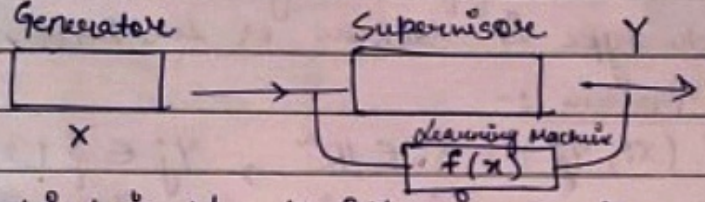
9/1/23

IT-60

Machine Learning

- Assignments
- Surprise Quiz
- In-sem
- End-Sem



Relationship b/w X & Y is random.

Chaotic behaviour → random nature.

(label present) → Supervised learning → supervisor respond is must.

(label not present i.e Y not present) → Unsupervised " → " " " not must.

Semi-supervised learning → using few labels to get info.

Problem of Classification

How to classify different groups

features used to differentiate b/w diamond or gold (eg.)

X_1	X_2	X_3	$\xrightarrow{f}$ Y	
Colour	light intensity	Density	(label data)	
2,1,0	5	9.2	-1	(diamond)
1,2,1	6	1.6	1	(gold)
1,0,1	3	1.2	1	
1,1,1	5	0.7	1	
1,1,1	4	0.2	-1	(diamond)
1,0,1,	5	0.1		

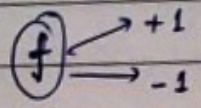
(supervised learning)

→ Now this will be using ML model to predict which is gold | diamond.

11/1/23

Classification Problem

We need variables or label data to classify types of fish.



eg.- length, height, no. of fins, weight, color. to decide whether fish type is salmon or sea bass.

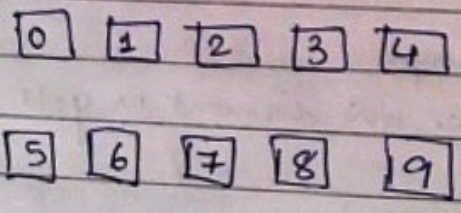
Multiclassification Problem :-

→ x Training set $\{(x_i, y_j) : x_i \in \mathbb{R}^n, y_j \in \{1, 2, \ldots, k\}$ $i = 1, 2, \ldots l\}$ which can efficiently predict

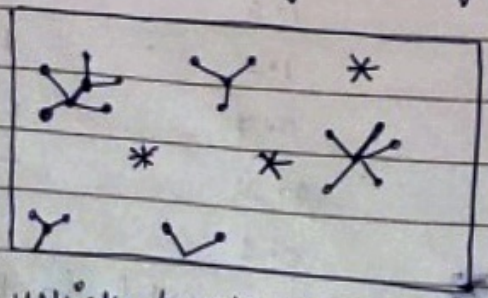
* For binary classification → 1 or 2, +1 or -1 etc.

For every data pt., we have vectors in $\mathbb{R}^n$.

10-class classification (∵ 10 numbers)



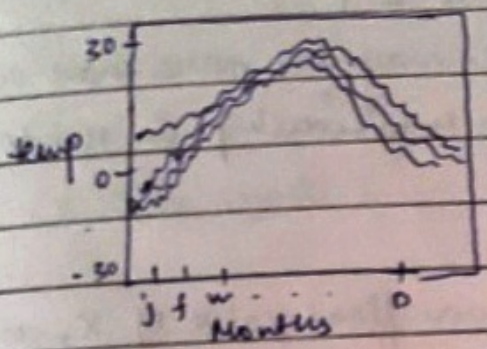
Graph-Data : → eg. instagram data



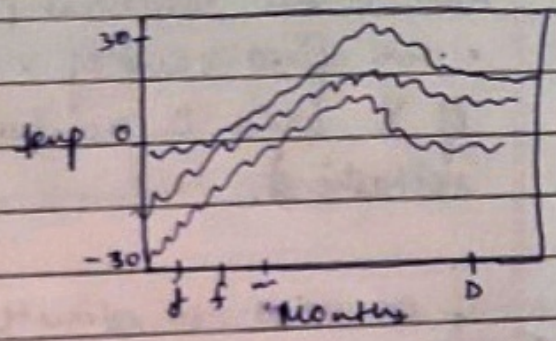
To refer which people are students, which are not, for this we need graph data.

Functional Data : eg. - weather data, EEG data (signal)

Rainy station



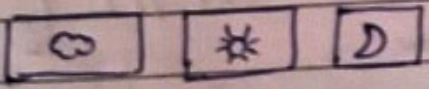
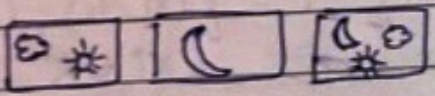
Dry station



Functional data → when x is a function not a vector.

eg.→ traffic → which roads are prone to accidents.

Multi-class and multi-label classification

Multi-class		Multi-label
c = 3	Samples	
(100) (010) (001)		
	labels (t) [0 0 1] [1 0 0] [0 1 0]	labels (t) [1 0 1] [0 1 0] [1 1 1]

Multi-table → movie (components - genre, actor, director etc.)

→ wikipedia page (components → academic, author, entertainment etc.)

- Y is a vector
- categories are present (more than 1, so multi-label).

Regression Problem (one parameter)

Y X

Heights (in cm) , weight (in kg)

want to see relationship b/w X & Y

- For fixed value of X, there may be more than one value of Y, which is random, so relationship is random or stochastic.

If we want to estimate Y for a fixed value of X, we should check central tendency (mean, median, mode).

$$f(x) = E(Y/X)$$

more than one parameter.

we have to estimate income of a person.

having gender, education, seniority, age, work class, income

$$T = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_l, Y_l)\}.$$

$X_i \in R^n$

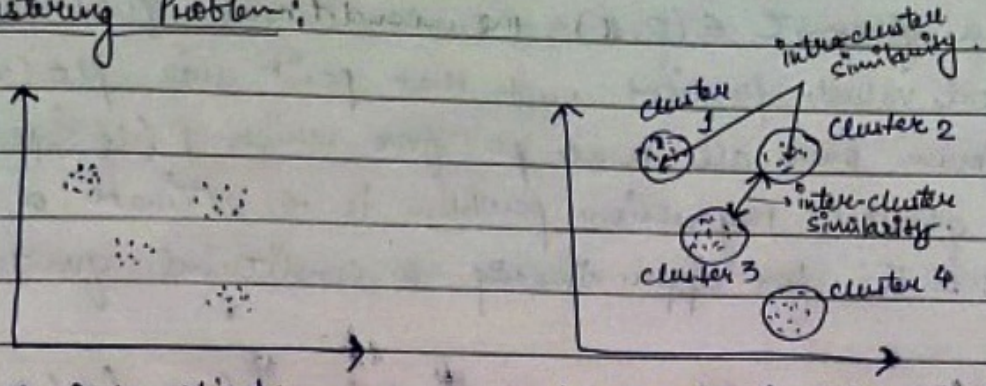
$Y_i \in R.$

$$f(x) \rightarrow Y$$

problem → to estimate fn x of data to give approx value of Y.

if we want to predict, both income & saving, then this is the case of multi-output regression problem.

Clustering Problem:



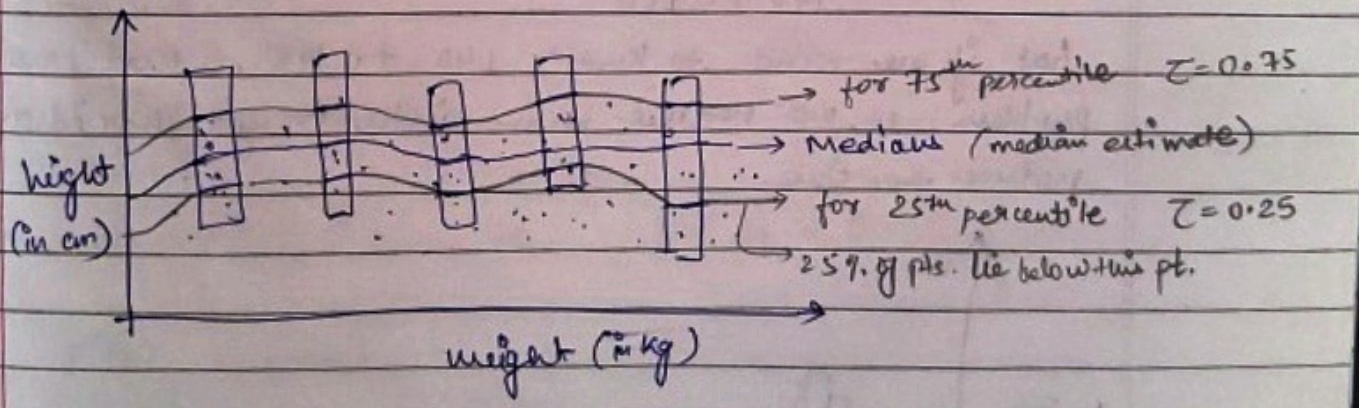
a. Data objects b. clustered data objects.

16/1/23 Most common problem is regression problem as any classification problem can be classified as a regression problem.

Measures of central tendency. → Mean, Median, Mode

Mean is affected by outliers.

For a fixed value of weight, there may be diff^n. values of height.

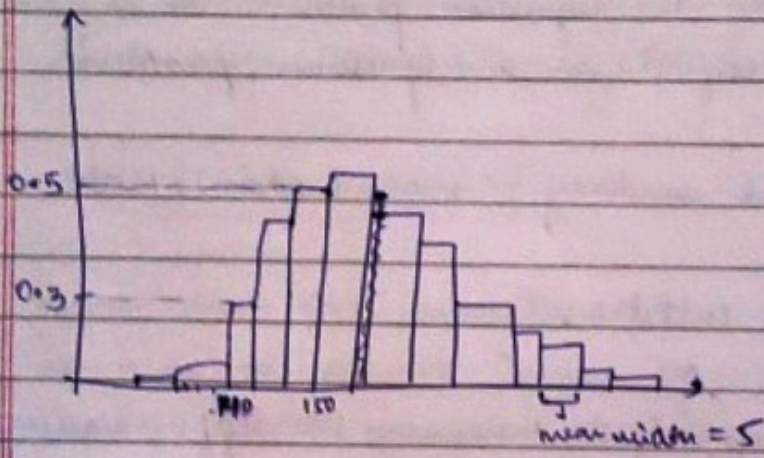


$f(X) = \text{median}(Y/X)$

For a given $\tau \in (0,1)$, the conditional quantile μ_τ, is a real valued function such that point wise $\mu_\tau(x)$ is the minimum over all real μ for which $P(y \leq \mu / x) = \tau$

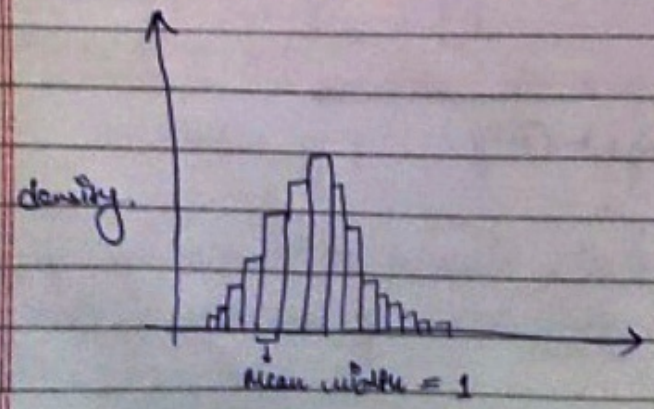
The quantile regression problem is to estimate a fn. $f_\tau \in F$ for approximating a conditional quantile $\mu_\tau(x)$

Conditional quantile → τ % of data points should be lying below $f_\tau(x)$.



0.3 + 0.5 = 0.8 % of prop. of students lies under 140 to 150.

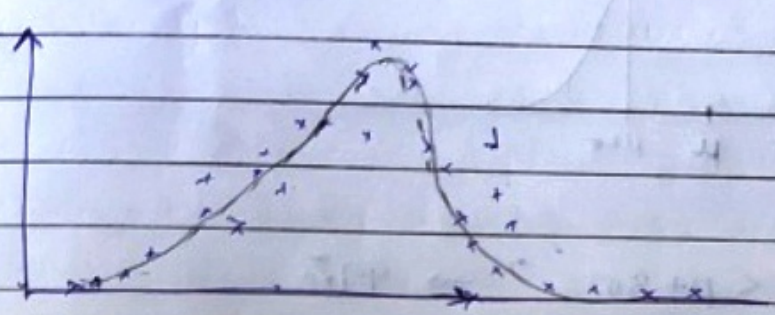
But if we want to know 140 to 148, that creates problem. so, we reduce mean width to get clear idea, ie, reduce the size



if width is unique, then y axis gives relative frequency.

Density = prop. × width

density curve se prob. nikalna h to add kar dage dono ka prop.



$f(x)$, $[a, b]$

$$a f(a) + (a+h) f(a+h) + \dots$$

$$\int_a^b f(x)\,dx.$$

Properties:

→ $f(x) \geq 0$

→ $\int_{-\infty}^{\infty} f(x)\,dx = 1.$

→ $P(X = a) = 0$ ie $\int_a^a f(x)\,dx = 0$

⇒ $F'(x) = f(x)$

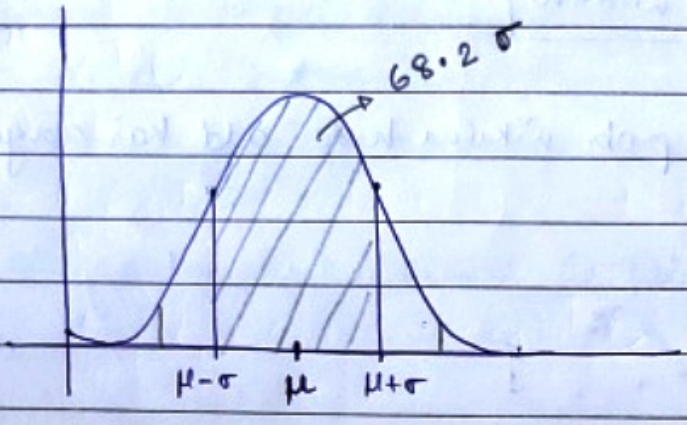
→ $P(a \leq X \leq b) = F(b) - F(a)$

→ $F(a) = P(X \leq a)$

→ $F(b) = P(X \leq b)$

18/1/23

Normal Distribution



$\mu - 3\sigma < x < \mu + 3\sigma \quad \rightarrow 99\%$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \sigma > 0, \quad -\infty < \mu < \infty$$

$$\min_c \sum_{i=1}^{n} (x_i - c)^2, \quad x_i \in R \qquad \text{find } c \rightarrow \text{centroid}$$

Mean → sum of squares of deviation is minimum.

$$\frac{d}{dc} \sum_{i=1}^{n} (x_i - c) = 0$$

$$\Rightarrow \boxed{c = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Manhattan distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2}$$

$$d(x,y) = \sum_{i=1}^{n} |x_1-y_1| + |x_2-y_2| + \cdots + |x_n-y_n|$$

$$d(x,y) = \left[\sum_{i=1}^{n} (x_1-y_1)^p + (x_2-y_2)^r + \cdots + (y_n-y_n)^p\right]^{1/p}$$

(A, d)
↳ distance

(i) → distance is always symmetric, ie, $d(x,y) = d(y,x) \quad \forall\ x,y \in A$.

(ii) → distance from $(x,y) = 0$
ie, $d(x,y) = 0 \iff x = y$.

(iii) $d(x,y) + d(y,z) \geq d(x,z)$, $x,y,z \in R$
ie, sum of 2 sides of distance is always greater than the third side.

→ Minimising sum of Euclidean distance.

$$\min_{c \in R^2} \sum_{i=1}^{2} (x_i - c)^T (x_i - c), x_i \in R^2$$

$$c = \frac{1}{n} \sum x_i$$

$$x_i = [x_i, y_i]$$

$$\min_{c \in R^2} \sum_{i=1}^{2} (x_i, y_i)$$

$$[c_1, c_2] = \frac{1}{n} \sum_{i=1}^{2} (x_i, y_i)$$

$$\min_{\substack{c \in R^2 \\ (c_1,c_2)}} (x_1-c_1)^2 + (y_1-c_2)^2 + (x_2-c_1)^2 + (y_2-c_2)^2 + \cdots + (x_n-c_1)^2 + (y_n-c_2)^2$$

Let $\frac{\partial L}{\partial c_1} = 0 \Rightarrow c_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$

var. of projected data ? pt. on d.
median, mode of n data pts.

$x^T x \rightarrow$ dot product of x^T and x.

$$\frac{1}{n}\sum_{i=1}^{n} x_i^T d$$

$$= \frac{1}{n}(x_1^T d + x_2^T d + \cdots + x_n^T d)$$

$$= \frac{1}{n}(x_1 + x_2 + \cdots + x_n)^T d$$

$$= \underbrace{\mu^T d}_{\text{mean of actual data}} \qquad d \rightarrow \text{(projected in d.)}$$

* $\mu^T d$: mean of projected data in d.
Take the mean of actual data & project it on d.
(d is unit vector).

Saare data points ko ek unit vector pe project kia to bivariate or n-variate data ke scalar points mile unka mean nikala. (formula se pata chala mean value pt. ko unit vector pe project karne se bivariate data ka mean milta h.

19/1/23

$x_1 \in \mathbb{R}$
$x_2 \in \mathbb{R}$
$\vdots$
$x_n \in \mathbb{R}$

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$E(\bar{x}) = \mu$$ $\bar{x}$ is unbiased estimator of μ.

$x \uparrow$ $Y \downarrow$ → covariance −ve.
$x \uparrow$, $Y \uparrow$ → covariance +ve.

$$\text{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

−+	++
−−	+−

If most data pts. lie b/w 1st & 3rd quadrant → +ve covariance
" " " " 2nd & 4th " → −ve covariance.

$$\text{Corr}(X,Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$

$r(x,y)$
$-1 \le r \le 1$.

x_1 y_1 → x_1 → $d^T x_1$ → $(d^T x_1 - d^T \mu)^2$
x_2 y_2 → x_2 → $d^T x_2$ → $(d^T x_2 - d^T \mu)^2$
$\vdots$
x_n y_n → x_n → $d^T x_n$ $(d^T x_i - d^T \mu)^2$
μ_x μ_y $d^T \mu$

μ

Var. of projected matrix : $\frac{1}{n-1}\sum_{i=1}^{n}(d^T(x_i - \mu))^2$

$$= \frac{1}{n-1}(d^T(x_i - \mu))^T (d^T(x_i - \mu))$$

- inner product
- norm
- dist.
- dot product

Date
Page

Length of vector (Norm) :

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

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

* $\|x_2\|_2 = \left| \sqrt{\sum_{i=1}^{n} x_i^2} \right|$ (L_2 Norm of x) ∵ distance, so +ve

* $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (L_1 Norm of x)

* $\|x\|_p = \left(\sum_{i=1}^{n} x_i^p \right)^{1/p}$ (p^{th} norm of x)

* $\|x\|_\infty = \max_{i=1,2,\dots n} (x_i)$ (∞ norm)

eg. $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ $\quad \|x\|_1 = 1+2+3 = 6$

$\|x\|_2 = \left| \sqrt{1^2+2^2+3^2} \right| = \left| \sqrt{14} \right|$

* $\|x+y\| \leq \|x\| + \|y\|$

$\|x\|_2^2 = x^T x$

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$x^T x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$$x^T x = x_1^2 + x_2^2 + \cdots + x_n^2$$

$x^T \rightarrow w$.

$$= \frac{1}{L-1} \sum_{i=1}^{L} d^T (x_i - \mu)^2$$

x = data vector.

$$= \frac{1}{L-1} \left(d^T (x-\mu) \right) \left(d^T (x-\mu) \right)^T$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{bmatrix}$$

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ y_1 & y_2 & \cdots & y_L \end{bmatrix}$$

$$(X-\mu) = \begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 & \cdots & x_L - \mu_1 \\ y_1 + \mu_2 & y_2 - \mu_2 & & y_2 - \mu_2 \end{bmatrix}$$

$$= \frac{1}{L-1} d^T (x-\mu)(x-\mu)^T d$$

11/23

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ y_1 & y_2 & & y_L \end{bmatrix}$$

then projection, $x^T d$.

mean $\mu^T d$

$d^T \Sigma d$

$$= \frac{1}{L-1} \sum_{i=1}^{L} \left[(x^T d) - (d^T \mu_i) \right]^2$$

$$= \frac{1}{L-1} \sum_{i=1}^{L} \left[(x^T d) - (\mu^T d) \right]^2$$

$$= \frac{1}{L-1} (x^T d - \mu^T d)^T (x^T d - \mu^T d)$$

$$= \frac{1}{L-1} d^T (x^T - \mu^T)^T (x^T - \mu^T) d$$

$$= \frac{1}{L-1} d^T (x-\mu)(x-\mu)^T d.$$

Variance in d direction :

$$= \frac{1}{L-1} d^T (x-\mu)(x-\mu)^T d$$

$$\frac{1}{L}(x-\mu)(x-\mu)^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ y_1 & y_2 & \cdots & y_L \end{bmatrix} - \begin{bmatrix} \mu_L \\ \mu_2 \end{bmatrix}$$

$$(x-\mu)(x-\mu)^T = \frac{1}{L} \begin{bmatrix} x_1-\mu_1 & x_2-\mu_1 & \cdots & x_L-\mu_1 \\ y_1-\mu_2 & y_2-\mu_2 & \cdots & y_L-\mu_2 \end{bmatrix} \begin{bmatrix} x_1-\mu_1 & y_1-\mu_2 \\ x_2-\mu_1 & y_2-\mu_2 \\ \vdots & \vdots \\ x_L-\mu_1 & y_L-\mu_2 \end{bmatrix}$$

$$= \frac{1}{L-1} \begin{bmatrix} \sum_{i=1}^{L}(x_i-\mu_1)^2 & \sum_{i=1}^{L}(x_i-\mu_1)(y_i-\mu_2) \\ \sum_{i=1}^{L}(x_i-\mu_1)(y_i-\mu_2) & \sum_{i=1}^{L}(y_i-\mu_2)^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$$

↓

Variance - Covariance Matrix.

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ y_1 & y_2 & \cdots & y_L \end{bmatrix} \in R^2$$

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_L \\ y_1 & y_2 & \cdots & y_L \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$$

Variance $= d^T \Sigma d$.

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{L1} \\ x_{12} & x_{22} & \cdots & x_{L2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & & x_{Ln} \end{bmatrix} \qquad d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}$$

$$\|d\|_2 = 1.$$

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

→ Var Cov. Matrix for n components.

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

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

$$\mu_1 = \frac{2}{4} = \frac{1}{2}$$

$$\mu_2 = \frac{1}{2}$$

$$(A-\mu) = \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & 0.5 & -0.5 & 0.5 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$(A-\mu)^T = \begin{bmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$(A-\mu)(A-\mu)^T = \begin{bmatrix}$$

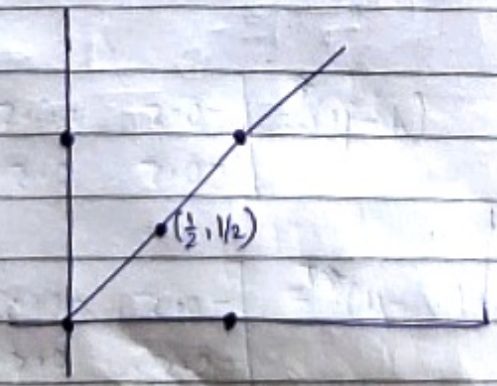
$$(A-\mu)(A-\mu)^T = \begin{bmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & 0.5 & -0.5 & 0.5 \end{bmatrix}$$

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

$$\Sigma = \frac{1}{L-1} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad L = 4$$

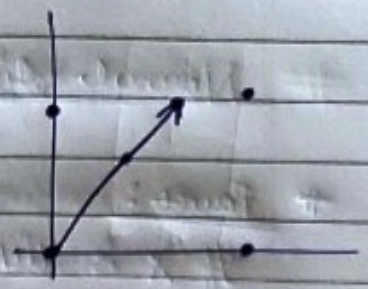
$$= \frac{1}{3} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix}$$



$$\begin{bmatrix} -0.5 & -0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$d = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$



$$d^T \Sigma d = \begin{bmatrix} 3/5 & 4/5 \end{bmatrix} \begin{bmatrix} 1/3 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}$$

$$= \begin{bmatrix} 3/5 & 4/5 \end{bmatrix}_{1\times 2} \begin{bmatrix} 3/15 \\ 4/15 \end{bmatrix}$$

$$= \frac{9}{75} + \frac{16}{75} = \frac{25}{75} = \frac{1}{3}$$

$$\boxed{d^T \Sigma d = \frac{1}{3}} \rightarrow \text{Variance of projected data pt. in } d.$$

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}_{4\times 2} \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix}_{2\times 1} = \begin{bmatrix} 0 \\ 4/5 \\ 3/5 \\ 7/5 \end{bmatrix}$$

A $\qquad$ d $\qquad$ $\downarrow$

projection in d dirⁿ.

25/1/23

Normal distribution,

* **Result:**
If X is distributed as $N_p(\mu, \Sigma)$ then any linear comb^n. of variables $a'X = a_1X_1 + a_2X_2 + \cdots + a_pX_p$ is distributed as $N(a'\mu, a'\Sigma a)$.
Also if $a'X$ is distributed as $N(a'\mu, a'\Sigma a)$ for every a, then X must be $N_p(\mu, \Sigma)$.

* If X is distributed as $N_p(\mu, \Sigma)$, the q linear combinations,

$$A_{(q\times p)} X_{(p\times 1)} = \begin{bmatrix} a_{11}X_1 + \cdots + a_{1p}X_p \\ a_{21}X_1 + \cdots + a_{2p}X_p \\ \vdots \\ a_{q1}X_1 + \cdots + a_{qp}X_p \end{bmatrix}$$

In 1-D

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$\Sigma \rightarrow \begin{bmatrix} 8.002 \\ \text{var (x dir}^n\text{)} \end{bmatrix}$ → var. cov. matrix. (let)
var (y dir^n)

mean = 0, 0

$(x-\mu)^T \Sigma^{-1} (x-\mu)$

* dist. of $\left(\frac{x-\mu}{\sigma}\right)$ for pt → Mahalanobis distance.

Variance → $(x-\mu)^T \Sigma^{-1} (x-\mu)$

x mean se kitna S.D. dur h → given by $(x-\mu)^T \Sigma^{-1} (x-\mu)$

For multi-Dimension: $X \sim N_p(\mu, \Sigma)$

$$f(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$ (in p-dimension)

$-\infty < x_i < \infty, \quad i = 1, 2, \cdots, p.$

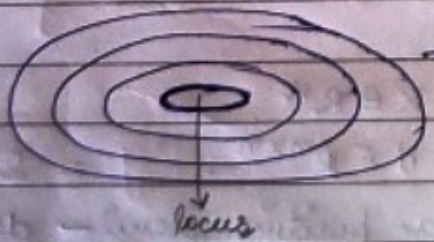
where, Σ → positive definite

6/2/23

Plotting Multivariate Normal Density.

$\sigma = (x-\mu)^T (x-\mu)$

→ in



→ contour plot

focus

for symmetric distribution → contour is circle otherwise, it is ellipse.

Bivariate Normal Density :

Evaluating for $p = 2$ variate normal density in terms of individual parameters,

$\mu_1 = E(X_1)$

$\mu_2 = E(X_2)$

$\text{Var}(X_1) = \sigma_{11}$

$\text{Var}(X_2) = \sigma_{22}$

$\text{Corr}(X_1, X_2) = \rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$

Result: If Σ is positive definite, so that Σ^{-1} exists. then

$$\Sigma e = \lambda e$$

$$\Rightarrow \Sigma^{-1} e = \frac{1}{\lambda} e.$$

so (λ, e) is an eigenvalue - eigenvector pair for Σ corresponding to the pair $\left(\frac{1}{\lambda}, e\right)$ for Σ^{-1}.

Also Σ^{-1} is positive definite.

If $X^T H X \geq 0$, $X \in R^n$, $H \in R^{n \times n}$.

then H is said to be a positive semi- definite matrix.

Constant probability density contour :-

$= \{$all x such that $(x-\mu)^T \Sigma^{-1}(x-\mu) = c^2\}$

= surface of an ellipsoid centered at μ.

Contours of constant density for the p-dimensional normal distrib^n. are ellipsoids defined by x such that.

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = c^2$$

These ellipsoids are centered at μ & have axes $\pm c\sqrt{\lambda_i} e_i$ where $\Sigma e_i = \lambda_i$ for $i = 1, 2, \cdots, p$.

Convex Sets :

$C \subseteq \mathbb{R}^n$ is convex if $tx + (1-t)y \in C$ for any $x, y \in C$ & $0 \leq t \leq 1$.

convex set

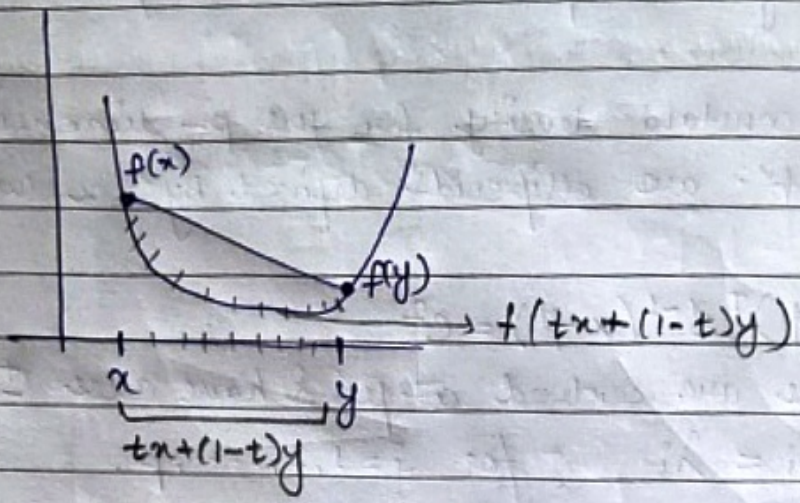
not convex set.

Convex function :

$f : \mathbb{R}^n \longrightarrow R$ is convex if dom(f) (the domain of f) is a convex set, & if $f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$ for any $x, y \in$ dom (f) & $0 \leq t \leq 1$.

$tx + (1-t)y$ → pt. on axis.

$\Rightarrow f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$



Convex function :

$f(x)$ is convex fn.
then $f(x)$ would be concave fn.
also, $\alpha f(x)$ where $\alpha \geq 0$ is convex fun.
$f_1(x)$ & $f_2(x)$ is convex then,
$f_1(x) + f_2(x)$ would be convex fn.
then, $\max(f_1(x) + f_2(x))$ is convex fn.

⇒ $y = x^2$ → convex fn. ($\because \frac{d^2y}{dx^2} = 2 > 0$)

⇒ $y = -x^2$ → concave fn.

⇒ $y = \sin x$ → concave on interval $[0, \pi]$.

8/2/23

Date
Page

Strictly convex function :

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(x)$$

where $0 \leq \lambda \leq 1$.

Concave function :

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(x)$$

where $0 \leq \lambda \leq 1$.

Strictly concave function :

$$f(\lambda(x) + (1-\lambda)y) > \lambda f(x) + (1-\lambda) f(x)$$

where $0 \leq \lambda \leq 1$.

Gradient and Derivatives :-

- For a fn. $f(x) = f(x_1, x_2, \ldots, x_n)$, and a unit vector $u = (u_1, \ldots, u_n)$, the directional derivative is

as $$\nabla_u f(x) = \lim_{h \to 0} \left(\frac{f(x+hu) - f(x)}{h} \right)$$

$f(x)$ is differentiable implies that $\nabla_u f(x)$ is well defined for all x & u can be obtained as

$$\nabla f = \begin{bmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{bmatrix}$$

↓ gradient (rate of change)

& the directional derivative of f at pt. x, for any dirn. (unit vector) u can be obtained as,

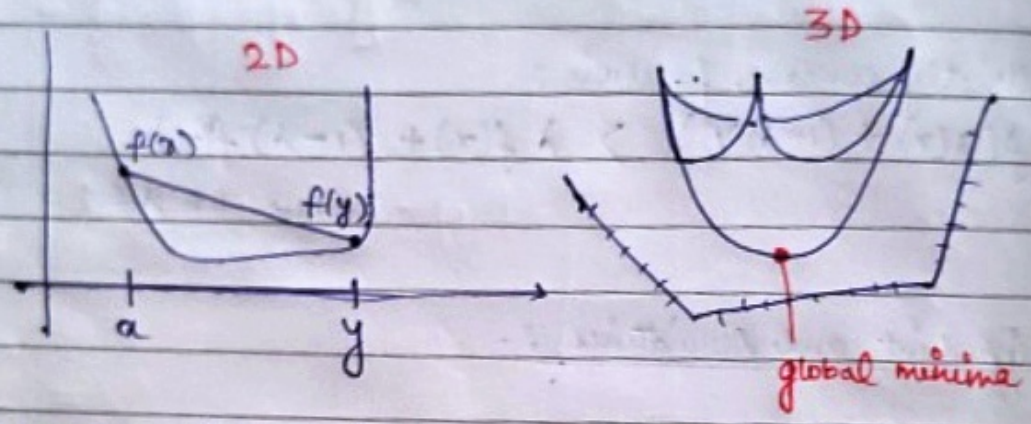
$$\nabla_u f(x) = u^T \nabla f$$

Unconstrained Convex Optimization

Consider the optimization problem

$\min_x f(x)$; $f: R^n \to R$ is convex & smooth,

then the necessary & sufficient condition for optimal soln. x_0 is $\nabla f(x) = 0$ at $x = x_0$



$$(\nabla f(x))^T = \left[\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \ldots, \frac{\partial L}{\partial x_n}\right]$$

If d dirⁿ. me pta karna h, then,

$d^T \nabla f(x)$, where d is unit vector & $\|d\|_2 = 1$

Find the dirⁿ. u such that, $d^T \nabla f(x)$ is maxᵐ (u) such that $\|u\|_2 = ?$

* if $\max_b a^T x$ (we have to ~~minimise~~ maximise b, then b should be in dirⁿ. of $a^T x$, then $a^T b$ will be maxᵐ).

⇒ So, in dirⁿ. of $\nabla f(x)$, then $d^T \nabla f(x)$ will be sharply increasing, ie, in dirⁿ. of gradient.

⇒ $\frac{\nabla f(x)}{\|\nabla f(x)\|}$

* for minimum, the dirn. of gradient will be -ve.
ie if $\min_b a^T b$, then, $b = -a^T$.

$$\Rightarrow \frac{\nabla f(x)}{\|\nabla f(x)\|} = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

Gradient-Descent Algorithm:
- An iterative algorithm.
- The -ve gradient dirn. is dirn. of steepest descent.

★ Method of Gradient Descent:
(i) initialise $x_0 = x_{start} \in R^n$
(ii) repeat $x^{(k+1)} = x^{(k)} - \gamma_k \nabla f(x^{(k)})$
until $\|\nabla f(x^{(k)})\| \leq \varepsilon$.

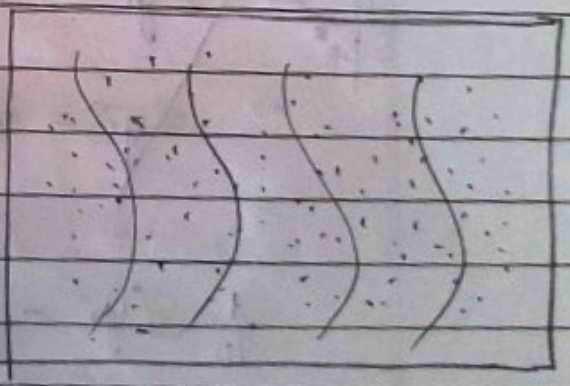
k = no. of steps
γ_k = eta / gamma
↓
tells how much we are moving in a dirn.
(step length)

Regression:

$$Y_i = E(Y_i/X) + \varepsilon_i$$

for mean,

$$f(Y_i) = E(Y/X)$$



Assumptions:
(i) all X_i & Y_i are i.i.d.
(ii) $Y_i = E(Y_i/X) + \varepsilon_i$
(iii) $E(\varepsilon_i) = 0$ (i.e, mean of error term = 0).

↓
random in nature

Task: predict balance using income.

- Training set.
- Testing set

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\varepsilon = [y - (\beta_0 + \beta_1 x_1)]^2$$ we have to minimize error.

$$\min_{\beta_0, \beta_1} \frac{1}{10} \sum_{i=1}^{10} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \end{cases}$$

$$\frac{\partial}{\partial \beta_0} = 0$$

$$\Rightarrow \frac{2}{10} \Sigma (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{\partial^2}{\partial \beta_0^2} \Rightarrow \Sigma y_i = \Sigma (\beta_0 + \beta_1 x_i) = 0$$

$$\Rightarrow$$

$$\frac{1}{10} \Sigma (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial \beta_0} = \frac{2}{10} \Sigma (y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\Rightarrow -\Sigma y_i + \beta_0 + \beta_1 x_i = 0$$

$$\Rightarrow \Sigma y_i = \beta_0 + \beta_1 x_i \quad \Rightarrow \hat{\beta}_0 = \Sigma (y_i - \hat{\beta}_1 x_i)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial}{\partial \beta_1} = \frac{2}{10} \Sigma (y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$= -\Sigma (y_i x_i + \beta_0 x_i + \beta_1 x_i^2 = 0$$

$$\Rightarrow \Sigma y_i x_i = \beta_0 x_i + \beta_1 x_i^2$$

$$\hat{\beta_1} = \frac{\Sigma y_i x_i - \beta_0 x_i}{x_i^2}$$

$$\hat{\beta_1} = \frac{\Sigma y_i}{x_i} - \frac{\beta_0}{x_i}$$

$$= \frac{\Sigma y_i}{x_i} - \frac{1}{x_i}(\Sigma y_i - \hat{\beta_1} x_i)$$

$$= \frac{\Sigma y_i}{x_i} - \frac{\Sigma y_i}{x_i} + \hat{\beta_1}$$

Fitting least sq. Regression polynomial

Our estimate is $wx+b \to f(x)$

$$\min_{w,b \in \mathbb{R}} \sum_{i=1}^{L} (y_i - (wx_i + b))^2$$

$$\min_{u \in \mathbb{R}^2} (Y - Au)^T (Y - Au)$$

$$\Delta u = \begin{bmatrix} wx_1 + b \\ wx_2 + b \\ \vdots \\ wx_L + b \end{bmatrix} \qquad A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_L & 1 \end{bmatrix} \quad ; \quad u = \begin{bmatrix} w \\ b \end{bmatrix}$$

$$= \begin{bmatrix} y_1 - (wx_1 + b) \\ y_2 - (wx_2 + b) \\ \vdots \\ y_L - (wx_L + b) \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix}$$

$$\min_{u \in \mathbb{R}^2} (Y - Au)^T (Y - Au) = \begin{bmatrix} y_1 - (wx_1 + b) \\ y_2 - (wx_2 + b) \\ \vdots \\ y_L - (wx_L + b) \end{bmatrix}^T \begin{bmatrix} y_1 - (wx_1 + b) \\ y_2 - (wx_2 + b) \\ \vdots \\ y_L - (wx_1 + b) \end{bmatrix}$$

$$\min_{u \in \mathbb{R}^2} (Y - Au)^T (Y - Au) = (y_1 - (wx_1 + b))^2 + (y_2 - (wx_2 + b))^2 + \cdots + (y_L - (wx_L + b))^2$$

$$\min_u (Y - Au)^T (Y - Au) \simeq J(u)$$

$$\nabla_u J(u) = 2A^T(Y - Au) = 0$$

$$\Rightarrow A^T(Y - Au) = 0$$

$$\Rightarrow \boxed{u = (A^TA)^{-1}A^Ty}$$

$$\frac{\nabla J(u)}{\nabla (Y - Au)} = \left(\frac{Y - Au}{\nabla u}\right)$$

$$\Rightarrow \frac{\partial}{\partial} \min_u (Y - AY)^T (Y - Au)$$

$$= -2(Y - Au) = 0$$

$$\Rightarrow A^T(Y - Au) = 0$$

$$u = (A^TA)^{-1}A^Ty$$

$$u = \begin{bmatrix} w \\ b \end{bmatrix} = \begin{bmatrix} B_1 \\ B_0 \end{bmatrix}$$

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_L & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix}$$

$\min x^Tx = F(x)$, $x \in R^n$

$\min \sum_{i=1}^{n} (x_1^2 + x_2^2 + \cdots + x_n^2)$

$\frac{\partial}{\partial x_1} = 2x_1$, $\frac{\partial}{\partial x_2} = 2x_2$

$\vdots$

$\# \frac{\partial (x^Tx)}{\partial x} = 2Ax$ (A → matrix)

$\# \nabla_x Ax = A^T$

or $\frac{\partial (Ax)}{\partial x} = A^T$

$$\begin{bmatrix} \hat{B_0} \\ B_1 \end{bmatrix} = (x^Tx)^{-1}x^Ty$$

$f(x) = 2.6 + 6.9x$

X	Y
1	0
0	1
1	1
0	1

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \left(\because A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}\right)$$

$$u = (A^T A)^{-1} A^T Y$$

$$A^T A = \underset{2\times4}{\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}} \underset{4\times2}{\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}}$$

$$= \begin{bmatrix} 2 & 2 \\ 2 & 48 \end{bmatrix}$$

$$(A^T A)^{-1} = \frac{1}{4}\begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$$

$$(A^T A)^{-1} A^T Y = \underset{2\times2}{\begin{bmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}} \underset{2\times4}{\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}} \underset{4\times1}{\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}}$$

$$= \underset{2\times2}{\begin{bmatrix} 1 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}} \underset{2\times1}{\begin{bmatrix} 1 \\ 3 \end{bmatrix}} = \begin{bmatrix} 1 - 3/2 \\ -\frac{1}{2} + \frac{3}{2} \end{bmatrix}_{2\times1}$$

$$(A^T A)^{-1} A^T Y = \begin{bmatrix} -1/2 \\ 1 \end{bmatrix}$$

RMSE:

RMSE → Root Mean Square Error

* $$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(y_i - f(x_i)\right)^2}$$

* $$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2}$$

Quadratic Fitting

$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

$f(x) = w_2 x^2 + w_1 x + b$

$$\min_{(w_2, w_1, b)} \sum_{i=1}^{L} \left(y_i - (w_2 x_i^2 + w_1 x_i + b) \right)^2$$

$$A = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & & \vdots \\ x_L^2 & x_L & 1 \end{bmatrix} \qquad u = \begin{bmatrix} w_2 \to \beta_2 \\ w_1 \to \beta_1 \\ \vdots \\ b \to \beta_0 \end{bmatrix}$$

$$\min_{u} (Y - Au)^T (Y - Au)$$

Cubic fitting

$$f(x) = w_3 x^3 + w_2 x^2 + w_1 x + b$$

$$\min_{(w_3, w_2, w_1, b)} \sum_{i=1}^{L} \left(y_i - (w_3 x^3 + w_2 x^2 + w_1 x + b) \right)^2$$

$$A = \begin{bmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ \vdots & & & \\ x_L^3 & x_L^2 & x_L & 1 \end{bmatrix}_{L \times 4} \qquad u = \begin{bmatrix} w_3 \\ w_2 \\ w_1 \\ b \end{bmatrix}$$

$f(x) = w_1 x_1 + w_2 x_2 + b$

* Quad-fitting in n dimension, $A = ?$, $u = ?$, $w = ?$

Date
Page

fitting with fifth order polynomial

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

$$\min_{(w_4, w_3, w_2, w_1, b)} \sum_{i=1}^{L} \left(y_i - (w_4 x^4 + w_3 x^3 + w_2 x^2 + w_1 x + b) \right)^2$$

$$A = \begin{bmatrix} x_1^4 & x_1^3 & x_1^2 & x_1 & 1 \\ x_2^4 & x_2^3 & x_2^2 & x_2 & 1 \\ \vdots & & \vdots & & \\ x_L^4 & x_L^3 & x_L^2 & x_L & 1 \end{bmatrix}_{L \times 5} \qquad u = \begin{bmatrix} w_4 \\ w_3 \\ w_2 \\ w_1 \\ b \end{bmatrix}$$

$$u = (A^T A)^{-1} A^T y$$

when $f(x) = w_1 x_1 + w_2 x_2 + b$ (ie, x_1 & x_2)

$$f(x) = w_1 x_1 + w_2 x_2 + b$$

$$\min_{(w_2, w_1, b)} \sum_{i=1}^{L} \left(y_i - (w_1 x_1 + w_2 x_2 + b) \right)^2$$

* then, $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$,

then, $w^T x + b = f(x)$

ie,

$f(x) = w^T x + b = w_1 x_1 + w_2 x_2 + b$

or, $\min_{(w \in R^2, b \in R^2)} \sum_{i=1}^{L} \left[y_i - (w^T x + b) \right]^2$

$$\min_{w \in R^2,\, b \in R} \sum_{i=1}^{L} \left(y_i - (w^T x_i + b) \right)^2$$

$$\min_{(w \in R^2, u)} (Y - Au)^T (Y - Au)$$

where, $A = \begin{bmatrix} x_{11} & x_{12} & 1 \\ x_{21} & x_{22} & 1 \\ \vdots & \vdots & \vdots \\ x_{L1} & x_{L2} & 1 \end{bmatrix}$ $u = \begin{bmatrix} w_2 \\ w_1 \\ \vdots \\ b \end{bmatrix}$, $b \in R$

$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in R^2$ $x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iL} \end{bmatrix} \in R^2$

$$u = (A^T A)^{-1} A^T Y$$

\# $u = (A^T A)^{-1} A^T Y = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix}$

$X_1 = 5.6$, $X_2 = 37$. Predict balance.

$\therefore f(x) = w_1 x_1 + w_2 x_2 + b$

$= 0.7 x_1 + 0.2 x_2 + 0.1$.

ie, x_1 contributes most in balance.

\# Improving the prediction for $M = 8$

$$\min \frac{1}{10} \sum_{i=1}^{10} \left[y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \beta_6 x_i^6 + \beta_7 x_i^7 \right) \right]^2$$

$X^T H X + f^T X + b$

$\Rightarrow X^T H X + f^T X + b \quad ; \quad w^T x + b \; ; \; w \in R^n, \; b \in R$

H = Sq. matrix, $= \begin{bmatrix} w_5 & w_3/2 \\ \frac{w_3}{2} & w_4 \end{bmatrix} = H$

$$f = \begin{bmatrix} w_2 \\ w_1 \end{bmatrix}$$

$$f(x) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = w_5 x_1^2 + w_4 x_2^2 + w_3 x_2 x_1 + w_2 x_2 + w_1 x_1 + b$$

$$\Rightarrow \min_{(w_5, w_4, \cdots w_1, b)} \sum_{i=1}^{L} \left[y_i - \left(w_5 x_{i1}^2 + w_4 x_{i2}^2 + w_3 x_{i1} x_{i2} + w_2 x_{i2} + w_1 x_{i1} + b \right) \right]^2$$

$$A = \begin{bmatrix} x_{11}^2 & x_{i2}^2 & x_{i1} x_{i2} & x_{i1} & x_{i2} & 1 \\ x_{21}^2 & x_{22}^2 & x_{21} x_{22} & x_{21} & x_{22} & 1 \\ \vdots & & & & & \vdots \\ x_{L1}^2 & x_{L2}^2 & x_{L1} x_{L2} & x_{L1} & x_{L2} & 1 \end{bmatrix}$$

$$u = \begin{bmatrix} w_5 \\ w_4 \\ w_3 \\ w_2 \\ w_1 \\ b \end{bmatrix}$$

$$u = (A^T A)^{-1} A^T Y$$

When $X \in \mathbb{R}^n$

ie $X \in \mathbb{R}^n$, $Y \in \mathbb{R}$

$$A = \begin{bmatrix} x_{11}^n & x_{12}^n & \cdots & x_{1n}^n & x_{11}x_{21}^{n-1} & x_{11}^m x_{12}^{n-2} & \cdots & 1 \\ & & & & & & & \\ & & & & & & & \\ x_{L1}^n & x_{L2}^2 & \cdots & x_{Ln}^n & x_{L1}x_{L2}^{n-1} & x_{L1}^2 x_{L2}^{n-2} & \cdots & 1 \end{bmatrix}$$

$\frac{(m+n)!}{m!\,n!}$

9/2/23

#1. $T = \{(x_i, y_i) : x_i \in R^n,\ y_i \in R,\ i = 1, 2, \ldots L\}$

$f(x) \sim Y$

The regression problem is to find $f(x)$ which can approximate y values well. For this we need to minimize error.

$$\min_f \sum_{i=1}^{L} (y_i - f(x_i))^2 .$$

$L(x, Y, f)$

$$\min_f \sum_{i=1}^{L} L(x_i, y_i, f)$$

$$\sum_{i=1}^{L} (y_i - f(x_i))^2$$

One of the type of loss fn :-

- least square loss function :-

$$L(x_i, y_i, f) = (Y - f(x_i))^2$$

#2. Find a fn. $f \in F$ such that $\sum_{i=1}^{L} (y_i - f(x_i))^2$.

→ set of linear fn. in R^n

$$F = \{ w^T x + b : w \in R^n,\ b \in R \}$$

$$\min_{w \in R^n,\ b \in R} \sum_{i=1}^{L} (y_i - (w^T x_i + b))^2 .$$

$$\begin{bmatrix} w \\ b \end{bmatrix} = (A^T A)^{-1} A^T Y$$

$$A = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} & 1 \\ x_{21} & x_{22} & \ldots & x_{2n} & 1 \\ \vdots & & & & \\ x_{L1} & x_{L2} & \ldots & x_{Ln} & 1 \end{bmatrix}$$

In n dimension, for m no. of parameter, no. of possible terms:

$$= \frac{(m+n)!}{m!\,n!}$$

For Quadratic,

$$\frac{(2+n)!}{2!\,n!}$$

$$= \frac{(2+n)(1+n)\,n!}{2!\,n!} \qquad = \frac{(n+1)(n+2)}{2}$$

For cubic, $\frac{(3+n)!}{3!\,n!}$

$$A = \begin{bmatrix} X_{11}^m & X_{12}^m & \cdots & X_{1n}^m & \cdots & 1 \\ X_{21}^m & X_{22}^m & \cdots & X_{2n}^m & \cdots & 1 \\ \vdots & & & & & \\ X_{L1}^m & X_{L2}^m & \cdots & X_{Ln}^m & \cdots & 1 \end{bmatrix}$$

$$\begin{bmatrix} w_K \\ w_{K-1} \\ \vdots \\ w_1 \\ b \end{bmatrix} = (A^T A)^{-1} A^T Y$$

$$K = \frac{(m+n)!}{m!\,n!} - 1$$

13/2/23

Training set

$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$ $\quad x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $i = 1, 2, \ldots, L$

→ to fit a line / hyperplane

(linear) → $f(x) = w^T x + b$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$

$= w_k x_1^m + w_{k-1} x_2^m + \cdots + w_{k-(n+1)} x_n^m + \cdots$

$+ w_{k-n} x_1^{m-1} x_2 + \cdots + b.$

(non-linear) $f(x) = w^T \phi(x)$

$= w_1 \phi_1(x) + w_2 \phi_2(x) + \cdots + w_m \phi_m(x)$

where, $\phi_1(x) = x_1^m$

$\phi_2(x) = x_2^m$

$\vdots$

$\phi_{m-1}(x) = x_n$

$\phi_m(x) = 1$ (may be 'b')

$\phi_k(x) = x_1^{r_1} x_2^{r_2} x_3^{r_3}$; s.t. $r_1 + r_2 + \cdots + r_n \leq m$

(non-linear) → $f(x) = w^T \phi(x)$

$= \begin{bmatrix} w_m \\ w_{m-1} \\ \vdots \\ w_1 \\ w_0 \; (b) \end{bmatrix}^T \begin{bmatrix} \phi_m(x) \to x_n^m \\ \phi_{m-1}(x) \\ \vdots \\ \phi_0(x) \to 1 \end{bmatrix}$

$= \begin{bmatrix} w_m \\ w_{m-1} \\ \vdots \\ w_1 \\ b \end{bmatrix}^T \begin{bmatrix} \phi_m(x) \\ \phi_{m-1}(x) \\ \vdots \\ 1 \end{bmatrix}$

$= w^T \phi(x) + b \quad ; w \in \mathbb{R}^m, b \in \mathbb{R}$

for Quadratic :

$$\begin{bmatrix} w_2 \\ w_1 \\ \vdots \\ b \end{bmatrix} \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & & \\ x_L^2 & x_L & 1 \end{bmatrix}$$

$$f(x) = w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + b x_1$$

$\downarrow \phi_1(x)$ $\downarrow \phi_2(x)$ $\downarrow \phi_3(x)$ $\downarrow \phi_4(x)$ $\downarrow \phi_5(x)$ $\downarrow \phi_6(x)$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ b \end{bmatrix} \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \phi_5(x) \\ 1 \end{bmatrix}$$

$$\min_{w \in R^n, b \in R} \sum_{i=1}^{L} \left(Y_i - (w^T x_i + b)\right)^2 \longrightarrow \text{linear}$$

$$\min_{w \in R^m, b \in R} \sum_{i=1}^{L} \left(Y_i - (w^T \phi(x_i) + b)\right)^2$$

where $\phi(x_i) = \begin{bmatrix} \phi_m(x) \\ \phi_{m-1}(x) \\ \vdots \\ \phi_1(x) \end{bmatrix}$

take derivative & equate it to 0.

$$\nabla_w J(w, b) = 0$$

$$\begin{bmatrix} w \\ b \end{bmatrix} = (A^T A)^{-1} A^T Y$$

$$A = \begin{bmatrix} \phi_m(x_1) & \phi_{m-1}(x_1) & \cdots & \phi_1(x_1) & 1 \\ \phi_m(x_2) & \phi_{m-1}(x_2) & \cdots & \phi_1(x_2) & 1 \\ \vdots & & & & \vdots \\ \phi_m(x_4) & \phi_{m-1}(x_4) & & \phi_{m-1}(x_4) & 1 \end{bmatrix}$$

$\phi_2(x_1)$ $\phi_2(x_1)$ $\phi_3(x_1)$

$$\begin{bmatrix} x_{11}^{\perp} & x_{12}^{\perp} & x_{13}^{\perp} & x_{11}x_{12} & x_{12}x_{13} & \cdots \\ x_{21}^{\perp} & x_{22}^{\perp} & x_{23}^{\perp} & x_{21}x_{22}, & \cdots & \\ x_{31}^{\perp} & x_{32}^{\perp} & x_{33}^{\perp} & x_{31}x_{32}, & \cdots & \end{bmatrix}$$

$\phi_1(x_2)$

$x_1 \in \mathbb{R}^3$, $x_2 \in \mathbb{R}^3$, $x_3 \in \mathbb{R}^3$

$$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \end{bmatrix} \qquad x_2 = \begin{bmatrix} x_{21} \\ x_{22} \\ x_{23} \end{bmatrix}$$

$$f(x) = w^T \phi(x) + b$$ $w \in R^w$, $b \in R$

Gaussian function :-

m diff. basis function

$$\phi_m(x) = e^{-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma_m}\right)^2}$$

$$\phi_{m-1}(x) = e^{-\frac{1}{2}\left(\frac{x-\mu_{m-1}}{\sigma_{m-1}}\right)^2}$$

$$\phi_1(x) = e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}$$

$$\Rightarrow \begin{bmatrix} w \\ b \end{bmatrix} = (A^T A)^{-1} A^T Y$$

where,

$$\Rightarrow A = \begin{bmatrix} \phi_m(x_1) & \phi_{m-1}(x_1) & \cdots & \phi_1(x_1) & 1 \\ \phi_m(x_2) & \phi_{m-1}(x_2) & \cdots & \phi_1(x_2) & 1 \\ \vdots & & & & \\ \phi_m(x_L) & \phi_{m-1}(x_L) & \cdots & \phi_1(x_L) & 1 \end{bmatrix}_{L \times (m+1)}$$

$e^{-\frac{1}{2}\left(\frac{x_1-\mu_m}{\sigma_m}\right)^2}$ $e^{-\frac{1}{2}\left(\frac{x_1-\mu_{m-1}}{\sigma_{m-1}}\right)^2}$

where, $\phi_m(x_1) = e^{-\frac{1}{2}\left(\frac{x_1-\mu_m}{\sigma_m}\right)^2}$

$$\phi_1(x_L) = e^{-\frac{1}{2}\left(\frac{x_L-\mu_1}{\sigma_1}\right)^2}$$

Basis

Sigmoidal function:

$$\phi_m(x) = \frac{1}{1 + e^{-(\beta_m^T x + \beta_m^0)}}$$

($x \in R^n$; $\beta_m \in R^n$; $\beta_m^0 \in R$)

β_m^0 is not power 0, it is a constant.

$$\phi_{m-1}(x) = \frac{1}{1 + e^{-(\beta_{m-1}^T x + \beta_{m-1}^0)}}$$

($\beta_{m-1} \in R^n$; $\beta_{m-1}^0 \in R$)

$$\phi_1(x) = \frac{1}{1 + e^{-(\beta_1^T x + \beta_1^0)}}$$

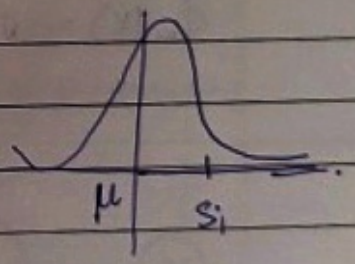
β_1^0 is not power 0, it is a constant.

Gaussian

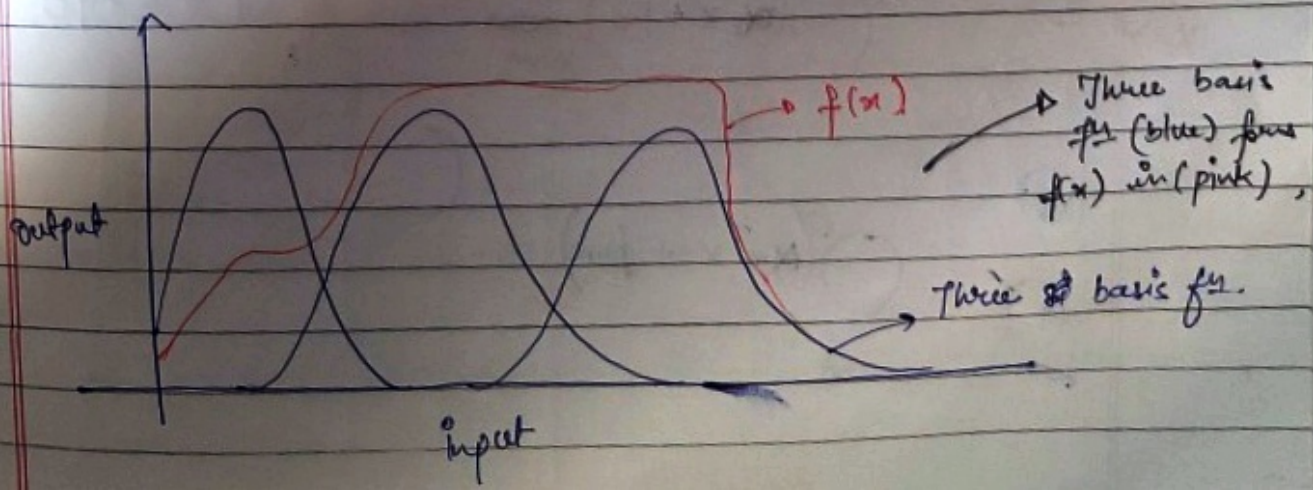
m-types of Gaussian functions.

$$\phi_j(x) = e^{-\left(\frac{1}{2\sigma_j^2} \|X - \mu_j\|^2\right)}$$

$$\Rightarrow \phi_1(x) = e^{-\frac{1}{2\sigma_1^2} \|X - \mu_1\|^2}$$



$$\phi_m(x) = e^{-\frac{1}{2\sigma_m^2} \|X - \mu_m\|^2}$$



20/2/22

$$\begin{bmatrix} w \\ b \end{bmatrix} = (A^T A)^{-1} A^T y$$

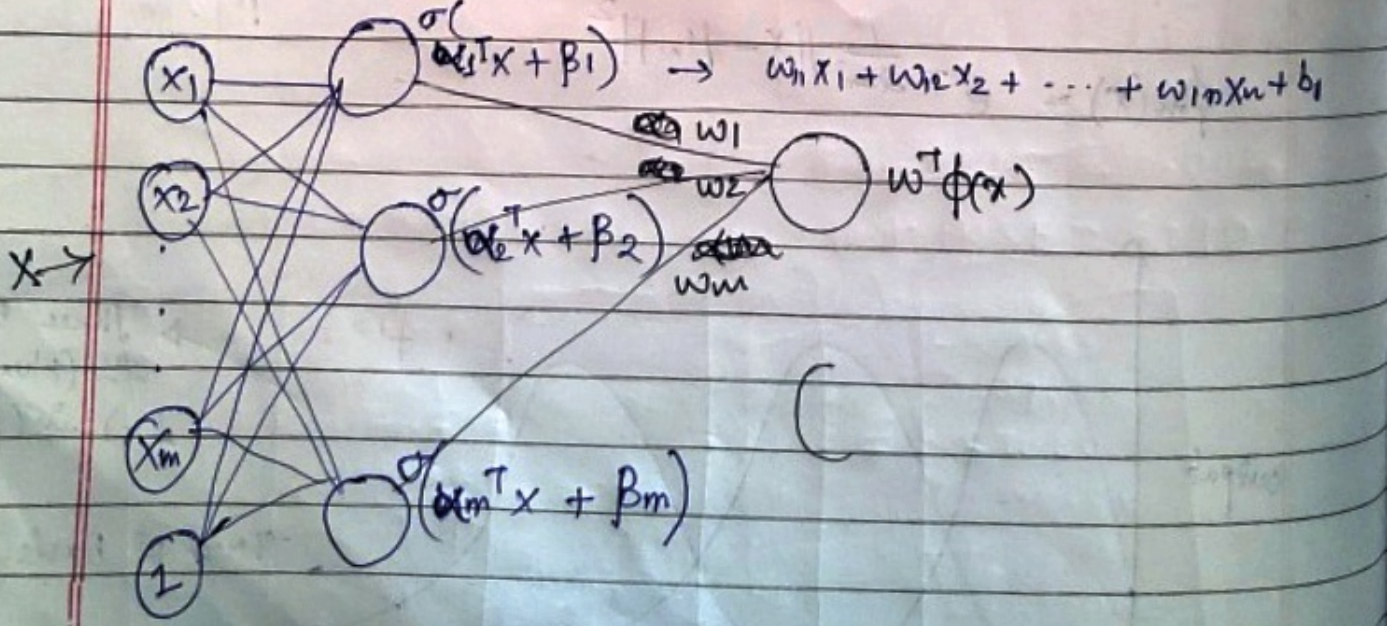
can be obtained by

Sigmoidal fn :

$$\phi_1(x) = \sigma(\alpha_1, \beta_1, x) = \frac{1}{1 + e^{-(\alpha_1^T x + \beta_1)}}$$

$$\phi_2(x) = \sigma(\alpha_2, \beta_2, x) = \frac{1}{1 + e^{-(\alpha_2^T x + \beta_2)}}$$

$$\phi_m(x) = \sigma(\alpha_m, \beta_m, x) = \frac{1}{1 + e^{-(\alpha_m^T x + \beta_m)}}$$



$$w_1\left(\sigma(\alpha_1^T x + \beta_1)\right) + w_2\left(\sigma(\alpha_2^T x + \beta_2)\right) + \cdots + w_m\left(\sigma(\alpha_m^T x + \beta_m)\right) + b$$

→ bias point

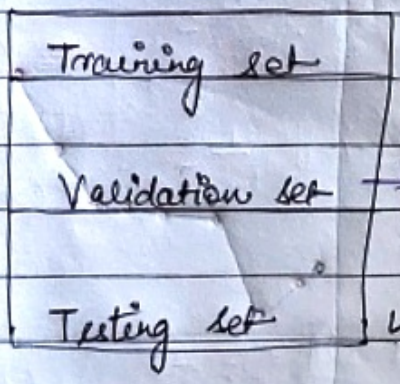
$$f(x) = w^T \phi(x) + b$$

for m data pts → $w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_m\phi_m(x) + b$

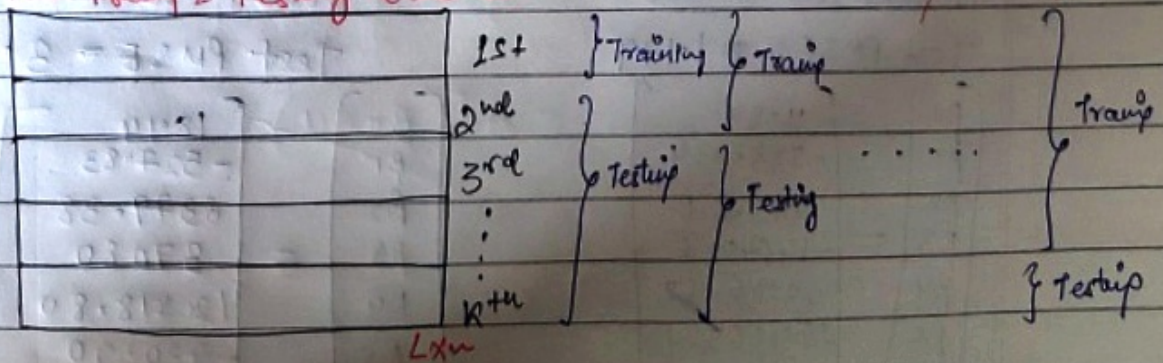
$$f(x) = w^T \phi(x) + b$$

for (m+1) data pts. → $w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_m\phi_m(x) + w_{m+1}\phi_{m+1}(x) + b$

* if we consider $w_{m+1}\phi_{m+1}(x)$ as 1, then, for (m+1) data pts., f(x) will be equal to for m data pts.



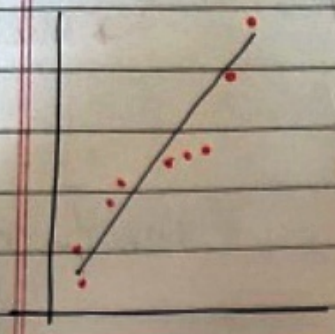
* K-fold Training & Testing criteria



So, we have (K) RMSE, since we have K^{th} data pts.

22/2/23

M = 2

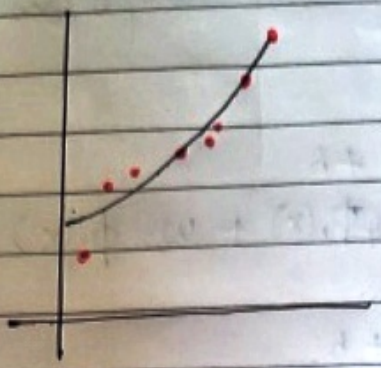


Train RMSE = 0.5947

Test RMSE = 0.9426

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 2.6 \\ 6.9 \end{bmatrix}$$

M = 3

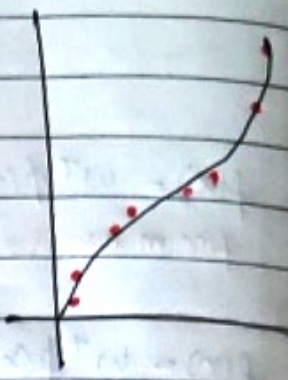


Train RMSE = 0.4980

Test RMSE = 0.7711

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 3.36 \\ 2.51 \\ 4.58 \end{bmatrix}$$

M = 5

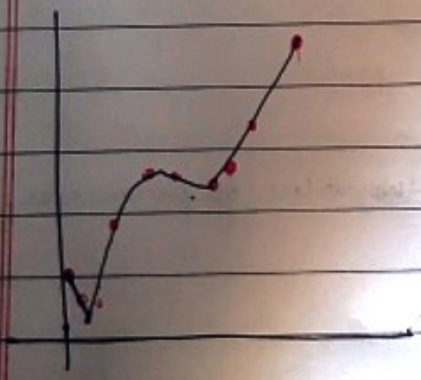


Train RMSE = 0.4937

Test RMSE = 0.9811

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 2.62 \\ 11.14 \\ 15.58 \\ 9.19 \\ 4.26 \end{bmatrix}$$

M = 8

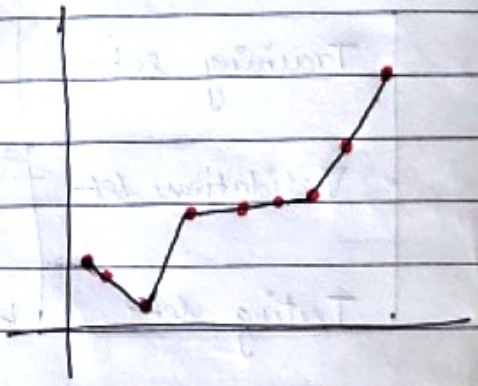


Train RMSE = 0.1186

Test RMSE = 1.179

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix} = \begin{bmatrix} 11.89 \\ -283.034 \\ 3015.61 \\ -14643.7 \\ 38006.62 \\ -54565.9 \\ 40844.45 \\ -12458.6 \end{bmatrix}$$

M = 9



Train RMSE = 0.0026

Test RMSE = 3.90

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} = \begin{bmatrix} 18.14 \\ -527.83 \\ 6379.38 \\ 87080 \\ 120578.80 \\ -230390 \\ 256860.2 \\ 154208 \\ 308542.75 \end{bmatrix}$$

* we can see that value β_i is very large.

* problem of overfitting ⑧ increases, exponentially, with ↑ in dimensionality

* higher order of β, has more or very large values.

for M=8,

* $\left[\min \frac{1}{10}\left\{\sum_{i=1}^{10}\left(y_i-(\beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_7 x_i^7)\right)^2\right\}+\left\{\frac{\lambda}{2}(\beta_0^2+\beta_1^2+\cdots+\beta_2^2+\cdots+\beta_7^2\right\}\right]$ ← user defined

λ = regularisation parameter.

minimising sq. of second norm.

↑ value of λ, fn. will tend to constant.

* we ↓ the value of $\beta_0, \beta_1, \beta_2 \ldots$ by λ regularisation technique.

* We use regularisation technique

- to convert flexibility
- to decrease value of β_i, to control overfitting.

* one of method to find λ → grid search method.

$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_L,y_L)\}$

$x_i \in R^n$, $y_i \in R$

$f(x)=w^T \phi(x)+b$ ← diff. basis function

* Optimisation Problem →

$$\min_{w,b} \frac{\lambda}{2} w^T w + \frac{1}{2}\sum_{i=1}^{2}\left(y_i-(w^T\phi(x_i)+b)\right)^2$$

$w_1^2+w_2^2+\cdots$; $(w^T\phi(x_i)+b)$ ← least sq.

$\frac{d}{du}(u^T u)=2u$

$$\min_{w,b} \frac{\lambda}{2} u^T I_0 u + \frac{1}{2}\left((Y-Au)^T(Y-Au)\right)$$

derivative wrt u $= -A^T(Y-Au)$

$I_0 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & & 0 \end{bmatrix}$ $\quad u = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \\ b \end{bmatrix}_{(m+1)\times 1}$

$I_0 u^T u = \begin{bmatrix} w_1^2 \\ w_2^2 \\ \vdots \\ b^2 \end{bmatrix}$ ~~$\frac{\partial (AX)}{\partial x} = 2AX$~~

$u^T \Theta I_0 v = \begin{bmatrix} w_1^2 & 0 & \cdots & 0 \\ 0 & w_2^2 & \cdots & 0 \\ \vdots & \ddots & & 0 \\ 0 & & b & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$

$u = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \\ b \end{bmatrix} = (\lambda I_0 + A^T A)^{-1} A^T y$

$J(w) = w_1^2 + w_2^2$

$\nabla w = \frac{\partial J(w)}{\partial w_1} = 2w_1 = 0$

$\nabla w = \frac{\partial J(w)}{\partial w_2} = 2w_2 = 0$

✓* $\frac{\partial}{\partial x}(x^T A x) = 2Ax.$

✓* $\frac{\partial}{\partial v}\left(\frac{1}{2}\left[(y-Au)^T(y-Au)\right]\right) = \frac{1}{2} 2 \cdot (y-Au)(-A)^T$

$= -A^T(y-Au)$

✓* $\frac{\partial}{\partial u}(u^T u) = 2u$

✓* $\frac{\partial}{\partial u}\left(\frac{\lambda}{2} u^T I_0 u\right) = \frac{\lambda}{2} 2 u I_0 = \lambda I_0 u$

* Regularised least square Regression model.

* Grid search method

2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}

λ

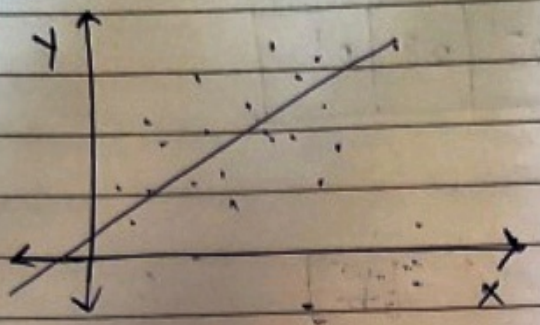
Training set → $\begin{bmatrix} w \\ b \end{bmatrix}$

Validation set

* We choose one value of λ, then perform prediction on training set & then validation set and record the value.
We then choose another value of ~~Foo~~ λ, then perform same thing on Training & then on validation set.
Save Best value of Training set & validation set, we choose that λ for our Regularised least square Regression model.

23/2/23

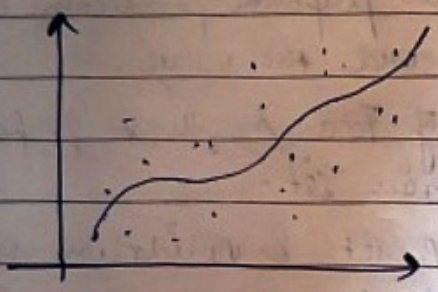
$f(x) = w^T x + b \quad , x \in R^n, w \in R^n, b \in R.$



(linear)

$f(x) = w^T \phi(x) + b,$ $\quad w \in R^m$

$$\phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_m(x) \end{bmatrix} \quad m \to \infty$$

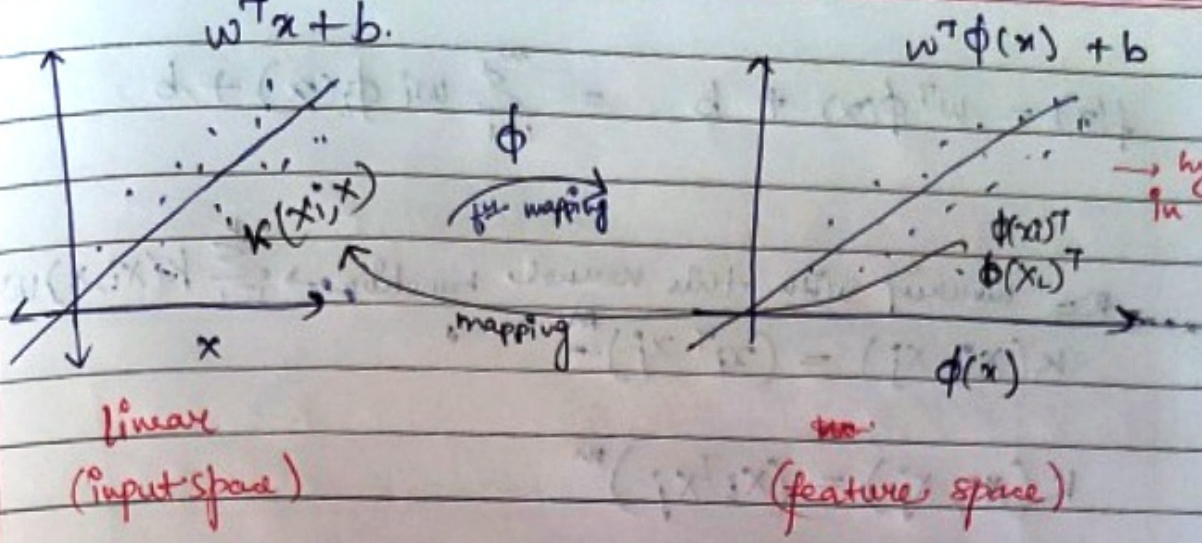


(non-linear)

* mapping x to $\phi(x)$ when we work with non-linear ϕ. So, at last, we are working with linear function $(w^T \phi(x) + b)$

$w^T \phi(x) + b$ is a line in m-dimension. for $\phi(x) = \phi_1(x), \ldots \phi_m(x)$

* jitne no. of pts in $\phi(x)$, utne ka dimension hoga.



Linear (input space)

(feature space)

$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$

$x_i \in R^n$

$y_i \in R$

$W = \sum_{i=1}^{L} v_i \phi(x_i)$

$= v_1 \phi(x_1) + v_2 \phi(x_2) + \ldots + v_L \phi(x_L)$

* In higher dimension space, w can be obtained by linear comb^n of L training pts. in feature space.

→ $w^T \phi(x) + b$

⇒ $\left(\sum_{i=1}^{L} v_i \phi(x_i)\right)^T \phi(x) + b$

⇒ $\sum_{i=1}^{L} v_i \phi(x_i)^T \phi(x) + b$

* Kernel generated function

$\sum v_i K(x_i, x) + b$

↳ this will tell in higher dim^n how $\phi(x)$ looks like.

$$f(x) = w^T \phi(x) + b = \sum_{i=1}^{m} w_i \phi_i(x) + b$$

working with this kernel function $\rightarrow \sum_{i=1}^{L} K(x_i, x) w_i + b$

$$K(x_i, x_j) = (x_i^T x_j)^m$$

$$K(x_i, x_j) = (x_i^T x_j)^m$$

* If we are working with Gaussian basis fn & have taken m no. of basis, then we use this type of Kernel fn

$$K(x_i, x_j) = \exp \frac{-\|x_i - x_j\|^2}{\sigma^2} = \phi(x_i)^T \phi(x_j)$$

→ Kernel fn. for Gaussian basis fn

27/2/23

Fitting the linear function

$$f(x) = w_1 x_1 + w_2 x_2 + b. \longrightarrow \text{linear}$$

$$f(x) = w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2 + w_4 x_1 + w_5 x_2 + b$$

Kernel function

x_1	x_2
-	-
-	-
-	-

ϕ

x_1	x_2	x_3
-	-	-
-	-	-
-	-	-

$$\phi(x) = \phi\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \end{bmatrix}$$

$w_1x_1^2 + w_2x_2^2 + \sqrt{2}\,w_3x_1x_2 + b$ → linear fn. in 3-D.

$w_1x_1^2 + w_2x_2^2 + \sqrt{2}\,w_3x_1x_2 + w_4x_1 + w_5x_2 + b$

↳ Linear fn. in 5D

$\phi(x)^T\phi(y) = ?$

$$\begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} y_1^2 \\ y_2^2 \\ \sqrt{2}y_1y_2 \\ y_1 \\ y_2 \end{bmatrix} = (x_1y_1)^2 + (x_2y_2)^T + 2x_1x_2y_1y_2 + x_1y_1 + x_2y_2$$

$$= (x^Ty)^2 + (x^Ty)$$

$\therefore \phi(x)^T\phi(y) = (x^Ty)^2 + (x^Ty)$

$\boxed{K(x,y) = (x^Ty)^2 + x^Ty}$ → Kernel Generated fn.

$\sum_{i=1}^{} K(x_i, x)\, u_i + b$

	x_1	x_2
x_1	5.9	3
x_2	6.9	3.1
x_3	6.6	2.9
x_4	4.6	3.2
x_5	6	2.2

$\boxed{\sum_{i=1}^{} K(x_i, X)\, u_i + b}$

$u_i = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ b \end{bmatrix}$

$$K\left(\begin{bmatrix} 5.9 \\ 3 \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \left((5.9x_1 + 3x_2)^2 + (5.9x_1 + 3x_2)u_1 + b\right.$$

u_i → we need to find. $\begin{bmatrix} w_2 \\ w_1 \\ \vdots \\ b \end{bmatrix}$

b → bias term

* For every kernel fn., we have a kernel matrix.

$T = \{(x_1, y_1), (x_2, y_2) \cdots, (x_L, y_L)\}$ $x_i \in R^n$, $y_i \in R$.

Kernel Matrix:

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_L) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_L) \\ \vdots & & & \\ K(x_L, x_1) & K(x_L, x_2) & & K(x_L, x_L) \end{bmatrix}$$

- always +ve semi definite
- symmetric matrix

$K(X, Y) = (c + X^T Y)^2$

$c \rightarrow$ constant (user defined)

$K(X, Y) = (c + X^T Y)^m \rightarrow$ for polynomial of order m.

$$K(X, Y) = \exp^{\frac{-\|X-Y\|^2}{2\sigma^2}}$$

$\sigma^2 \rightarrow$ user defined

This kernel fn can be used to learn any type of function in any dimension

→ Implicit map Kernel $\rightarrow$ feature space.

$$\min_{(u,b)} \frac{\lambda}{2} u^T u + \sum_{i=1}^{L} \left(y_i - \left(\sum_{j=1}^{L} K(x_j, x_i) u_j + b \right) \right)^2$$

$$\begin{bmatrix} u \\ b \end{bmatrix} = (A^T A + \lambda I)^{-1} A^T y$$

$$A = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_L) & 1 \\ K(x_2,1) & K(x_2,x_2) & \cdots & K(x_2,x_L) & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ K(x_L,x_1) & K(x_L,x_2) & \cdots & K(x_L,x_L) & 1 \end{bmatrix}$$

18/23

Given the training set,

$T = \{(x_i, y_i) : x_i \in R^n, y_i \in R\}$

$f(x) = w^T x + b$

$w \in R^n$

$b \in R$.

$= \min_{(w,b)} J(w,b)$

We are solving

$$\min_{w,b} \frac{\lambda}{2} w^T w + \frac{1}{2L} \sum_{i=1}^{L} (y_i - (w^T x_i + b))^2.$$

- $\frac{d(x^T A x)}{dx} = 2Ax$

$\frac{d(\lambda w^T w)}{dw} = \frac{\lambda \cdot 2w}{2} = \lambda w$

Gradient Descent Method

$$\frac{\partial}{\partial w} = \nabla_w J(w,b) = \lambda w - \frac{2}{L} \sum_{i=1}^{L} x_i (y_i - (w^T x_i + b))$$

$$\frac{\partial}{\partial b} = \nabla_b J(w,b) = -\frac{2}{L} \sum_{i=1}^{L} (y_i - (w^T x_i + b))$$

$\begin{bmatrix} w^0 \\ b^0 \end{bmatrix} \in R^{n+1}$ ⟶ Initialise

Repeat

$$\begin{bmatrix} w^{k+1} \\ b^{k+1} \end{bmatrix} = \begin{bmatrix} w^k \\ b^k \end{bmatrix} - \eta \underbrace{\begin{bmatrix} \nabla_w J(w^k, b^k) \\ \nabla_b J(w^k, b^k) \end{bmatrix}}_{\text{Gradient}}$$

Until $\left\| \begin{bmatrix} \nabla_w J(w^k, b^k) \\ \nabla_b J(w^k, b^k) \end{bmatrix} \right\|_2 \leq \varepsilon$

→ * For any basis function :-

$$\min_{w,b} \frac{\lambda}{2} w^T w + \sum_{i=1}^{L} \left(y_i - (w^T \phi(x_i) + b) \right)^2 = \min_{(w,b)} J(w,b)$$

$$\nabla_w J(w,b) = \lambda w - \frac{2}{L} \sum_{i=1}^{L} \left(y_i - (w^T \phi(x_i) + b) \right) \phi(x_i)$$

$$\nabla_b J(w,b) = - \frac{2}{L} \sum_{i=1}^{L} \left(y_i - (w^T \phi(x_i) + b) \right)$$

* For Kernel Generated function :-

$$\min_{u,b} \frac{\lambda}{2} u^T u + \frac{1}{L} \sum_{i=1}^{L} \left(y_i - \left(\sum_{j=1}^{L} K(x_j, x_i) u_j + b \right) \right)^2$$

$$f(x) = \sum_{j=1}^{L} k(x_j, x) u_j + b$$

M 4-D ← $\nabla_u J(u,b) = \lambda u - \frac{2}{L} \sum_{i=1}^{L} \left(y_i - \left(\sum_{j=1}^{L} K(x_j, x_i) u_j + b \right) \right) K(x_j, x_i)$

M 1-D ← $\nabla_b J(u,b) = - \frac{2}{L} \sum_{i=1}^{L} \left(y_i - \left(\sum_{j=1}^{L} k(x_j, x_i) u_j + b \right) \right)$

$W = \begin{bmatrix} w \\ b \end{bmatrix}^{*} =$ (101 ×1)

$K = [\ ]_{100 \times 100}$

Date
Page

	x_1	x_2	y
x_1	0	0	0
x_2	0	1	0
x_3	1	1	0
x_4	1	0	1

$K(U,V) = (U^T V)^2$

↑ kernel fn

≠

~~$K(X,X) \neq (X^T X)^2$~~

write, Gradient Descent Algo.

→ Kernel matrix, $K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) \cdots K(x_1,x_L) \\ K(x_2,x_1) & K(x_2,x_2) \quad K(x_2,x_L) \\ \vdots \\ K(x_L,x_1) & K(x_L,x_2) \cdots K(x_L,x_L) \end{bmatrix}$

$K = \begin{bmatrix} \cancel{K(0,0)}\,0 & \cancel{K(0,0)}\,0 & \cancel{K(0,1)}\,0 & \cancel{K(0,1)}\,0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}_{4\times4}$

$\alpha_i = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$ $b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

3 $y_1 - \left(\sum_{j=1}^{4} K(x_j, x_1)\, \alpha_j + b\right) \sum_{j=1}^{4} K(x_j, x_1)$

2 $y_2 - \left(\sum_{j=1}^{4} K(x_j, x_2)\, \alpha_j + b\right)$

$0 - \Big(\big(K(x_1,x_1)\alpha_1 + K(x_2,x_1)\alpha_2 + K(x_3,x_1)\alpha_3 + K(x_4,x_1)\alpha_4\big) + b\Big)\big(K(x_1,x_1) \cdot K(x_2,x_1) \cdot K(x_3,x_1) \cdot K(x_4,x_1)\big)$

$= 0 - [0+0+0+0+b](0)$

$= 0$

$(x_1^T x_1)^2$ $(x_2^T x_1)^2$ $(x_2^T x_2)^2$

$\begin{bmatrix} 0 \\ 0 \end{bmatrix}_{2\times1} \begin{bmatrix} 0 & 0 \end{bmatrix}_{1\times2}$ $\begin{bmatrix} 0 \\ 1 \end{bmatrix}_{2\times1} \begin{bmatrix} 0 & 1 \end{bmatrix}_{1\times2}$

$x_1^T = [0\ \ 0]$

$x_2^T = [0\ \ 1]$

$x_3^T = [1\ \ 1]$

$x_4^T = [1\ \ 0]$

$(x_2^T x_1)^2 = [0\ 1] \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0^2 = 0$

$(x_2^T x_2)^2 = [0\ 1]_{1\times2} \begin{bmatrix} 0 \\ 1 \end{bmatrix}_{2\times1} = 1^2 = 1.$

13/3/23

Date
Page

Least Square Kernel Regression.

$$\min_{(u,b)} \frac{\lambda}{2} u^T u + \sum_{i=1}^{l} \left(y_i - \left(\sum_{j=1}^{l} K(x_j, x_i) u_j + b \right) \right)^2$$

$$\begin{bmatrix} u \\ b \end{bmatrix} = (A^T A)^{-1} A^T Y$$

$$A = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_l) & 1 \\ \vdots & & & & \\ K(x_l, x_1) & K(x_l, x_2) & \cdots & K(x_l, x_l) & 1 \end{bmatrix}$$

First let us consider

$$J(u, b, x_k, y_k) = \left(y_k - \left(\sum_{j=1}^{l} K(x_j, x_k) u_j + b \right) \right)^2$$

$$\frac{\partial J_u(u, b, x_k, y_k)}{\partial u} = -2 \left(y_k - \left(\sum_{j=1}^{l} K(x_j, x_k) u_j + b \right) \sum_{j=1}^{l} K(x_j, x_k) \right)$$

Differentiating this term wrt u.

$$\sum_{j=1}^{l} K(x_j, x_k) u_j = K(x_1, x_k) u_1 + K(x_2, x_k) u_2 + \cdots + K(x_l, x_k) u_l + b$$

$$= \begin{bmatrix} K(x_1, x_k) \\ K(x_2, x_k) \\ \vdots \\ K(x_l, x) \end{bmatrix}$$

$$\therefore \frac{\partial J(u, b, x_k, y_k)}{\partial u} = -2 \left(y_k - \left(\sum_{j=1}^{l} K(x_j, x_k) u_j + b \right) \right) \begin{bmatrix} K(x_1, x_k) \\ K(x_2, x_k) \\ \vdots \\ K(x_l, x) \end{bmatrix}$$

now wrt b,

$$\frac{\partial J(u,b,x_k,y_k)}{\partial b} = -2\left(y_k - \left(\sum_{j=1}^{l} K(x_j, x_k) u_j + b\right)\right)$$

Gradient Descent Least Square Kernel Regression

Algorithm : gradient descent method :-

Initialise $u^0 = u^{start} \in R^l$ and $b \in R$

Repeat

$$u^{(j+1)} = u^{(j)} - \eta_k\left(\lambda u + \sum_{i=1}^{l} \frac{\partial J(u,b,x_k,y_k)}{\partial u}\right)$$

η_k = step length

$$b^{(j+1)} = b^{(j)} - \eta_k\left(\sum_{i=1}^{l} \frac{\partial J(u,b,x_k,y_k)}{\partial b}\right)$$

Until $\ominus$

$$\left\| \begin{bmatrix} \lambda u + \sum_{i=1}^{l} \frac{\partial J(u,b,x_k,y_k)}{\partial u} \\ \sum_{i=1}^{l} \frac{\partial J(u,b,x_k,y_k)}{\partial b} \end{bmatrix} \right\| \leq \varepsilon$$

Stochastic Gradient Descent Least Square Regression

$$\min_{(w,b)} \frac{\lambda}{2} \cdot w^T w + \sum_{i=1}^{l} (y_i - (w^T x_i + b))^2$$

$$\frac{\partial}{\partial w} = \lambda w - 2 \sum_{i=1}^{l} (y_i - (w^T x_i + b)) x_i$$

$$J(w, b, x_k, y_k) = (y_k - (w^T x_k + b))^2$$

$$\nabla_w J(w, b, x_k, y_k) = -2(y_k - (w^T x_k + b)) x_k$$

$$\frac{\partial J(w, b, x_k, y_k)}{\partial b} = -2(y_k - (w^T x_k + b))$$

Algorithm : Stochastic Gradient Descent Method

* At every iteration, value of η_k should be changed.

Initialise $w^0 = w^{start} \in R^l$ and $b^0 \in R$

Repeat

Randomly select subset B from Training set T.

$$w^{(j+1)} = w^{(j)} - \eta_k \left(\lambda w + \sum_{(x_k, y_k) \in B} \frac{\partial J(w^{(j)}, b^{(j)}, x_k, y_k)}{\partial w} \right)$$

$$b^{(j+1)} = b^{(j)} - \eta_k \left(\sum_{(x_k, y_k) \in B} \frac{\partial J(w^{(j)}, b^{(j)}, x_k, y_k)}{\partial b} \right)$$

Until : $\left\| \begin{bmatrix} w^{(j+1)} \\ b^{(j+1)} \end{bmatrix} - \begin{bmatrix} w^{(j)} \\ b^{(j)} \end{bmatrix} \right\| \leq \varepsilon$

One of the method to reduce the value of eta.

$\eta = 0.2$

for $i = 1 : 1000$

$\eta_i = \frac{\eta}{(1 + 0.0008)}$

end.

other methods to reduce the value of η (eta)

- Line search method
- Momentum concept

Stochastic Gradient Descent, Kernel Method Regression (least square)

Algorithm : Stochastic gradient descent method

Initialize $x^0 = u^{start} \in R^2$ and $b \in R$

Repeat

Randomly select subset B from Training set T.

$$u^{(j+1)} = u^{(j)} - \eta_k \left(\lambda u + \sum_{(x_k, y_k) \in B} \frac{\partial J(u, b, x_k, y_k)}{\partial u} \right)$$

$$b^{(j+1)} = b^{(j)} - \eta_k \left(\sum_{(x_k, y_k) \in B} \frac{\partial J(u, b, x_k, y_k)}{\partial b} \right)$$

Until $\left\| \begin{bmatrix} u^{(j+1)} \\ b^{(j+1)} \end{bmatrix} - \begin{bmatrix} u^{(j)} \\ b^{(j)} \end{bmatrix} \right\| \leq \varepsilon$

(x_1, y_1) (x_2, y_2) . . . (x_k, y_k)

$f(x_i)$ $f(x_2)$ $f(x_k)$

- $RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - f(x_i))^2}$

- $MAE = \frac{1}{k}\sum_{i=1}^{k}|y_i - f(x_i)|$

- $NMSE = \frac{SSE}{SST} = \frac{\sum_{i=1}^{k}(y_i - f(x_i))^2}{\sum_{i=1}^{k}(y_i - \bar{y})^2}$ → Normalised Mean Sq. Error

SST → Sum of Sq. Training Deviation

Variance of predicted fx →

$$\frac{1}{k^2}\left(\sum_{i=1}^{k} f(x_i) - f(\bar{x})\right)^2$$

- $R^2 = \frac{\sum_{i=1}^{k}(y_i - \bar{y})^2}{\sum_{i=1}^{k}(f(x_i) - f(\bar{x}))^2}$

- $Sparsity = \frac{\#(w_i = 0)}{\#(w_i)}$ → $\frac{\text{no. of } (w_i = 0)}{\text{no. of } w_i}$

Sparsity Calculates how many $w_i = 0$, ie it Calculates simplicity of the model and how many total w_i

let,

$w_1 = 2.8$

$w_2 = 0$

$w_3 = 0$

$b = 0.4$

sparsity = 2/3 (here) $= \frac{\#(w_i = 0)}{\#w_i}$ $\quad \#w_i = 0$ & total $w_i = 3$

* Sparsity should be high is preferable.

& if $w_1 = 2.8$

$w_2 = 1.2$

$w_3 = 0.8$

$b = 0.4$,

then sparsity = 0, because no $w_i = 0$.

15/3/23

* MLE

Let $\{X_1, X_2, \ldots, X_n\} \sim D(\theta)$ (any distrb^n.)

$\{39.5, 68, 78, 105, 101, 82, \ldots\} \sim N(\mu, \sigma)$

we select parameter θ, for which has max^m. prob. given sample D.

ie, $\max_\theta P(\theta/D)$

$$= \max_\theta P(D/\theta) \cdot P(\theta)$$

Likelihood — expert value (prior info) → sometimes not given, so we maximise only $P(D/\theta)$

$$= \max_\theta P(D/\theta)$$

pts. in D are i.i.d.

$$\max_\theta P(D/\theta) = \max_\theta P(X_1, X_2, \ldots, X_n/\theta)$$

$$\max_\theta P(D/\theta) = \max_\theta \prod_{i=1}^{n} P(X_i/\theta) \cdot = \max_\theta \sum_{i=1}^{n} \log P(X_i/\theta)$$

* Let $\{X_1, X_2, \ldots\} \sim N(\mu, \bar{\sigma})$

$\bar{\sigma}$ is any const., then estimate μ.

$$\max_{\theta \mu} \left(\sum_{i=1}^{n} \log P(X_i/(\mu, \bar{\sigma})) \right) \longrightarrow \text{in case of normal distribution,}$$

$$J(\mu) = \max_\mu \sum_{i=1}^{n} \log \frac{1}{\bar{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\bar{\sigma}}\right)^2}$$

$$J(\mu) = \max_\mu \sum_{i=1}^{n} \log\left(\frac{1}{\bar{\sigma}\sqrt{2\pi}}\right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$$

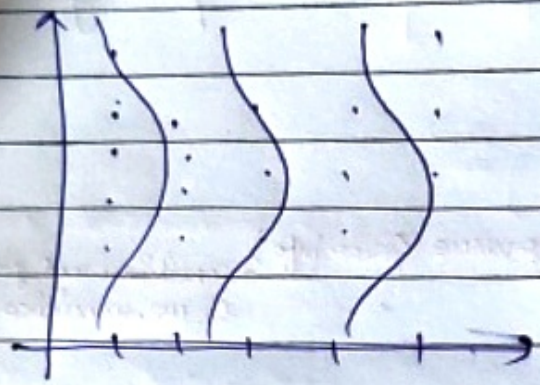
derivative $= 0$, ie $\frac{\partial J(\mu)}{\partial \mu} = 0$

$$\frac{\partial J(\mu)}{\partial \mu} \quad \frac{2}{2} \sum (x_i - \mu) = 0$$

$$\mu = \frac{1}{n} \Sigma x_i \qquad \boxed{\mu = \bar{x}}$$

* If we take μ as const. & want to estimate σ & maximize σ, then following same step we get as,

then $\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

→ 

$y_i = f(x_i) + \varepsilon_i$

$E(\varepsilon_i) = 0$

$E(y_i/X) \rightarrow w^T x_i + b$

$w^T \phi(x_i) + b$ (for basis fn.)

$y_i/X \sim N(w^T \phi(x_i) + b, \sigma)$

$\varepsilon_i = y_i - (w^T x_i + b) \sim N(0, \sigma)$

In regression problem, we have,

$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$

$x_i \in R^n, \quad y_i \in R.$

$Y_i/X_i \sim N(w^T \phi(x_i), \sigma)$

$\max P((Y_1, Y_2, \ldots, Y_L)/(X_1, X_2, \ldots, X_L))$

$\max P(Y_1/X_1, Y_2/X_2, \ldots, Y_L/X_L)$

$\max \prod_{i=1}^{L} P(Y_i/X_i)$

$= \max \prod_{i=1}^{L} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(Y_i - w^T\phi(x_i))^2}{2\sigma^2}} \quad (\because Y_i/X \sim N(w^T\phi(x_i), \sigma))$

$\propto \max \sum_{i=1}^{L} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(Y_i - w^T\phi(x_i))^2}{2\sigma^2}}\right)$

$= \max \sum_{i=1}^{L} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \sum_{i=1}^{L} \frac{(Y_i - w^T\phi(x_i))^2}{2\sigma^2}$

$\max -\sum_{i=1}^{L} (Y_i - (w^T\phi(x_i)))^2$

$\min_w \sum_{i=1}^{L} (Y_i - w^T\phi(x_i))^2$ → least squared loss fn.

Advantages :

This is eq. fn of w & b, smooth fn., solving system of equation is easy.

For solving $Y/X \sim N(\,)$, then, only this loss fn. will work.

Disadvantages

It is sensitive to outliers.

↓

To solve this, we need to solve or go for weighted least square regression model.