x^T x = \sum_{i=1}^{d} x_i^2$ (norm)
  → x & y are $\perp$ to each other if their dot product is 0.
- $f : A \to B$ (domain) (co-domain)
  1D : $f : \mathbb{R} \to \mathbb{R}$
  d dim : $f : \mathbb{R}^d \to \mathbb{R}$
  $G_f \subseteq \mathbb{R}^{d+1}$
  $G_f = \{(x, f(x)) : x \in \mathbb{R}^d\}$
- contour plots of 2 dimensional functions
  ↳ heat map (when fully coloured)
  ↳ it has ∞ contours so more information.

L2.2 Univariate Cal : Continuity & Differentiability

- $f : \mathbb{R} \to \mathbb{R}$
  is continuous at $x^* \in \mathbb{R}$ if for all sequences x $x_1, x_2 \ldots$ converges to x^*. we have that $f(x_i)$ converges to $f(x^*)$
  $\lim_{i \to \infty} x_i = x^* \Rightarrow \lim_{i \to \infty} f(x_i) = f(x^*)$
  $\lim_{x \to x^*} f(x) = f(x^*)$

Eg. $f(x) = \text{sign}(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases}$
$x^* = 0$

- a func^n is said to be continuous if it is continuous at all points in its domain
- a func^n $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $x^* \in \mathbb{R}$ if $\lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*}$ exists.
  $= f'(x^*)$

- If f is not continuous at x^* $\Rightarrow$ f is not differentiable at x^*

$$f'(x^*) = \lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*}$$

$$= \lim_{h \to 0} \frac{f(x^* + h) - f(x^*)}{h}$$

L2.3 Derivatives & linear approx.

- Let $f : \mathbb{R} \to \mathbb{R}$ be a diff. func^n

$$f'(x^*) = \lim_{x \to x^*} \frac{f(x) - f(x^*)}{x - x^*}$$

$$f'(x^*) \approx \frac{f(x) - f(x^*)}{x - x^*} \quad (\text{around } x = x^*)$$

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*)$$

$$\boxed{f(x) \approx L_{x^*}[f](x)}$$

Eg. $f(x) = x^2$, $x^* = 1$

$$L_{x^*}[f] = f(x^*) + f'(x^*)(x - x^*)$$
$$= 1^2 + 2(x-1)$$
$$= 2x - 1$$

- $G_{L_{x^*}[f]}$ is a tangent to the graph of G_f at the point $x^*, f(x^*)$.

Eg. $f(x) = \sin(x)$, $x^* = 0$

$f'(x) = \cos(x)$

$f'(x^*) = 1$

$f'(x^*) = 0$

$f(x) = 0 + 1(x - 0)$

$= x$

$\sin x \approx x$ if $x = 0$

Eg. $f(x) = e^x$, $x^* = 0$

$e^x = e^0 + (x - 0) \cdot 1$

$\approx 1 + x$ around $x = 0$

Eg. $\ln(1 + x)$ around $x^* = 0$

$\approx x$

Eg. $f(x) = (1 + x)^a$ around $x^* = 0$

$\approx 1 + ax$

L2.4 Multivariate Calculus: app & adv rules

- Linear App. $\rightarrow f(x) \approx L_{x^*}[f](x)$

Quadratic App. $\rightarrow f(x) \approx L_{x^*}[f](x) + \frac{1}{2} f''(x^*)(x - x^*)^2$

- (higher order approx. are better than linear approx.)

- $f(x) = x^2$

$x^2 = (x^*)^2 + 2x^*(x - x^*) + \frac{1}{2} \cdot 2 \cdot (x - x^*)^2$

$= x^2$ (only for quadratic eq^n)

Eg. higher order approx. of e^x around $x^* = 0$

$1 + x + \frac{x^2}{2}$

- $f(x) = g(x) \cdot h(x)$ Prod. rule

$f'(0) = g'(0)(h(0)) + h'(0) g(0)$

- $f(x) = g(h(x))$ Chain rule

$f'(0) = g'(h(0)) h'(0)$

Eg. $\frac{e^{3x}}{\sqrt{1+x}}$ around $x = 0$

$\approx (1+3x)\left(1 - \frac{x}{2}\right)$

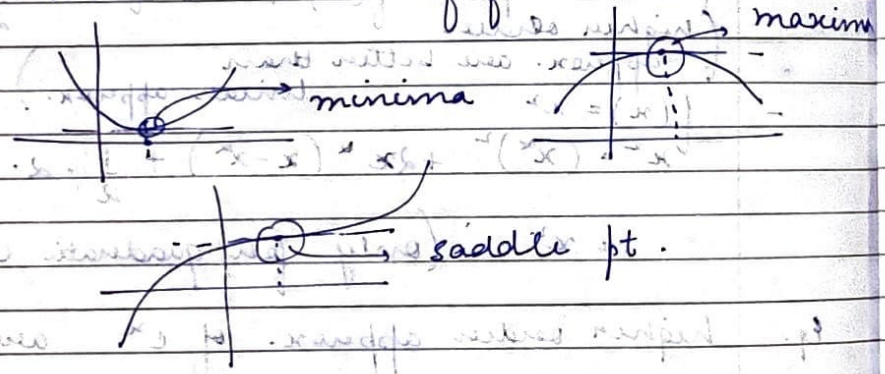
$\approx 1 + \frac{5}{2}x$

Eg. $e^{\sqrt{1+x}}$ around $x = 1$

$\approx e^{\sqrt{2}} + \frac{e^{\sqrt{2}}}{2\sqrt{2}}(x-1)$

- $L_{x^*}[f] = f(x^*) + f'(x^*)(x - x^*)$

$f'(x^*) = 0 \iff x^*$ is a critical point of f



L 2.5 Multi. Cal. – Lines & planes in higher dim. plane

(i) A line in $R^n \subseteq R^d$

(ii) A line thru the pt. $u \in R^d$ along the vector $v \in R^d$

$= \{x \in R^d : x = u + \alpha v \text{ for } \alpha \in R\}$

similarly line thru $(u, u') \in R^d$

u along $u' - u$

u along $u - u'$

- A $(d-1)$ dimensional hyperplane $\subseteq R^d$.
- A hyperplane normal to the vector $w \in R^d$ with value $b \in R$

$= \{x \in R^d : w^T x = b\}$

Eg. Line thru $\binom{1}{1}$ along $\binom{1}{2}$ —//—

- Tuples as Points & as Vectors (locations) (directions)
- $f : R^2 \to R \quad f(x_1, x_2) = x_1^2 + x_2^2$

$\frac{df}{dx_1}(v) = \lim_{\alpha \to 0} \frac{f\left(v + \begin{bmatrix}\alpha \\ 0\end{bmatrix}\right) - f(v)}{\alpha}$

$= \frac{f(v_1 + \alpha, v_2) - f(v_1, v_2)}{\alpha}$

$\frac{df}{dx_i}(v) = \lim_{\alpha \to 0} \frac{(v + \alpha e_i) - f(v)}{\alpha}$

- Gradients

$f : R^d \to R$

$\frac{df}{dx} = \left[\frac{df}{dx_1}(v), \ldots, \frac{df}{dx_d}(v)\right]$

$\nabla f(v) = \left[\frac{df}{dx}\right]^T$

Eg. $d = 2$. $f(x) = x_1^2 + x_2^2$ $\quad \frac{df}{dx_1}(v) = 2v_1$

$\frac{df}{dx_2} \Rightarrow v = 2v_2 \quad \therefore \nabla f(v) = \begin{bmatrix}2v_1 \\ 2v_2\end{bmatrix}$

Eg. $f(x) = x_1 + 2x_2 + x_3 \Rightarrow$ Linear func^n $\quad \nabla f(v) = \begin{bmatrix}1 \\ 2 \\ 1\end{bmatrix}$

L 2.1

MC: Linear approximation & app.

$f: \mathbb{R} \to \mathbb{R}$

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*)$$

$L_{x^*}[f](x)$

$u \in \mathbb{R}^d, \quad x \in \mathbb{R}^d$

$$f(x) \approx f(u) + \nabla f(u)^T (x - u)$$

$L_u[f](x)$

- $f: \mathbb{R}^2 \to \mathbb{R}$

$$f(y_1, y_2) \approx f(u_1, u_2) + \frac{\delta f}{\delta x_1}(u) \cdot (y_1 - u_1)$$

Eg. $f(x_1, x_2) = x_1^2 + x_2^2 \qquad \nabla f_x = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$

(i) App. f around $(6, 2)$

$f(u) = 40, \quad \nabla f(u) = \begin{bmatrix} 12 \\ 4 \end{bmatrix}$

$$f(x) = 40 + [12, 4] \begin{bmatrix} x_1 - 6 \\ x_2 - 2 \end{bmatrix}$$

$$= 12x_1 + 4x_2 - 40$$

- The graph of $L_u[f]$ is tangent to the graph of f at the pt. $(u, f(u))$
- Directional derivative

$$D_u[f](u) = \lim_{\alpha \to 0} \frac{f(u + \alpha u) - f(u)}{\alpha}$$

$$= \nabla f(u)^T u$$

- Cauchy-Schwarz Inequality

$$-\|a\| \cdot \|b\| \leq a^T b \leq \|a\| \cdot \|b\| \qquad a_1 \dots a_d \quad b_2 \dots b_d$$

$a = \alpha b, \ \alpha < 0$; $a = \alpha b, \ \alpha > 0$

- dirn of Steepest accent

maximize $D_u[f](u) = \nabla f(u)^T u$

- descent dirns

$f: \mathbb{R}^d \to \mathbb{R}$

what are valid dirc^n u ?

$u: D_u[f](u) < 0$

descent dirc^n : $\{u \in \mathbb{R}^d : \nabla f(u)^T u < 0\}$

- If $f(x)$ is minimised at $u \implies \nabla f(u) = 0$

$\{u : \nabla f(u) = 0\} \implies$ Critical Point

Week-3

L3.1 4 Fundamental Subspaces

- Column space $C(A)$

$$A = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix} \quad C(A) = \text{span}(u_1 \cdots u_n)$$

solving $Ax = b$: $b \in C(A)$

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

4 eqn & 3 unknowns

col 3 = col 1 + col 2

$C(A) = \text{span}(\text{col}1, \text{col}2)$

→ 2D subspace of R^4

- Null space $N(A)$

$N(A) = \{ x \mid Ax = 0 \}$

A = as before → LC of columns of A should give 0 vector

$$\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \in N(A)$$

$N(A)$ is a line in R^3

→ If A is invertible, $N(A)$ has '0' only & $C(A)$ is the whole space

$Ax = b$ has a unique soln, $x = A^{-1}b$

Else, $N(A) = x_n \neq 0$ then, more solns

→ Gaussian elimination to get $N(A)$

$$\begin{bmatrix} 1 & 2 & 2 & 2 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 8 & 10 \end{bmatrix} \longrightarrow \begin{bmatrix} \textcircled{1} & 2 & 2 & 2 \\ 0 & 0 & \textcircled{2} & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$x_2 = 1$, $x_4 = 0$, $x_3 = 0$, $x_1 = -2$

$$\begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ -2 \\ 1 \end{bmatrix} \in N(A)$$

Rank = no. of pivot col = 2

Nullity = no. of free var. = 2

Rank = $\dim(C(A))$, Nullity = $\dim(N(A))$

A has n col. ⇒ rank + nullity = n

- Row space : col space of A^T → span of rows of A

col. rank = $\dim(C(A))$ $\quad R(A) = C(A^T)$

row. rank = $\dim(R(A))$

col. rank = row. rank

- Left Null space $N(A^T)$

$$m \times n \rightarrow \underbrace{\begin{bmatrix} y_1 \cdots y_n \end{bmatrix}}_{N(A^T)} \begin{bmatrix} A \end{bmatrix} = \begin{bmatrix} 0 \cdots 0 \end{bmatrix}$$

- A is m x n matrix

$\dim(C(A)) + \dim(N(A)) =$ no. of col = n

$\dim(C(A^T)) + \dim(N(A^T)) =$ m = no. of rows

$A \Rightarrow N(A^T) = (1, 1, -1)$

Eg. $A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$ $\quad m = n = 2$ $\quad C(A) = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ rank = 1

$N(A) = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ $\quad R(A) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $\quad \dim(R(A)) = 1$

nullity = 1

$N(A^T) = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ $\quad \dim(N(A^T)) = 1$

L3.2 Orthogonal Vectors & Subspaces

- Length of a vector
  $\|x\|^2 = x_1^2 + x_2^2$
  same for general
- Orthogonal → $x \perp y$ if $x^T y = 0$ (inner prod.)
  $$\boxed{x^T y = \sum_{i=1}^{n} x_i y_i}$$
  - 0 is orthogonal to every x
  - if $(u_1 \ldots u_n)$ are mutually orthogonal, then this is a linear indep. set
- Orthonormal → orthogonal + their unit lengths are 1
  (u, u) " if $u^T u = 0$ & $\|u\| = \|v\| = 1$
- Orthogonal subspaces
  V, V "
  if $x^T y = 0 \quad \forall x \in U, y \in V$
- Orthogonality w.r.t. 4 fundamental spaces
  ① $R(A) \perp N(A)$
  ② $C(A) \perp N(A^T)$

L3.3 Projections

- projection onto a line
  $p = \hat{x}\, a$
  $\hat{x} = \frac{a^T b}{a^T a} \rightarrow \boxed{p = \left(\frac{a^T b}{a^T a}\right) a}$
- Cauchy - Schwarz Inequality
  $|a^T b| \leq \|a\| \|b\|$

Eg. $a = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ $P = \frac{a a^T}{a^T a} = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$

$= \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$

→ P is symmetric
→ P^2 is P
→ $C(P) \Rightarrow a = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ $N(P) \Rightarrow$ plane orthogonal to a
Rank $r(P) = 1$

L3.4 Least sq. & projecⁿ onto a subspace

- $2x = b_1$, $3x = b_2$, $4x = b_3$; sys. is solvable if b is on line $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$
  minimize avg. error
  $E^2 = (2x - b_1)^2 + (3x - b_2)^2 + (4x - b_3)^2$
  $\frac{dE^2}{dx} = 0$
  $\Rightarrow \hat{x} = \frac{2b_1 + 3b_2 + 4b_3}{4 + 9 + 16} = \frac{a^T b}{a^T a}$

→ Taking derivative & finding the minima turns out to be as same as performing an projecⁿ

- $Ax = b$, A is $m \times n$ $m > n$
  → projecⁿ of b onto $C(A)$
  $\boxed{A^T A \hat{x} = A^T b}$ → this leads to $\hat{x}$ that minimizes $\|Ax - b\|^2$

→ columns of A are linearly indep.
Then, $A^T A$ is invertible

→ $b \in C(A)$ i.e., $b = Ax$
$p = b$

→ $b \in N(A^T)$
$p = 0$

→ A is eq. & invertible $\Leftrightarrow C(A) = R^n$
$p = b$

→ A is rank 1
$\hat{x} = \frac{a^T b}{a^T a}$

- Projection matrix,

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

symmetric $P^T = P$; $P^2 = P$

So, projection matrix is sym. & satisfies $P^2 = P$, as the converse is also true.

L 3.5 Example of Least Squares

- 1D

$(x_1, b_1) \dots (x_m, b_m)$

$b_i = \theta x_i + \theta' \quad i = 1 \dots m$ — (1)

$A\theta = b$ where $\theta = \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix}$

$A\theta = b$ may be inconsistent

∴, minimize $E^2 = \|b - A\theta\|^2$.

$(\hat{\theta}', \hat{\theta}'') = \arg\min \|b - A\theta\|^2$

Eg. $\begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \theta' \\ \theta'' \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$ not consistent since, $b \notin C(A)$

$A^TA\hat{\theta} = A^Tb$

$A^TA = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}$ $\quad A^Tb = \begin{bmatrix} 6 \\ 5 \end{bmatrix}$

$A^TA\hat{\theta} = A^Tb$

$\theta' = \frac{4}{7}$ & $\theta'' = \frac{9}{7}$

$\hat{\theta} = \begin{bmatrix} 4/7 \\ 9/7 \end{bmatrix}$

Line → $\frac{4}{7}x + \frac{9}{7}$

Projections : $P_1 = \frac{4}{7}(+1) + \frac{9}{7} = \frac{5}{7}$,

$P_2 = \frac{13}{7}$ & $P_3 = \frac{17}{7}$

- original data is not on line, $E^2 > 0$

$E^2 = \|b - A\hat{\theta}\|^2$, $e = \left(\frac{-2}{7}, \frac{-6}{7}, \frac{4}{7}\right)$
$= e^2$

Week 4

L4.1 Linear & Polynomial Regression

- $\{(x_1, y_1) \cdots (x_n, y_n)\}$ $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

$$L(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T \theta - y_i)^2$$

Minimize L : define $A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$ ← Feature matrix

so, $L(\theta) = \frac{1}{2} (A\theta - y)^2$

$\Rightarrow (A^T A)\theta = (A^T y) \Rightarrow$ least sq. eq. soln

$\theta = (A^T A)^{-1} A^T y$ ⇒ if A is full rank or $A^T A$ is invertible

- Maximum likelihood & least sq —

$x \rightarrow$ Model $\rightarrow y = \theta^T x + \varepsilon$ ↳ zero-mean noise

$D = \{(x_i, y_i), i = 1 \cdots n\}$ iid $\varepsilon \sim$ gaussian with mean & var $(\frac{1}{\beta})$

Max. likelihood

$$L(\theta) = \prod_{i=1}^{n} \frac{\sqrt{\beta}}{\sqrt{2\pi}} \exp\left(-\frac{\beta}{2} (y_i - \theta^T x_i)^2\right)$$

- Polynomial regression

1 D data

Transformed features :

$$\phi(x) = \sum_{j=0}^{m} \theta_j \phi_j(x), \quad \phi_j(x) = x^j$$

$$\hat{y}(x) = \theta^T \phi(x)$$

using these features, perform linear regression

$$(A^T A)\theta = A^T y$$

$$L(\theta) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T \theta - y_i)^2 + \lambda \|\theta\|^2$$

L4.2 Eigenvalues & Eigenvectors

- ordinary differentiated motion

- single eqn $\frac{du}{dt} = au$

soln $\Rightarrow u(t) = e^{at} u(0)$

- is unstable if $a > 0$
- is neutral stable if $a = 0$
- is stable if $a < 0$

- 2 D — parameter

$v(t) = e^{\lambda t} y$, $w(t) = e^{\lambda t} z$

or $u(t) = e^{\lambda t} x$, $x = \begin{bmatrix} y \\ z \end{bmatrix}$

$\boxed{Ax = \lambda x}$ ⇒ eigenvalue eqn

- For a matrix A, λ is an eigenvalue & $x \neq 0$ is an eigenvector if $Ax = \lambda x$.

If x is an eigenvector, then A either stretches / shrinks it, but A does not Δ the direction of x.

if $\lambda = 0$: $Ax = 0 \Rightarrow x \in N(A)$

Eg. $B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $x = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

- Finding the eigenvalues :

$Ax = \lambda x \Rightarrow (A - \lambda I)x = 0$

$\boxed{\det(A - \lambda I) = 0}$

↳ charac. poly. of matrix A

If A is $n \times n$, the deg(charac. poly.) n

"n" roots of charac. poly. = eigenvalues

For eigenvectors $x \in N(A - \lambda I)$

Eg. $A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ $\det(A - \lambda I) = 0$

$\lambda^2 - 6\lambda + 8 = 0$: $\lambda_1 + \lambda_2 = 6$, $\lambda_1 \lambda_2 = 8$

$\lambda_1 = 4, \lambda_2 = 2$

$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

- Sym. matrix has real eigenvalues.
- a 2×2 matrix where we do not get independent eigenvectors

$x \in N(A) \iff Ax = 0x$

L4.3 Diagonalizing ⁿ of matrix

- A matrix A is diagonalizable if it is invertible matrix S such that

$S^{-1}AS = \Lambda$ ← diagonal matrix

- This is tied to having enough independent eigenvectors
- if λ_1, λ_2 are the eigenvalues with corresp. eigenvectors x_1, x_2 & $\lambda_1 \neq \lambda_2$, then $\{x_1, x_2\}$, is a linearly indep. set

and extension: if $\lambda_1 \dots \lambda_n$ are distinct, then eigenvectors are linearly indep.

Eg. $A = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$ $A - \lambda I = \begin{bmatrix} 1-\lambda & 4 \\ 2 & 3-\lambda \end{bmatrix}$

Is A diagonizable?

Yes, it has 2 distinct eigenvalues

$\lambda^2 - 4\lambda - 5 = 0$

$\lambda_1 = 5, \lambda_2 = -1$

$\lambda_1 = 5$ $A - \lambda_1 I = \begin{bmatrix} -4 & 4 \\ 2 & -2 \end{bmatrix}$ $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\lambda_2 = -1$ $A - \lambda_2 I = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$ $x_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

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

$\Rightarrow \boxed{S^{-1}AS = \Lambda}$ ✓

- $S^{-1}AS = \Lambda$ or $A = S\Lambda S^{-1}$ Is S Unique? No
- $A = S\Lambda S^{-1}$
- Powers of A:

$A^2 x = A(Ax) = \lambda Ax = \lambda^2 x$

$S^{-1}A^2 S = \Lambda^2 \Rightarrow$ Yes in

The vectors for a general $k \geq 1$

- Not all matrices are diagonizable

L4.4 Solving Fibonacci Seq. using diagonalizing ⁿ

Fibonacci $F_{k+2} = F_{k+1} + F_k$ What is F_{100}?

$u_k = A^k u_0$ sol ⁿ to $u_{k+1} = A u_k$

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

$\lambda_1 = \frac{1+\sqrt{5}}{2}$, $\lambda_2 = \frac{1-\sqrt{5}}{2}$

$x_1 = \begin{bmatrix} \lambda_1 \\ 1 \end{bmatrix}$, $x_2 = \begin{bmatrix} \lambda_2 \\ 1 \end{bmatrix}$

$F_k \approx \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^k$

$F_{100} \approx \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^{100}$

$= \frac{1}{\sqrt{5}} (1.618)^{100}$

helps to understand linear recurrence relation — diagonaliza

L45 Orthogonally diagonalizable matrices

- A is a real sym. matrix, then
  1. eigenvalues of A are real
  2. eigenvectors corresp. to diff. eigenvalues are linearly indep.
  3. A is orthogonally diagonizable

$A = Q \lambda Q^T \quad Q^T Q = I$

- any real matrix is not necessarily diagonalizable, but a real sym. is is diagonalizable

Week-5

L5.1. Complex Matrices

- i.e. $\mathbb{C}^n$: complex counterpart of $\mathbb{R}^n$
- i.e. $x_1, \ldots x_m \in \mathbb{C}^n$
- Addⁿ: $(a+ib) + (c+id) = (a+c) + i(b+d)$
- Multiplicⁿ: $(a+ib)(c+id) = (ac-bd) + i(bc+ad)$
- complex conjugate of $(a+ib)$ is $(a-ib)$
- Linear combinⁿ: $c_1 x_1 + c \ldots + c_k x_k = 0$
- In $\mathbb{R}^n$, $\|x\|^2 = x^T x$ (length, inner product)
- $x \cdot y = \bar{x}^T y = \bar{x}_1 y_1 + \ldots + \bar{x}_n y_n$

  Also, $\bar{x}^T y \neq y^T x$
- Length of a complex vector

  For $x \in \mathbb{C}^n$, $\|x\| = \bar{x}^T x$

  $\rightarrow x \cdot y = \overline{y \cdot x}$

  $\rightarrow x \cdot (cy) = c(x \cdot y)$

  $\rightarrow (cx) \cdot y = \bar{c}(x \cdot y)$

  } from def^n of inner prod.

  $x \cdot y = \bar{x}^T y = \sum_{i=1}^{n} \bar{x}_i y_i$
- Conjugate transpose:

  A^* = conj. transpose of A

  $A^* = \bar{A}^T = \overline{A^T}$
- Real matrix A, $A^* = A^T$ Real case —

  $\rightarrow (A^*)^* = A$ | $\rightarrow (A^T)^T = A$

  $\rightarrow (AB)^* = B^* A^*$ | $\rightarrow (AB)^T = B^T A^T$
- Inner prod, $\boxed{x \cdot y = x^* y}$
- A matrix is Hermitian if $A^* = A$,

  These matrices are similar as similar matrices in complex vector space

L5.2 Hermitian matrices

Eg. $A = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}$ $\quad A^* = \bar{A}^T = \begin{bmatrix} 2 & 3+3i \\ 3-3i & 5 \end{bmatrix}^T = A$

- Diagonal entries of hermitian matrix are real
- Properties —

① If A is Hermitian, all eigenvalues are real.

Eg. $|A - \lambda I| = \begin{vmatrix} 2-\lambda & 3-3i \\ 3+3i & 5-\lambda \end{vmatrix} \Rightarrow \lambda^2 - 7\lambda - 8 = (\lambda - 8)(\lambda + 1)$

$\lambda = 8, -1$

② If A is hermitian, then eigenvectors corresponding to diff. eigenvalues, are orthogonal ⇒ thus, linearly independent

$Ax = \lambda_1 x$ $\quad Ay = \lambda_2 y$ $\quad \lambda_1 \neq \lambda_2 \Rightarrow x \cdot y = \bar{x}^T y = 0$

Eg. $(A - 8I)x = \begin{bmatrix} -6 & 3-3i \\ 3+3i & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \quad x = \begin{bmatrix} 1 \\ i+1 \end{bmatrix}$

$(A + I)y = 0 \Rightarrow y = \begin{bmatrix} 1-i \\ -1 \end{bmatrix}$

$x \cdot y = \begin{bmatrix} 1 & 1-i \end{bmatrix} \begin{bmatrix} 1-i \\ -1 \end{bmatrix} = 0$

- The equivalent of hermitian matrix is real symmetric matrices.
- All real symmetric are hermitian.
- If eigenvalues are distinct, the A is diagonizable

L5.3 Unitary Matrices

A matrix is unitary, if its square & has orthonormal columns.

Real: $Q^T Q = I \Rightarrow Q$ is orthogonal case $Q^{-1} = Q^T$

Complex: $U^* U = I \Rightarrow U$ is unitary $U^{-1} = U^*$

Eg. $U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

- Properties — $U \rightarrow$ Unitary matrix

① "Length unchanged": $\|Ux\| = \|x\|$

② Eigenvalues of U have absolute value 1
then, $|\lambda| = 1$

③ Eigenvectors corresponding to different eigenvalues of a unitary matrix U are orthogonal.

- For a hermitian matrix A, we can find unitary matrix U

$A = U \lambda U^*$, $\lambda \Rightarrow$ diagonal matrix with eigenvalues of A

L5.4 Diagonalisation of Hermitian Matrix - I

- $A^* = \bar{A}^T$

Matrix A is hermitian if $A^* = A$

Matrix U is unitary if $U^* U = I$ & U is Sq.

- A matrix 'A' is unitarily diagonizable if there exists a unitary matrix U

$A = U \lambda U^*$

- Schur's Theorem - Any $n \times n$ matrix is similar to upper triangular matrix T, there exists unitary matrix U

$$A = UTU^*$$

Eg: $A = \begin{bmatrix} 5 & 8 & 16 \\ 5 & 0 & 9 \\ -3 & -5 & -10 \end{bmatrix}$

$$p(\lambda) = -(\lambda - 1)(\lambda + 3)^2 \quad \lambda_1 = 1, \ \lambda_2 = -3$$

Eigenvector $z_1 = \begin{bmatrix} -2 \\ -1 \\ 1 \end{bmatrix}$

$$T = \begin{bmatrix} 1 & -3/\sqrt{2} & -12/\sqrt{3} \\ 0 & -3 & \sqrt{6} \\ 0 & 6 & -3 \end{bmatrix}$$

LB 5

Part 2

- Spectral Th. - A hermitian matrix "A" is unitarily diagonizable

$$U^* A U = D$$

↳ diagonal matrix with real nos.

$$U^* A U = T$$

$$T^* = U^* A^* U = U^* A U = T$$

↳ lower triangular ↳ upper triangular

Eg. $A = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}$ $\quad A^* = A$

$$p(\lambda) = (\lambda - 1)(\lambda - 4) \Rightarrow \lambda_1 = 1, \ \lambda_2 = 4$$

$$z_1 = \begin{bmatrix} -1+i \\ 1 \end{bmatrix} \qquad z_2 = \begin{bmatrix} 1-i \\ 2 \end{bmatrix}$$

- Spectral Th. - A real symmetric matrix A is orthogonally diagonizable, there exists a matrix Q

$$Q^T A Q = D, \quad Q^T Q = I$$

↳ diag. matrix with real entries

- Hermitian $\Rightarrow$ unitary diagonizable

$\not\Leftarrow$

but, the converse is not true

Week-6

L6.1 SVD (Singular Value Decomposition)

- A is $n \times n$ real symmetric matrix
(i) All eigenvalues are real
(ii) A is orthogonally diagonalizable

$$A = Q \Lambda Q^T$$

Every matrix can't be diagonalized, but any real $n \times n$ matrix can be decomposed in the SVD form.

$$A = Q_1 \Sigma Q_2^T$$

$m \times n$ $\quad$ $m \times m$ $\quad$ $n \times n$

Q_1 & Q_2 are orthogonal

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}, \text{ where } D = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_m \end{bmatrix}$$

(1)
- $A = Q_1 \Sigma Q_2^T$
- $AA^T = Q_1 \Sigma \Sigma^T Q_1^T$

eigen decomposition of AA^T into Q_1

(2) $A^T A = Q_2 \Sigma^T \Sigma Q_2^T$

eigen decomposition of $A^T A$ into Q_2

L6.2 Eg. of SVD

Eg. $A = \begin{bmatrix} \sqrt{2} & 1 \\ 0 & \sqrt{2} \end{bmatrix}$ Find SVD?

Is matrix A diag.? No

Eigenvalue = $\sqrt{2}, \sqrt{2}$ (repeated)

$A - \sqrt{2} I = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ eigenvector

& they aren't linearly indep.

Find SVD of $\begin{bmatrix} \sqrt{2} & 1 \\ 0 & \sqrt{2} \end{bmatrix}$

(i) Eigenvalues & eigenvec. of $A^T A$

(ii) $\sigma_1 y_1 = A x_1$, $\sigma_2 y_2 = A x_2$

to find y_1, y_2

(iii) $Q_1 = \begin{bmatrix} | & | \\ y_1 & y_2 \\ | & | \end{bmatrix}$, $Q_2 = \begin{bmatrix} | & | \\ x_1 & x_2 \\ | & | \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}$

(i) $A^T A = \begin{bmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 3 \end{bmatrix}$ $\lambda_1 = 4, \lambda_2 = 1$

$\sigma_1 = \sqrt{\lambda_1} = 2$

$\sigma_2 = \sqrt{\lambda_2} = 1$

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

Eigenvectors of $A^T A \Rightarrow \begin{bmatrix} -2 & \sqrt{2} \\ \sqrt{2} & -1 \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$ &

$\begin{bmatrix} 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} \begin{bmatrix} \sqrt{2} \\ -1 \end{bmatrix}$,

Normalize, $x_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ \sqrt{2} \end{bmatrix}$ $x_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} \sqrt{2} \\ -1 \end{bmatrix}$

$Q_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & \sqrt{2} \\ \sqrt{2} & -1 \end{bmatrix}$

(ii) $y_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} \sqrt{2} \\ 1 \end{bmatrix}$, $y_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -\sqrt{2} \end{bmatrix}$

(iii)

L6.3 +ve definiteness

- $f(x,y) = 2x^2 + 4xy + y^2$

$\frac{\partial f}{\partial x} = 4x + 4y = 0$ $\quad \frac{df}{dy} = 4x + 2y = 0$

$(x,y) = (0,0)$ both $\frac{df}{dx}, \frac{df}{dy}$ are 0

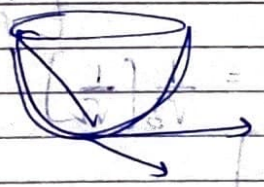
Hence, $(0,0)$ is a stationary point of f

$\frac{d^2f}{dx^2} = 4 \quad \frac{d^2f}{dx\,dy} = 4, \quad \frac{d^2f}{dy^2} = 2$

$\Rightarrow f$ has a minima at $(0,0)$

- Every quadratic func^n of the form $(ax^2 + 2bxy + y^2)$ has a stationary point at $(0,0)$

- A func^n f that vanishes at $(0,0)$ and is strictly +ve at other pts is called positive definite. $\Rightarrow f > 0$



$(ax^2 + 2bxy + cx^2)$

Conditions —

① if $f > 0$, then $a > 0$

② if $f > 0$, then $c > 0$

Further,

$$f(x,y) = \underbrace{a\left(x + \frac{b}{a}y\right)^2}_{\geq 0} + \underbrace{\left(c - \frac{b^2}{a}\right)y^2}_{\geq 0}$$

③ if $f > 0$, $\quad ac > b^2$

Combining all these 3 conditions,

$f(x,y) = ax^2 + 2bxy + cy^2$ is +ve definite iff $a > 0$ & $ac > b^2$

- if $ac = b^2$, then | +ve semi-def. $a > 0$ | -ve semi-def. $a < 0$

- we have a saddle point at $(0,0)$ if $ac < b^2$

- $ax^2 + 2bxy + cy^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$

$v = \begin{bmatrix} x \\ y \end{bmatrix} \quad A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$

$\Rightarrow ax^2 + 2bxy + cy^2 = v^T A v$

① $f(x,y) = 2x^2 + 4xy + y^2 \Rightarrow$ saddle pt. at origin

② $f(x,y) = 2xy \Rightarrow$ saddle at origin

L6.4 +ve Definite Matrices

- $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is +ve df. if $a > 0$, $ac - b^2 > 0$

corres. func^n is $f(v) = v^T A v \quad v = \begin{bmatrix} x \\ y \end{bmatrix}$

- if $a > 0$, $ac - b^2 > 0$, then both eigenvalues are +ve

$ac - b^2 = \det(A) = \lambda_1 \lambda_2 > 0$

$\text{Trace}(A) = \lambda_1 + \lambda_2 = a + c > 0$

- a real sym. matrix is +ve df. if

(i) $v^T A v > 0 \quad \forall\, v \in R^n, \; v \neq 0$

(ii) all eigenvalues of A are > 0

Week-7

L7.1 Principal Comp. Analysis

- PCA (Idea) – Project given data onto a lower dimensional subspace such that –
  (i) reconstruction error is minimized
  (ii) variance of proj. data is maximized

$$x_i = \sum_{j=1}^{d} (x_i^T u_j) u_j$$

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i - \tilde{x}_i\|^2$$

$$\tilde{x}_i = \sum_{j=1}^{m} z_{ij} u_j + \sum_{j=m+1}^{d} (x_i^T u_j + \beta_j) u_j$$

L7.2 (Contd.)

Lagrangian

$$L(u, \lambda) = u^T C u + \lambda (1 - u^T u)$$

$$D_u L(u, \lambda) = 0 \Rightarrow Cu = \lambda u \quad (\text{eqn})$$

$$u^T C u = \lambda$$

$$C = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

PCA

① data : $\{x_1 \dots x_n\}$

② Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$C = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

③ find eigenvalues with corresponding eigenvectors

④ projected data

$$\tilde{x}_i = \sum_{j=1}^{m} (x_i^T u_j) u_j + \sum_{j=m+1}^{d} x_i^T u_j$$

L7.3 PCA as maximizing variance

- x_i projection along u is $(x_i^T u) u$

$D = \{x_1 \dots x_n\}$ $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$x_i \in \mathbb{R}^d$

$(x_i^T u) u$, mean $= (\bar{x}^T u) u$

Variance is $(x_i^T u - \bar{x}^T u)^2$

$$\frac{1}{n}\sum_{i=1}^{n} (x_i^T u - \bar{x}^T u)^2 \rightarrow \text{maximize this over } u$$

A calculus argument,

$$\max_u \frac{u^T C u}{u^T u}$$

In vector form,

$$u^T u \, Cu = (u^T C u) u \Rightarrow Cu = \left(\frac{u^T C u}{u^T u}\right) u$$

or $Cu = \lambda u$

To maximize $\frac{u^T C u}{u^T u}$, choose λ to be the largest eigenvalue of C & let u be the corresponding eigenvector.

L7.4 PCA in higher dimensions

- PCA req. finding the eigenvectors of

$$C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T,$$

a $d \times d$ matrix $\quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\text{rank}(C) \leq n$

$\therefore \lambda_i$ is an eigenvalue of $\frac{1}{n}AA^T$

So, instead of working with $d \times d$ matrix $C = \frac{1}{n}A^TA$ to find its eigenvalues/vectors,

it is enough to find eigenvalues/vectors of $\frac{1}{n}AA^T$, which is a $n \times n$ matrix.

Week-8

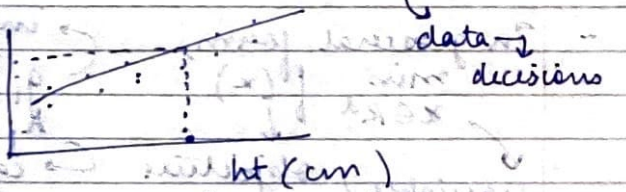
L8.1 Pillars of ML

- noise / uncertainty

P_1 Linear algebra ← struc./relationship btw. data

P_2 Probability

P_3 Optimization → data → decisions

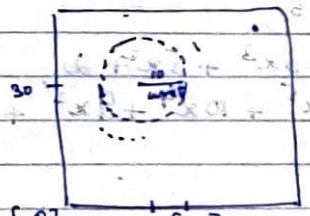


- ht & wt. are linearly related

L8.2 Intro to Optimizn

- we care about finding the 'best' classifier
  - ↳ least 'lost'
  - ↳ maximum 'reward'

Eg.



- a cone at $\begin{bmatrix}20\\20\end{bmatrix}$
- cone is held to 10 units radius rope
- a ⊥ force passes thru $\begin{bmatrix}25\\0\end{bmatrix}$

- grass on the field at $\begin{bmatrix}40\\40\end{bmatrix}$

"How close the cone get to the grass?"

- dist. from grass $\begin{bmatrix}40\\40\end{bmatrix}$ say the cone is $\begin{bmatrix}x_1\\x_2\end{bmatrix}$ $\quad d = (x_1-40)^2 + (x_2-40)^2$

↓ make this as small as possible

$$\min_{\begin{bmatrix}x_1\\x_2\end{bmatrix}} (x_1-40)^2 + (x_2-40)^2$$

Rope restric^n

$$(x_1 - 20)^2 + (x_2 - 30)^2 \leq 10^2$$

Fence restric^n

$$x_1 \leq 25$$

- In general form,

$$\min_{x \in R^d} f(x)$$

(x: variable / parameter; f: objective)

$g_i(x) \leq 0 \quad \forall i = 1 \dots k$ ← inequality constraint

$h_j(x) = 0 \quad \forall j = 1$ ← equality constraint

L8.3 Solving an unconstrained opt. problem – 1

Unconstrained Opt.

$$\min_{x \in R} (x-5)^2$$

Sol^n $\longrightarrow$ $x = 5$

- $\min_{x \in R} 3x^6 + 2x^5 + 3x^3 + 5x^2 + 2$

$\frac{dx}{dx} = 18x^5 + 10x^4 + 9x^2 + 10x = 0$

- $\min_{x \in R} (x-5)^2$

⇒ Start with $x_0 \in R$ (arbitrary)

for $t = 1, \dots T$

- update $x_{t+1} = x_t + d$ → direc^n

'good direc^n' 'close to min'

given x, what is a good direc^n?

if $x > 5$ then ← $d < 0$

if $x < 5$ then → $d > 0$

⇒ direc^n d will depend on x, i.e, d must be a func^n of x.

$f(x) = (x-5)^2$ $x_1 = 8$

$f'(x) = 2(x-5)$ ⇒ $2(8-5) = 6$

⇒ $x > 5$ ⇒ $f'(x) > 0$

⇒ $x < 5$ ⇒ $f'(x) < 0$

Choose, $d = -f'(x)$

$f(x) = (x-5)^2$ $f'(x) = 2(x-5)$ $d = -f'(x)$

⇒ but, in the value oscillates from 0 to 10 but atleast the direc^n is giving good

L8.4 Part -2

$$x_{t+1} = x_t - \eta_t f'(x_t)$$

($f'(x_t)$: direc^n; η_t: step size (scalar quantity +ve))

Now, how to choose a step size

1st attempt $\eta_t = 1, \eta_1 = \frac{1}{2} \dots \eta_t = \frac{1}{2^t}$ ... X

≈ 2

2nd attempt $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4} \dots$

$\eta_t = \frac{1}{t+1}$ → good, step size seq

$\approx \infty$

L8.5 Basic Algo: Gradient Descent

$$\min_{x \in R} f(x)$$

Algo - Gradient Descent (1st Order Algo.)

- Initialize at $x_0 \in \mathbb{R}$

for $t = 1, 2 \ldots$

$$x_{t+1} = x_t - \eta_t(-f'(x_t)) \quad \text{where } \eta_t = \frac{1}{t+1}$$

- Properties -

① if $\eta_t = \frac{1}{t+1}$, the algo converges

② gradient descent converges to local minima

- there few funcns in which local minima $\equiv$ global minima
  - convex functions

L8.6 Gradient descent & taylor series

$$x_{t+1} = x_t + \eta_t(-f'(x_t))$$

↳ what is so spl. abt. this?

- Taylor series

$$f(\underbrace{x+\eta d}_{\tilde{x}}) = f(x) + \eta d\, f'(x) + \eta^2 \frac{d^2}{2} f''(x) + \ldots$$

- The local info. gives global info.

- $f(x+\eta d)$
  - d: direcn
  - η: small +ve step size

evaluations are all at x.

the further higher order terms

$$f(x+\eta d) \approx f(x) + \eta d\, f'(x)$$

$$\Rightarrow f(x+\eta d) - f(x) \approx \eta d\, f'(x)$$

func^n evalua^n at updated pt. along direcn d — func^n evalua^n at current pt.

want to choose a direcn s.t. we want this to be negative

η: small +ve const.

$$\eta d\, f'(x) < 0 \Rightarrow d f'(x) < 0$$

$$\therefore d f'(x) = -(f'(x))^2 < 0$$

L8.7 Gradient descent for multivariate funcns

- let's go to higher dimensions
- $f(x_1, x_2) = x_1^2 + 4x_2 + 8x_2^2$

Derivatives $\Leftrightarrow$ Gradient (vector of partial derivatives)

$$\nabla f\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = \begin{bmatrix} \frac{\delta f}{d x_1}\Big|_{x_1 = a} \\ \frac{\delta f}{d x_2}\Big|_{x_2 = b} \end{bmatrix}$$

Eg. $\nabla f\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}\right) = \begin{matrix} 2x_1 \\ 4 + 16x_2 \end{matrix}\Big|_{\substack{x_1=1 \\ x_2=3}} = \begin{bmatrix} 2 \\ 52 \end{bmatrix}$

- Pract. eq. $d(x_1, x_2) = (x_1 - 40)^2 + (x_2 - 40)^2$

$$\nabla d[x_1, x_2] = \begin{bmatrix} 2(x_1 - 40) \\ 2(x_2 - 40) \end{bmatrix}_{\begin{bmatrix} 5 \\ 2 \end{bmatrix}} = \begin{bmatrix} -70 \\ -76 \end{bmatrix}$$

$$-\nabla f\left(\begin{bmatrix} 5 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} 70 \\ 76 \end{bmatrix}$$

- Gradient descent : $\vec{x}_{t+1} = \vec{x}_t + \eta(-\nabla f(x_t))$

L8.8 Taylor series in higher dimensions

- Higher order taylor series,

$$f(x+\eta d) = f(x) + \eta\, d^T \nabla f(x) + \dots$$

$$f(x+\eta d) - f(x) \approx \eta\, d^T \nabla f(x)$$

$\rightarrow d^T \nabla f(x) < 0$

$\Rightarrow d = -\nabla f(x)$

- $-\nabla f(x)$ gives the steepest descent
- G.D is also called the steepest descent.

$$\min_{x \in R^d} f(x) \quad g(x) \leq 0$$

Week 9

L9.1 Constrained Optimiza^n - 1

- $\min_x f(x)$ s.t. $g(x) \leq 0$

→ x^* solves this, how to check whether it is indeed optimal

① $g(x^*) \leq 0$

② no "descent direc^n" should be a "feasible direc^n"

- "descent direc^n" → any direc^n that reduces the func^n value
- "feasible direc^n" → any direc^n that takes to a point that is feasible (for same step-size) → satisfying constraint $g(x) \leq 0$

- no descent dir^n is feasible dir^n
- $d^T \nabla f(x) < 0 \Rightarrow d$ is a descent direc^n
- If x^* satisfies $g(x^*) \leq 0$, then any descent direc^n for g is also a 'feasible direc^n'
- x^* can't be optimal for the $\nabla f(x^*)$, $\nabla g(x^*)$ dearen above

L9.2 Part - 2

- Necessary condition for optimality of x^*

$$\nabla f(x^*) = -\lambda \nabla g(x^*)$$

λ → Lagrange multiplier

Can be any arbitrary scalar

L 9.3 Method of Lagrange Multiplier, Projected Gradient descent

- method of lagrange multipliers

① $g(x^*) = 0$

② $\nabla f(x^*) = -\lambda \nabla g(x^*)$ for some λ

Eg. $f(x_1, x_2) = x_1^2 + 2x_2 + 4x_2^2$

$g(x_1, x_2) = x_1^2 + x_2^2 - 1$

$$\nabla f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_1 \\ 2+8x_2 \end{bmatrix} ; \quad \nabla g\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

$$\begin{bmatrix} 2x_1 \\ 2+8x_2 \end{bmatrix} = -\lambda \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

$2x_1 = -\lambda 2x_1$ — ①

$2+8x_2 = -\lambda(2x_2)$ — ②

from ① $\Rightarrow$ either $x_1 = 0$ or $\lambda = -1$

Case 1: $\lambda = -1$

$2+8x_2 = -(-1)2x_2 \Rightarrow x_2 = -1/3$

then $x_1 = \left\{ \frac{+\sqrt{8}}{3}, \frac{-\sqrt{8}}{3} \right\}$

Case 2: $x_1 = 0$

then $x_2 = 1$ or -1

To find $\min_x f(x)$ subject to $g(x) = 0$, substitute each pot. soln into f

$f(x) = x_1^2 + 2x_2 + 4x_2^2$

$f\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 6$ ✓ $\quad f\begin{bmatrix} \sqrt{8}/3 \\ -1/3 \end{bmatrix} = -\frac{4}{3}$

$f\begin{bmatrix} 0 \\ -1 \end{bmatrix} = 2$ $\quad f\begin{bmatrix} -\sqrt{8}/3 \\ -1/3 \end{bmatrix} = -2/3$

- It may not be true always possible to solve the sys. of eqⁿ that satisfy the Lagrange eqⁿ.

- Project Gradient descent

x_0 for $t = 1, \dots T$

$x_{t+1} = x_t - \eta \nabla f(x_t)$

Projection operator

$$\Pi(x) = \min_{y \in S} \|x - y\|_2^2$$

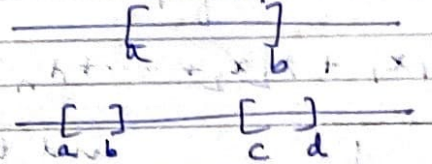
$S \leftarrow \{y : g(y) \le 0\}$ — In convex constant sets

$$x_{t+1} = \Pi\left(x_t - \eta \nabla f(x_t) \right)$$

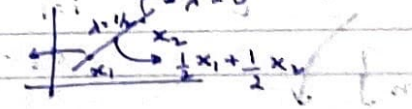
gradient

projection step

L 9.4 Intro. to Convexity



- A set $S \subseteq R^d$ is a convex set if $\forall x_1, x_2 \in S$, then $\lambda x_1 + (1-\lambda) x_2 \in S$ $\forall \lambda \in [0,1]$



- ~~Half spaces~~ Hyperplanes are convex sets

$S \subseteq R^d \quad \{x : w^T x = b\}$

- Half spaces are convex

$S \subseteq \mathbb{R}^d \quad \{x \in \mathbb{R}^d : w^T x \leq b\}$.

L 9.5 Prop. of Convex Sets

- Intersecⁿ of convex sets

$S_{1,2} = S_1 \cap S_2 = \{x : x \in S_1, x \in S_2\}$

→ $S_{1,2}$ is also convex

Eg. $\{x : Ax = b\}$ $A \in \mathbb{R}^{m \times d}$ $b \in \mathbb{R}^m$
$\in \mathbb{R}^d$

is it convex?

S is convex { intersecⁿ of hyperplanes is also a convex }

- Convex combinaⁿ

Let $S = \{x_1 \dots x_n\} \subseteq \mathbb{R}^d$. $z \in \mathbb{R}^d$ to be a convex combinaⁿ of pts. in S if $\exists \lambda_1 \dots \lambda_n$

$$z = \lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_n x_n$$

and they sum up to 1

- Convex Hull (S) = $\{z : z = \sum_{i=1}^{n} \lambda_i x_i$ for some $\lambda_1 \dots \lambda_n \geq 0$, $\sum \lambda_i = 1\}$

↓ this in itself is a convex set

$\{x_1 \dots x_n\}$

- Convex hull $\{x_1 \dots x_n\}$ as the intersecⁿ of all convex sets that contain $\{x_1 \dots x_n\}$

- Euclidean balls in $\mathbb{R}^d$

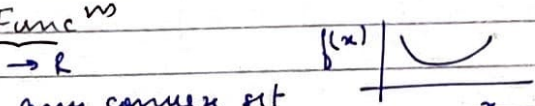
$B = \{x : \|x\|_2 \leq \theta\}$ → This is also a convex set

$\sqrt{\sum_{i=1}^{d} x_i^2}$

L9.6 Convex Funcⁿs

$f : \mathbb{R}^d \to \mathbb{R}$

↳ any convex set



$$epi(f) = \left\{ \begin{bmatrix} x \\ z \end{bmatrix} \in \mathbb{R}^{d+1} : z \geq f(x) \right\}$$

- A funcⁿ $f : \mathbb{R}^d \to \mathbb{R}$ is a convex funcⁿ if epi(f) is a convex set. ↳ $\mathbb{R}^{d+1}$

- or,

A funcⁿ $f : \mathbb{R}^d \to \mathbb{R}$ is convex iff $\forall x_1, x_2 \in \mathbb{R}^d$ and all $\lambda \in [0,1]$

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

- or,

Assume, $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable

f is convex iff

$$f(y) \geq f(x) + (y - x)^T \nabla f(x)$$

- or,

f is twice differentiable $f : \mathbb{R}^d \to \mathbb{R}$

$H \in \mathbb{R}^{d \times d}$

$H_{ij} = \frac{\delta^2 f}{\delta x_i \delta x_j}$ $f(x_1 \dots x_d) \to \mathbb{R}$

f is convex iff. H is a p.s.d matrix

↳ p.s.d ≡ semi-def. eigen(H) ≥ 0

L9.7 Prop. of Convex func^n

- If f is a convex func^n, then all local minima of f are also global minimum

$f(z) < f(x^*)$ → this is wrong
(z: global min, x^*: local min)

Week-10

L10.1 Prop. of Convex Func^n

- it continuous to last pt., there could be multiple global minimum
- Set of all global minima of a convex func^n is a convex set.
- necessary & sufficient cond^n for optimality of convex func^n

$\min_x f(x)$ — convex, differentiable

- If $f \to$ convex, differential func^n $R^d \to R$
  $x^* \in R^d$ is a global minimum of f
  iff. $\nabla f(x^*) = 0$
  and also, this convex if true.
- Add^{nal} prop. of convex func^{n}s —
  ① If $f: R^d \to R$, $g: R^d \to R$ are both convex
  then, $h(x) = f(x) + g(x)$ $\forall x$ is convex.
  ⇒ sum of convex func^{n}s is convex
  ② Compositions
  $f: R \to R$ be convex & non-decreasing func
  $g: R^d \to R$ be convex
  $h(x) = f(g(x)) \Rightarrow h := f \circ g$
  h is also convex.
  ⇒ Composi^n of convex with (convex + non dec.) is convex

③ $f: R \to R$ convex, $g: R^d \to R$ linear
$h = f \circ g \Rightarrow h$ is convex

- In general, if f & g are convex, then $h = f \circ g$ may not be convex

L10.2 App. of Optimizaⁿ in ML

- Linear regression: Dataset $\{\underbrace{x_1, \dots, x_n}_{\text{Training data}}, y_1, \dots, y_n\}$

$x_i \in \mathbb{R}^d \ \forall\, i$

$y_i \in \mathbb{R}$

Goal: $h : \mathbb{R}^d \to \mathbb{R}$ $\quad \hat{y}_{test} = h(x_{test})$, $x_{test} \in \mathbb{R}^d$

data → Alg → Model $h : \mathbb{R}^d \to \mathbb{R}$ → w → $\hat{y}_{test}$

h is linear, $h(x) = w^T x$ for some $w \in \mathbb{R}^d$

Performance: sum of square errors measure

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2$$

$$f(w) = \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2$$

$h_i(w) = (w^T x_i - y_i)^2 = f(g(w))$, $g(w) = w^T x_i - y_i$ (linear), $f(z) = z^2$ (convex)

$\Rightarrow$ hence, $f(w)$ is a convex func^n

$$f(w) = \frac{1}{2} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2$$

$$f(w) = \frac{1}{2} \| Xw - y \|_2^2$$

$$f(w) = \frac{1}{2} (Xw - y)^T (Xw - y)$$

$$\nabla f(w) = (X^T X) w - X^T y$$

Advantages → Analytical solution

Issue ← needs an 'inverse' computaⁿ — $O(d^3)$

↳ can use iterative procedures [gradient descent]

$$w^{t+1} = w^t - \eta_t \underbrace{\nabla f(w^t)}_{(X^T X) w - X^T y}$$

[does not involve an inverse]

→ Approximaⁿ of gradient by Stochastic Gradient descent:

↳ samples a small set of data points uniformly at random

↳ new dataset & compute gradient w.r.t w

$$\frac{1}{T} \sum_{t=1}^{T} w_t \longrightarrow w^*$$

L10.3 Revisiting Constrained Optimization

- unconstrained, f is convex, $\nabla f(x^*) = 0 \Rightarrow x^*$ is global opt.
- constrained

$$\min_x f(x) \quad \text{s.t.} \quad h(x) \le 0$$

- Lagrangian func^n

$$L(x, \lambda) = f(x) + \lambda h(x)$$

Fix $x \in \mathbb{R}^n$

$$\max_{\lambda \ge 0} L(x, \lambda) = \max_{\lambda \ge 0} f(x) + \lambda h(x) = \begin{cases} f(x) & h(x) \le 0 \\ \infty & h(x) > 0 \end{cases}$$

$$\min_x f(x) \quad \text{s.t.} \quad h(x) \le 0 \;=\; \min_x \left[\max_{\lambda \ge 0} L(x, \lambda) \right]$$

↳ primal problem

$$\max_{\lambda \geq 0} \left[\min_{x} L(x, y) \right] = \max_{\lambda \geq 0} \left\{ \min_{x} f(x) + \lambda h(x) \right\}$$

$$= \left[\max_{\lambda \geq 0} g(\lambda) \right]$$

↓ dual problem ; ↳ concave func^n

L 10.4 Rela^n btw. Primal & Dual prob. | KKT Cond^n

$$\min_{x} \max_{\lambda \geq 0} L(x, \lambda) \qquad \max_{\lambda \geq 0} \min_{x} L(x, \lambda)$$

Primal Dual

$$J(x) = \begin{cases} f(x) & h(x \leq 0) \\ \infty & \text{other} \end{cases}$$

For any $\lambda \geq 0$

$$L(x, \lambda) = f(x) + \lambda h(x)$$

Fix $\lambda \geq 0$

$$L(x, \lambda) \leq J(x) \quad \forall x$$

$$\max_{\lambda \geq 0} \min_{x} L(x, \lambda) \leq f(x^*)$$

$$\max_{\lambda \geq 0} [g(\lambda)] = g(\lambda^*) \leq f(x^*)$$

value at dual opt. $\leq$ value at primal opt.

Weak - Duality

- If f, h are convex, then Strong - Duality holds

$$f(x^*) = g(\lambda^*)$$

$$= \min_{x} \underbrace{f(x) + \lambda^* h(x)}$$

$$\leq f(x^*) + \lambda^* h(x^*)$$

- f, h are convex $\Rightarrow$ strong duality

x^*, y^* must satisfy

(a) $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$

(b) $\lambda^* h(x^*) = 0$

(c) $h(x^*) \leq 0$

(d) $\lambda^* \geq 0$

If x^*, y^* satisfies these conditions $\Rightarrow$ Local Optima

- KKT (Karush - Kuhn - Tucker) Conditions

$$\min f(x)$$

$$h_i(x) \leq 0 \quad \forall i = 1 \dots m$$

$$l_j(x) = 0 \quad \forall j = 1 \dots n$$

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{n} v_j l_j(x)$$

L 10.5 KKT cond^{n}s contd.

SVM (support vector machine) Optimization problem

$\min_{w,} \frac{1}{2}\|w\|^2 \longrightarrow$ quad. $\rightarrow$ convex

s.t. $w^T x_i y_i \geq 1 \quad \forall i$

↳ linear

$\Longrightarrow$ Convex

Week-11

11.1 Condi. RV

- Sample space $\rightarrow (\Omega, F, P)$
- $F \subseteq \{0,1\}^{\Omega}$
  - (i) $\Omega \in F$
  - (ii) $A \in F \Rightarrow A^c \in F$
  - (iii) $A_1, A_2 \ldots A_n \in F \Rightarrow \bigcup_{i=1}^{n} A_i \in F$
- $X : \Omega \rightarrow R$

Domain & range to be uncountable

Eg. Expt. - Waiting for bus, X : amt. of time you wait

7:00, 7:15, 7:30 - - - -

Reach the bus stop 7.10 & 7.20

- X $x \in R$

$f_x(x) = \frac{P(X \in [x, x+dx])}{dx}$ PDF

$P(X = x) = 0$

$F_x(x) : P(X \leq x)$ CDF

① $f_x(x) \geq 0$

② $\int_{-\infty}^{\infty} f_x(x)\,dx = 1$

③ $F_x(-\infty) = 0$

④ $F_x(\infty) = 1$

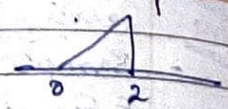
⑤ F_x is increasing

Eg.

$f_x(x) = \begin{cases} \frac{1}{15}\,(\text{min})^{-1} & x \in [0,15] \\ 0 & \text{otherwise} \end{cases}$

$F_x(x) = \begin{cases} 0 & x < 0 \\ x/15 & x \in [0,15] \\ 1 & x > 15 \end{cases}$

Eg. $f_x(x) = \begin{cases} x/2 & \text{if } x \in [0,2] \\ 0 & \text{otherwise} \end{cases}$



$$F_x(x) = P(X \le x) = \int_0^x f_x(x)\,dx = \frac{1}{2}\left[\frac{x^2}{2}\right]_0^x = \frac{x^2}{4}$$

L11.2 Conditional pdf

- (Ω, F, P) X, $A \subseteq \Omega$
- $$f_{X|A}(x) = \frac{P(X \in [x, x+dx] \mid A)}{dx}$$

Eg $f_x(x) = \begin{cases} x/2 & x \in [0,2] \\ 0 & \text{others} \end{cases}$

$A = X > 1$

$$P(A) = \int_1^2 \frac{x}{2}\,dx = \frac{3}{4}$$

$$f_{X|A}(x) = \begin{cases} 0 & \text{if } x < 1 \\ \dfrac{P(x \in [x, x+dx])}{P(A)\,dx} & \\ 0 & \text{if } x > 2 \end{cases}$$

$$= \begin{cases} 0 & \text{if } x < 1 \\ \frac{4x}{6} & \text{if } x \in [1,2] \\ 0 & \text{if } x > 0 \end{cases}$$

- funcⁿ of RV' $X \to Y = \frac{X}{2}$ or $Y = X^2$ or $Y = |X|$

Eg. $f_x(x) = \begin{cases} 1/2 & \text{if } x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$ $\quad Y = \frac{X}{2}$

$$f_Y(y) = \frac{P(x \in [y, y+dy])}{dy}$$

$$= \frac{P(x \in [2y, 2y+2dy])}{dy}$$

$\Rightarrow 0$ if $2y \notin [-1, 1]$

$\Rightarrow 2x, \frac{1}{2}$ if $y \in [-1/2, 1/2]$

Eg. $Y = |X|$ X is uniform $[-1, 1]$

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

L11.3 Expectation

- $X : \Omega \to \mathbb{R}$
- $$E[X] = \int_{-\infty}^{\infty} x \cdot f_x(x)\,dx$$

- Properties –

① $E[X+Y] = E[X] + E[Y]$

② $Y = g(x)$ $\quad E[Y] = \int_{-\infty}^{\infty} g(x) \cdot f_x(x)\,dx$

Eg. $X \sim \text{unif}[a,b]$ $\quad f_x(x) = \begin{cases} \frac{1}{(b-a)} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$

$$E[X] = \frac{(b+a)}{2}$$

Eg. $f_X(x) = \begin{cases} x/2 & \text{if } x \in [0,2] \\ 0 & \text{other} \end{cases}$

$$E[X] = \int_0^2 x \cdot \frac{x}{2}\, dx = \frac{4}{3}$$

- $\text{Var}[X] = E[(X - E(X))^2]$
$= E[X^2] - [E[X]]^2$

$\sqrt{\text{Var}[X]} = SD[X]$

- Prop -
1. $\text{Var}[X+Y] \neq \text{Var}[X] + \text{Var}[Y]$
2. $\text{Var}[aX] = a^2 \text{Var}[X]$
3. $\text{Var}[X] \geq 0$

Eg. $f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{other} \end{cases}$

$$E[X^2] = \frac{1}{3}[b^2 + a^2 + ab]$$

$$(EX)^2 = \frac{(b+a)^2}{4}$$

$$\boxed{\text{Var}[X] = \frac{(b-a)^2}{12}}$$

- X, A

$$E[X|A] = \int_{-\infty}^{\infty} x \cdot f_{X|A}(x)\, dx$$

Eg. $f_{X|A}(x) = \begin{cases} 1/5 & \text{if } x \in [0,5] \\ 0 & \text{other} \end{cases}$ $\quad E[X|A] = 2.5$

$$E[X|A^c] = 7.5$$

- $\boxed{E[X] = E[X|A] \cdot P(A) + E[X|A^c] + P(A^c)}$

$$E[X] = \frac{50}{8}$$

L11.4 Multiple RV

X, Y

- $f_{XY}(x,y) = \frac{P(X \in [x, x+dx], Y \in [y, y+dy])}{dx \cdot dy}$

↳ Joint Distribution / Density Func^n

$$F_{XY}(x,y) = P(X \leq x, Y \leq y)$$

Prop.
1. $f_{XY}(x,y) \geq 0$
2. $\int_x \int_y f_{XY}(x,y)\, dx\, dy = 1$
3. $F_{XY}(-\infty, -\infty) = 0$
4. $F_{XY}(\infty, \infty) = 1$

- $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy$
$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx$] marginal density

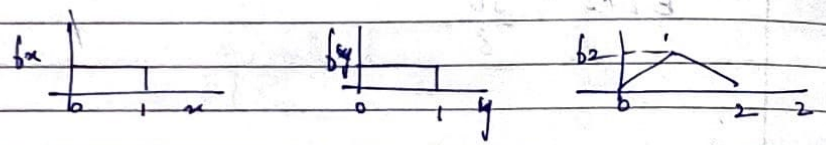
$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$] conditional
↳ $\neq 0$

L11.5 Indep. RV.

X, Y

$Z = X + Y$

$$f_Z(y) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(z-x)\, dx$$



- $Z = \max(X, Y)$

$$F_Z(y) = P(Z \le y)$$
$$= P(X \le y) \cdot P(X \le y)$$
$$= F_X(y) \cdot F_Y(y)$$

- $Z = \min(X, Y)$

$$F_Z(y) = P(Z \le y)$$
$$= P(X \le y \cup Y \le y)$$

- X, Y

$$Cov[X, Y] = E[XY] - E[X]E[Y]$$

$$\rho(X, Y) = \frac{Cov[X, Y]}{\sqrt{Var[X] \cdot Var[Y]}}$$

Independent $\Rightarrow$ Uncorrelated

Eg. $X \sim Unif.[-1, 1]$ $\quad Y = X^2$

$E[XY] = E[X^3] = 0$ $\quad E[XY] = 0$

$E[X] \cdot E[Y] = 0$

L11.6 Transformed RV

W, X

$Y = g(W, X)$ $\quad Z = h(W, X)$

$f_{YZ} = ?$

$$f_{YZ}(y, z) = P\big((Y, Z) \in (y, z)\text{ to }(y+dy, z+dy)\big)$$

L11.7 Uniform, Exponential, Normal

- $X \sim Unif(a, b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \end{cases}$$

$$E[X] = \frac{a+b}{2} \qquad Var[X] = \frac{(b-a)^2}{12}$$

- Exponential

$X \sim exp(\lambda)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & \text{other} \end{cases}$$

$$F_X(x) = 1 - e^{-\lambda x}$$

Memorylessness $\quad a > b$

$$P(X \ge a \mid X \ge b) = P(X \ge a - b)$$

$E[X] = \frac{1}{\lambda}$

$X \sim exp(\lambda)$

$Y \sim exp(\tau)$

$Z = \min(X, Y)$

$Z \sim exp(\lambda + \tau)$

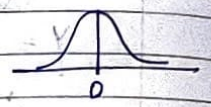
$F_Z(y) = 1 - e^{-(\lambda+\tau)y}$

$$f_Z(y) = \begin{cases} 0 & y < 0 \\ (\lambda+\tau)\, e^{-(\lambda+\tau)y} & \end{cases}$$

- Normal

$$Z \sim N(0,1)$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$



- Gaussian

$$X = \sigma Z + \mu \qquad Z \sim N(0,1)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \times \frac{1}{\sigma}$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$

$$X \sim N(\mu, \sigma^2)$$

$$E[X] = \mu \qquad Var(X) = \sigma^2$$

Week-12

L12.1 Bivariate & Multivariate Normal

- $Z_1 \sim N(0,1) \ \ldots \ Z_d \sim N(0,1)$ $\qquad Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_d \end{bmatrix}$

$$f_Z(z) = \frac{1}{2\pi}^{d/2} \exp\left(\frac{-1}{2}\|z\|^2\right)$$

- $X_1 = Z_1$; $X_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2$

$$X = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} Z \quad \Rightarrow Z = \begin{bmatrix} 1 & 0 \\ \frac{-\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{bmatrix}$$

A $\qquad$ A^{-1}

$$\det(A) = \sqrt{1-\rho^2} \qquad \det(A^{-1}) = \frac{1}{\sqrt{1-\rho^2}}$$

$$E[X_1] = E[X_2] = 0$$

$$\text{cov}[X_1, X_2] = E[X_1 X_2] = E[\rho Z_1^2 + \sqrt{1-\rho^2} Z_1 Z_2]$$

$$\text{Var}[X_1] = 1 \qquad \text{Var}[X_2] = \rho^2 + 1 - \rho^2 = 1$$

$$\text{cov}[X] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} = AA^T$$

Σ $\qquad$ $\det(\Sigma) = 1-\rho^2$

$$[\text{cov}[X]]^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

- $f_X(x) = f_Z(A^{-1}x) \cdot |\text{Det}(A^{-1})|$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right.$$

$$= f_{X_1}(x_1) \cdot f_{X_2|X_1}(x_2|x_1)$$

$\sim N(x_1|0,1)$ $\qquad$ $\sim N(x_2|\rho x_1, 1-\rho^2)$

- $X = AZ$

$$f_X(x) = f_Z(A^{-1}x) \cdot |\det(A^{-1})|$$

L12.2 Estimation of Parameters using ML

- $P = \{P_\theta : \theta \in \Theta\}$

$X_1 \cdots X_n$ drawn i.i.d from P_θ for $\theta \in \Theta$

- $L(\theta) = P(X_1 = x_1 \cdots X_n = x_n \mid \theta)$

$$= \prod_{i=1}^{n} f_{X_i}(x_i \mid \theta)$$

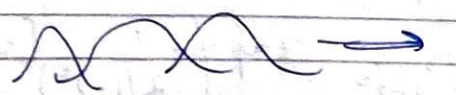
$$= \prod_{i=1}^{n} P_\theta x_i$$

- $\log(L(\theta)) = \sum_{i=1}^{n} \log(P_\theta(x_i))$
- $P = \{\text{Bern}(\theta) : \theta \in [0,1]\}$

$$P_\theta(x) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

- $R(\theta) = -\sum_{i=1}^{n} \log(P_\theta(x_i))$
- $\hat{\theta}_{ML} = \frac{a}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

L12.3 Gaussian mixture models & expectation max



→ multimodal

$$f_X(x) = \sum_{k=1}^{K} \Pi_k N(x \mid \mu_k, \Sigma_k)$$

L12.4 Laws of large nos.

- Markov inequality

$$P(X \geq t) \leq \frac{\mu}{t}$$

X is +ve RV, $E[X] = \mu$

- Chebyshev Inequality

$E[X] = \mu$, $\text{Var}[X] = \sigma^2$

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

- Hoeffding Inequality

$$P(|\bar{X}_n - \mu| \geq t) \leq \frac{\text{Var}[\bar{X}_n]}{\epsilon^2} = \frac{\sigma^2}{nt^2}$$

- Convergence in Probability

$X_1 \cdots$

$X_n \xrightarrow{P} X$ if

$$\lim_{n \to \infty} P(|X_n - X| \geq t) = 0 \quad \forall \epsilon > 0$$

= Law of large nos.

$X_1 \cdots$ $E[X_i] = \mu$

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \bar{X}_n \xrightarrow{P} \mu$$