8/01/21

Assignment → 30%

Mini-project → 30%

Mid-sem → 20%

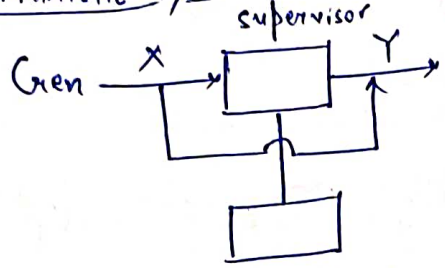
end-sem → 20%

presentation →

← Pattern-Recognition & Machine Learning →

18/01/21

Deterministic style →



features & labels →

features

labels.

$f_1, f_2, f_3 \ldots\ldots, f_n$ → $x_i \Rightarrow$

→ coal → ①.

→ Diamond → ②.

$Y_i \in \{1, 2\}$

↳ Here labels are represented in equeledean space of features in $\mathbb{R}^n$.

So → $T = \{(x_1, Y_1), (x_2, Y_2), \ldots\ldots (x_n, Y_n)\}$.

where $x_i \in \mathbb{R}^n$; and $Y_i \in \{1, 2\}$ ⇒ Example of 2-stone supervised learning.

Ex. ①. fish Recognition problem in duda Hart. } ⇒ Classification problem.

② Classification of Handwritten Numbers. }

③ Problem to find Expected income. } ⇒ Regression problem.

$T = \{(x_1, Y_1), (x_2, Y_2), \ldots\ldots (x_L, Y_L)\}$.

$x_i \in \mathbb{R}^n$, $Y_i \in \mathbb{R}$; $\forall i = 1, 2, 3, \ldots\ldots L$.

So → find a fn for Y for $x \in \mathbb{R}^n$

→ if ~~output~~ data is not labelled, then these problems are generally known as unsupervised learning. {if input data is not mapped to output labels.} ③. ⇒ {clustering problems.}

④.
→ Semi-supervised learning.

→ Regression prob. are more general & class^{n} problem is a particular prob. of regression.

20/01/2022 Linear Algebra →

① Metric space →

Let 'X' be a set. A metric $d: X * X \longrightarrow \mathbb{R}$ such that →

(i). $d(x,y) \geq 0$, $\forall x, y \in X$ and $d(x,y) = 0$ iff $x = y$.
(ii). $d(x,y) = d(y,x)$, $\forall x, y \in X$
(iii). $d(x,y) + d(y,z) \geq d(x,z)$, $\forall\ x, y, z \in X$

So → $A = \{(x_1, \ldots, x_n) : x_1, x_2, \ldots x_n \in \mathbb{R}\}$

Distance definition in A →

①. $\boxed{d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|}$: {Manhattan Distance}

②. $\boxed{d(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$ ⇒ $\boxed{(X-Y)^T \cdot (X-Y)}$ {Eudedean Distance}

③. $\boxed{d(x,y) = \max_{i=1,\ldots n} (|x_i - y_i|)}$; {Distance induced by infinity Norm}

④. $\boxed{d(X,Y) = X^T \Sigma^{-1} Y}$; {Mahalanobis Distance}

$$\Sigma = [x_1, \ldots x_n] \cdot \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \\ \sigma_{n1} & & \sigma_{nn} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Σ^{-1}

⇒ we can also write this as →

$$\Rightarrow \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2}$$

(2). Vector Space → all prop. of V.S. should satisfy.

⇓

* closed under addition
* " " scaler multiplication
* additive & multiplicative inverse should exist in set.

Ex → (1). set of all n-ordered tuples.

$$\mathbb{R}^n = \{(x_1, x_2, \ldots x_n) ; x_1, \ldots x_n \in \mathbb{R}\}.$$

→ $(P_1, P_2, \ldots, P_n)$

→ $(P_1', P_2', \ldots P_n')$

Ex → (2) set of all 2*2 matrix over the field of $\mathbb{R}$.

(3). → Subspaces → —

(4) → linearly dependence & independence →

A finite set of vectors is linearly dependent if at least one of them can be expressed as L.C. of others.

* → ie., that vector will lie in the Hyper plane of other vectors.

⑤. → span of a set → span(set) ⇒ Total no. of vectors that can be expressed using set.

⑥. → Basis of a set (or) space → Min.m no. of elements which can span over whole space.

⑦. → Inner Product space → Real 3 properties related to I.P.S. →

25|01|22 → all Rules & properties of basis →

⑧. linear mapping (or) linear transformation →

$$\forall x, y \in V, \forall \lambda, \psi \in R; \quad \boxed{\phi(\lambda x + \psi y) = \lambda \cdot \phi(x) + \psi \cdot \phi(y)}$$

⑨. Norm of a vector → (first Norm)

ⓐ. Manhattan Norm →

ⓑ. Euclidean Norm →

L_1- Norm
L_2- Norm

⑩. Bilinear Mapping →
(General inner product) →

⑪. standard inner product (Dot product) →

{ Note → Every G.I.P. can be written in form of a symmetric (+ve) semi-definite matrix.

$$\boxed{\langle x, Y \rangle = \hat{x}^T \cdot A \cdot \hat{Y}}$$

→ symmetric (+ve)- semi-definite matrix.

distance →

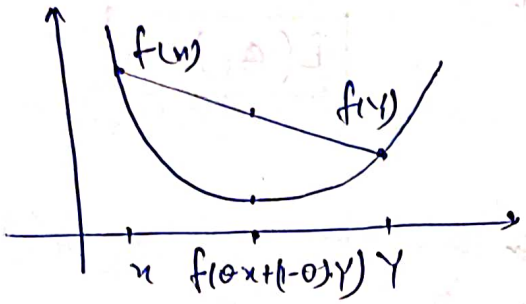
$$d(x,y) = \sqrt{\langle x-Y, x-Y \rangle} = \|x-Y\|$$

(12). Gradients & derivatives →

(13). Convex set →

(14). Convex function → {concept of convexity. ?}.

↳ for all values of 'x' & 'Y', this curve should have a lesser value than the line.



$$f(\theta x+(1-\theta).Y) \leq \theta.f(x)+(1-\theta).f(Y)$$

* (15). Constrained Convex Optimization → easier to find the minima.

(Ex) → find a fn to check how weight & height are related?

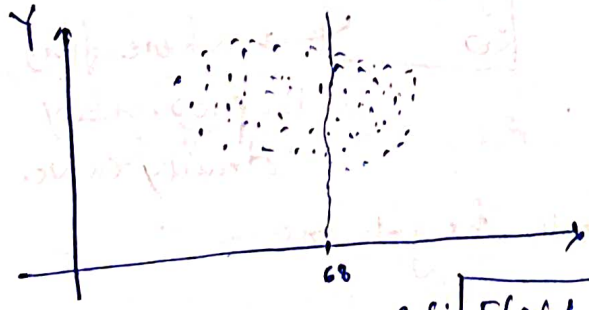
x (weight)	f	Y (height)
68	→	170
72	→	168
110	→	180
69	→	170
65	→	180
68	→	159

Ans → step(1) → Plot them first.

Density & Distribution of data

↳ check for a point

↳ checks in a interval.



So, we can cal.^{c} $F(Y/x)$ →

To cal.^{c} Y @ x = 68, we'll take mean of all the values.

→ Bcoz mean is a pts. where sum of square of deviations from ~~mean~~ that pt. is zero. ⇒ $\boxed{\text{MIN.} \sum_{i=1}^{n} (x_i - c)^2}$

So, mathematically →

$$\boxed{y_i = E(Y/x_i) + \epsilon_i}$$

→ distance from mean.

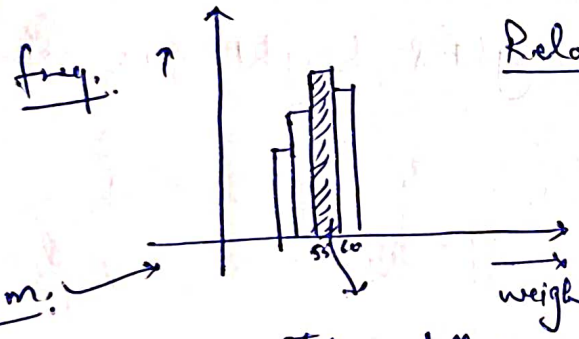
→ So, we want to find a fn that can generalise that.

$\boxed{E(\epsilon_i) = 0}$

** Assumptions of Regression Model →

27/01/2022

(Ex①) → Weight of students → 100 weights.

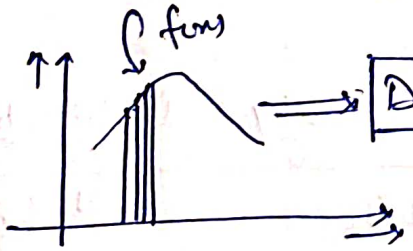


Relative freq. for (55-60) range → $\left(\frac{k}{n \times 5}\right)$

histogram:

→ This rep.n is not good bcoz data value (what it is exactly) will be lost in the range.

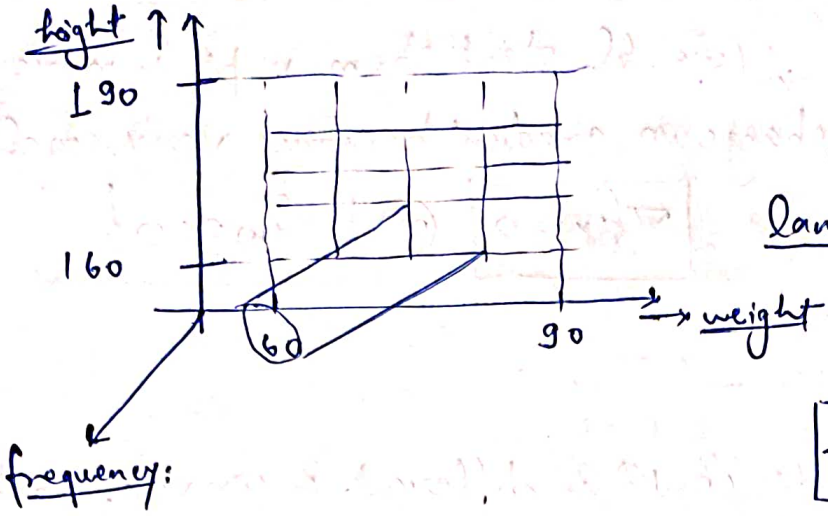
↳ To solve this, we can reduce the sizes of 'bin'.



⟹ Density Curve → $\int_{60}^{100} f(x).dx$ → where f(x) is the probability density curve.

Ex② → if in addition to prev. data, heights are also given.

⇒ We can map in 3-D.



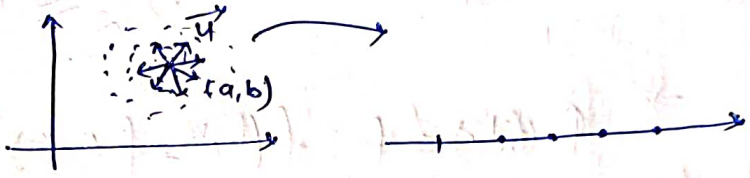
⇒ We'll get a landscape structure., that $f(x,y)$ will be the density fn.

$$f(x,y) = \frac{1}{\quad}$$

→ Convex Optimization →

Ex → $f: \mathbb{R}^2 \to \mathbb{R}$

$f(x,y) = x^2 + y^2$ ⟹



→ We want to move in the dir.^{n} such that change is max.^{m}

So → $\vec{u} \cdot \nabla f(a,b)$

$$\arg\max_{u} \nabla f(x)^T \cdot u$$

$$u = \frac{\nabla f(x)}{\|\nabla f(x)\|}$$ { Dir.^{n} of Max.^{m} ascent }.

r → x

find Y, s.t. $x^T Y \to$

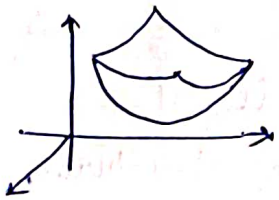
$$Y = \frac{x}{\|x\|}$$

So → Max.^{m} rate of change → $\frac{\nabla f(x)^T \cdot \nabla f(x)}{\|\nabla f(x)\|}$

⇒ $\|\nabla f(x)\|$

$$u = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$ { Dir.^{n} of Max.^{m} Descent }

Gradient Descent Algorithm →



(But f_x should be convex).

⇒ we'll start from ± pt. & move until change in gradient becomes very small.

$\nabla f(x) = 0$ (or) $\nabla f(x) \approx 0$

Definition → if 'f' is cont^{n} & different. & convex, such that →

$f: \mathbb{R}^n \to \mathbb{R}$;

$\| \nabla f(x) - \nabla f(y) \|_2 \leq \| x - y \|_2$

⇒ ~~f(x^k) ≤ f~~ $f(x^k - f(x^*)) \leq \frac{\| x^{(0)} - x^* \|_2^2}{2tk}$; $L > 0$; $t \to \gamma$

step length

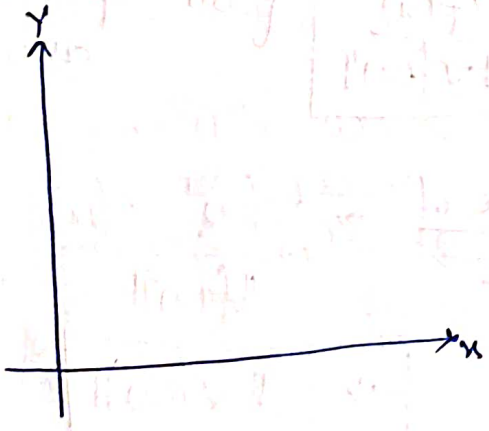
$x^0 \to x' \to \cdots \to x^k \to x^*$

x^0 → initial step.

x^k → k^{th} step

x^* → optimal

⇒



$f: X \to Y$

Target: $E(Y/X)$ →

$T = \{x_1, Y_1, \ldots, x_n, Y_n\}$.

$Y_i = f(x_i) + \epsilon_i$; $E(\epsilon) = 0$

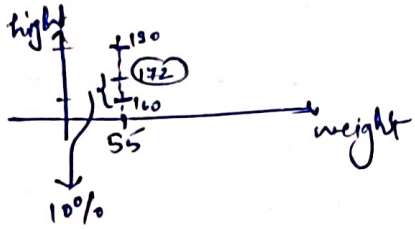
$E(Y/x_i)$

→ it is called Mean Regression Model.

⇒ Quantile Regression Model →

we are interested in fn 'f', such that for a given value of (x) → $\boxed{P(Y \leq f(x)) = \tau}$

Ex → given weight, what are lowest 10% values & top 10% value.



$\Rightarrow \boxed{P(172 \leq f(55)) = 0.1}$

→ This ~~inform~~ model will give more information.

Ex → (i). in Mean Regression Model → mean hight of all people having weight 55.

(ii). in Quantile Regression Model → it can give percentage boundary as well. based of value of τ.

⇒ ~~Empirical~~ Emperical loss minimization →

$\boxed{f : X \rightarrow Y}$; such that: $\boxed{Y = f^*(x_i) + \epsilon_i}$; $f^*(x_i) \rightarrow E(Y/x_i)$

To find the 'f', we can say that →

$\boxed{L(f(x_i) - Y_i)}$ should be minimum $\forall i$.

So → $L(f(x) - Y) = (f(x) - Y)^2$; ; $x_i \in \mathbb{R}^n$, $Y \in \mathbb{R}$

$L(f(x) - Y) = |f(x) - Y|$;

→ Considering L is an arbitary loss fn., then →

$$\sum_i L(f(x_i) - Y_i) = \int L(f(x_i) - Y_i) \cdot dP(x, Y)$$

$dP(Y/x) \cdot P(x)$.

This should be minⁿ.

⇓

① Parametric M.L. modelling.

② Non-Parametric M.L. " .

⇒ if we have L pts →

Emphirical Risk minimization

$$\left\{ \frac{1}{L} \cdot \sum_{i=1}^{L} L(Y_i - f(x_i)) \right\} = \frac{1}{L} \cdot \sum (Y_i - f(x_i))^2$$

↳

$$\int L(Y_i - f(x_i)) \cdot dP(x, Y) \qquad \int (Y_i - f(x_i))^2 \cdot dP(x, Y)$$

Structured Risk Minimization

↳ This is to be minimized.

$f \in F_1 = \{$ set of linear fn $\} = w^T x + b$

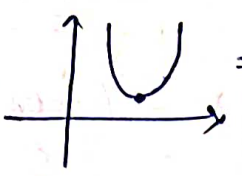
$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}; \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

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

$$L = (Y - f(x))^2 = u^2;$$

where $u = Y - f(x)$

least square loss fn.



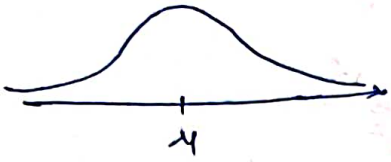
⇒ easy to find minima. bcoz convex fn

⇒ Now, we want to minimise →

$$\cdot \underset{\substack{w, b \\ \mathbb{R}^n \quad \mathbb{R}}}{\text{Min}} \quad \frac{1}{L} \cdot \sum_{i=1}^{L} (Y_i - (w^T x_i + b))^2$$

4/02/2022

Normal Distribution →



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

(μ, σ)

- 68% Data points → $(\mu-\sigma, \mu+\sigma)$
- 95% " " → $(\mu-2\sigma, \mu+2\sigma)$
- 99.7% " " → $(\mu-3\sigma, \mu+3\sigma)$

$X = (x, y)$

$H \quad w$

$$\bar{X} = (\bar{x}, \bar{y}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i, y_i)$$

Covariance of $(x,y) \Rightarrow$

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

$$Cov(x,Y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (Y_i - \bar{Y})$$

Que → How can be say that given set of data points will follow Normal Distribution?

$X = (x, y)$ → $X \sim N(\mu, \Sigma)$

$w^T X \sim N$

$$X = (x_1, x_2, \ldots\ldots x_n) \sim N(\mu, \Sigma)$$

$$w^T X \sim N(w^T \mu, w^T \Sigma w)$$

$\Rightarrow$

$X = (X, Y)$ { if X & Y are independent }.

$x = (x_1, x_2)$

So $\rightarrow$ $f(x) = f(x_1) \cdot f(x_2)$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$

$$\Rightarrow \frac{1}{(\sqrt{2\pi})^d \cdot |\Sigma|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right\}; \quad X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Σ $\rightarrow$ $\boxed{\text{cov}(X, Y) = 0}$

$\Rightarrow$ Probability ~~Distribution~~ Density fn of Guassian Distribution in 'd'-dim $\rightarrow$

$$f(x) = \frac{1}{(\sqrt{2\pi})^d \cdot |\Sigma|^{1/2}} \exp\left\{\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right\};$$

$x = (x_1, x_2, x_3, \ldots, x_n)$;

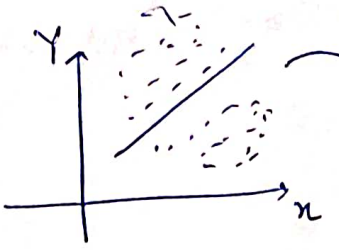
Contours $\rightarrow$ Contour pts. are where fn takes the constant value.

06/02/2022

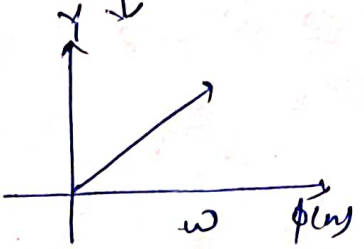
$\Rightarrow f(x) = w^T x + b$

$w \in R^m$

$b \in R$



$\Rightarrow \boxed{f(x) = w^T \phi(x)} \Rightarrow \phi(x) = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_n(x) \end{bmatrix} \rightarrow \perp \rightarrow \perp$



10/02/2022

Population 'θ' $\rightarrow \{x_1, x_2, x_3, \ldots, x_L\}$. $\rightarrow (\hat{\theta})$

$\boxed{N(\mu, \sigma)} \Rightarrow \boxed{P(\theta/T) = \frac{P(T/\theta) \cdot P(\theta)}{P(T)}}$

$\approx P(\theta/T) = P(T/\theta) = P(x_1, x_2, \ldots x_L/\theta)$

$\Rightarrow \prod_{i=1}^{L} P(x_i/\theta)^{-(\mu,\sigma)}$

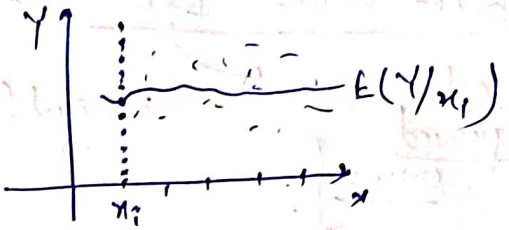
$\Rightarrow \prod_{i=1}^{L} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{ \frac{-1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}$.

$\text{Max} \log P(\theta/T) = \text{Max} \log \{ \quad \longrightarrow \quad \}$

$\boxed{\mu = \frac{\sum_{i=1}^{n} x_i}{n}}$; $\boxed{\sigma = \frac{\sum_{i=1}^{n} x_i - \mu}{n-1}}$;

$Y_i = E(Y/x_i) + \epsilon$; $\epsilon \sim N$; $E(\epsilon) = 0$

$f(x)$



→ Which loss fn to use in which case →

① So, we can say that (Y/x) is following Normal distribution with mean $f(x)$.

$$(Y/x) \sim N(f(x), \beta)$$

$\rightarrow w^T \phi(x)$

$w^T x$

$$P(Y/X, w) = N(Y/w^T\phi(x), \beta)$$

suppose → datapoints ⇒ $\{(x_1, Y_1), (x_2, Y_2), \ldots (x_L, Y_L)\}$.

MAX $P(Y_1/x_1, w, \beta) = N(Y_1/x_1, w, \beta)$

So → Max $\prod_{i=1}^{L} N(Y_i/w^T\phi(x_i), \beta)$

$$\Rightarrow \text{Max} \prod_{i=1}^{L} \frac{1}{\beta\sqrt{2\pi}} \cdot \exp\left\{ - \frac{(Y_i - w^T\phi(x_i))^2}{2\beta^2} \right\}$$

$$\Rightarrow \text{Max} \log \cdot \prod_{i=1}^{L} \frac{1}{\beta\sqrt{2\pi}} \cdot \text{——}$$

$$\Rightarrow \text{Max} \quad \frac{L}{2}\log\beta - \frac{L}{2}\cdot\log(2\pi) - \beta \cdot \sum_{i=1}^{L} (Y_i - w^T\phi(x_i))^2$$

$$\Rightarrow \text{Max} \quad - \sum_{i=1}^{L} (Y_i - w^T\phi(x_i))^2$$

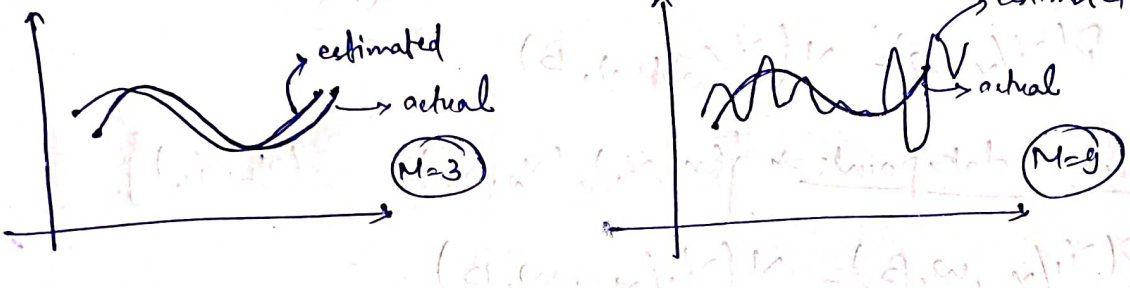
$$\text{Min}_{w} \quad \sum_{i=1}^{L} (Y_i - (w^T\phi(x_i)))^2$$

Problems with LSE (least squared error) → it is sensitive to the existence of outliers.

15th feb 2022

① → for higher degree, overfitting exict. bcoz we have made our curve ⓐ model more flexible, so, it is fitting on the noise values as well.

② → Overfitting occurs when our training sample size is small.



so, to correct this, we minimize as following →

$$\Rightarrow \min_{w,b} \frac{\lambda}{2} \|w\|_2^2 + \sum_{i=1}^{L} \left(y_i - (w^T \cdot \phi(x_i) + b) \right)^2;$$

we basically try to regularize the model.

so from Now onwards, we'll only this eq.ⁿ.

22/02/21 →

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

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

$$x \in \mathbb{R}^L = \sum_{i=1}$$

→ Gradient Descent →

&

→ Stochastic Gradient Descent →