CSE

Gradient descent

It is a optimization algorithm used to minimize the cost/loss function in machine learning.

$$\boxed{w = w - \alpha \frac{d}{dw} J(w,b)}$$

α = Learning rate

what is learning rate?

is like balance between moving fast and overshooting best solution

It is a parameter in machine learning and optimization algorithm (gradient descent) that determines the size of the steps taken during updates in training proces

TT-1 (*) what are potiential consequence of setting Learning rate high or low?

CSE

$$\boxed{w = w - \alpha \frac{d}{dw} J(w)}$$

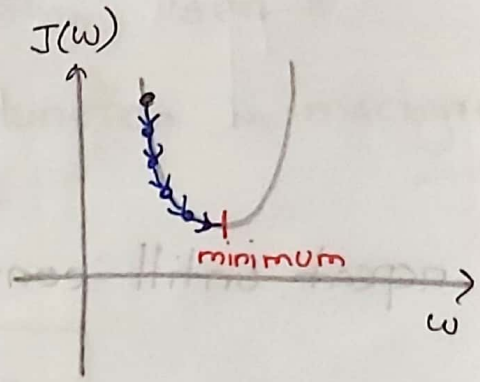
If α is too small

Gradient descent is too small

too slow

stuck in solution

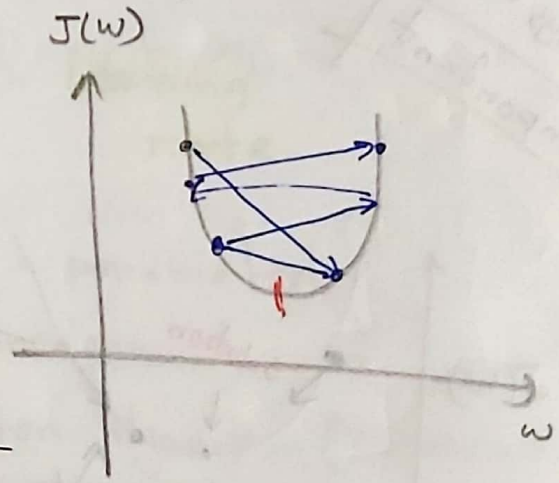
/ long time to reach



if α is too large

Gradient descent may overshoot and never reach the minimum

→ fail to converge, diverge



SWE-II

Q → is a model with Higher or lower variance better?

→ best using bias-variance tradeoff

SWE-17

CSE

Underfitting

→ when model is not able to capture the underlying trend of data

→ didnot learn enough from trained data

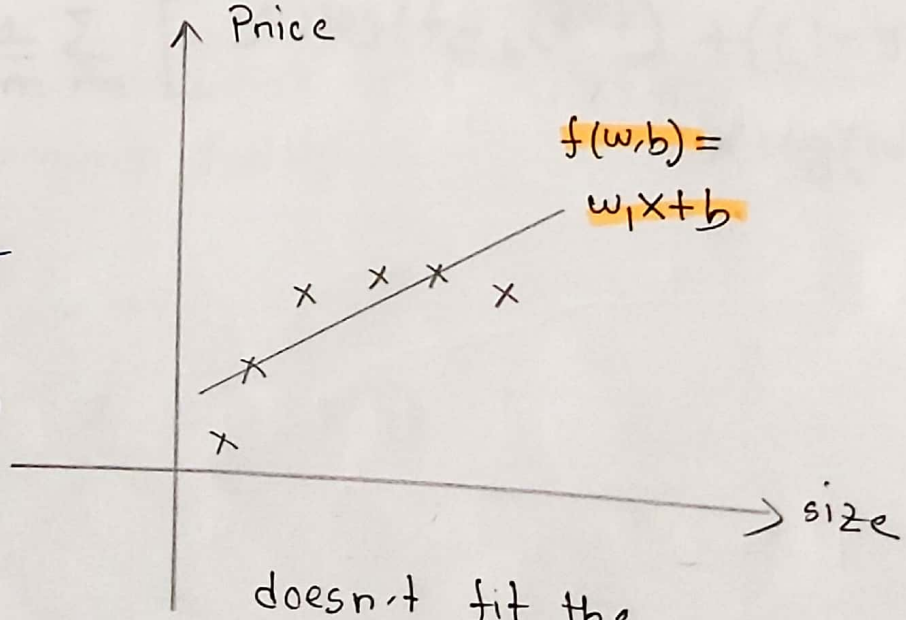
→ High bias → (too simple, incorrect prediction)

→ Low variance (doesn't adapt trained data)

→ Model has too few parameters / feature

advantage

- more stable
- less sensitive
- less chance overfitting



doesn't fit the training set well

SWE-17

what are causes of overfitting

16

CSE

overfitting

→ when model tries to coven all the data points / more than required data in dataset (unnecessary - irrelevant data)

→ caching noises and inaccurate values

→ Low bias (very complex , random fluctuations)

→ High variance (change prediction significantly)

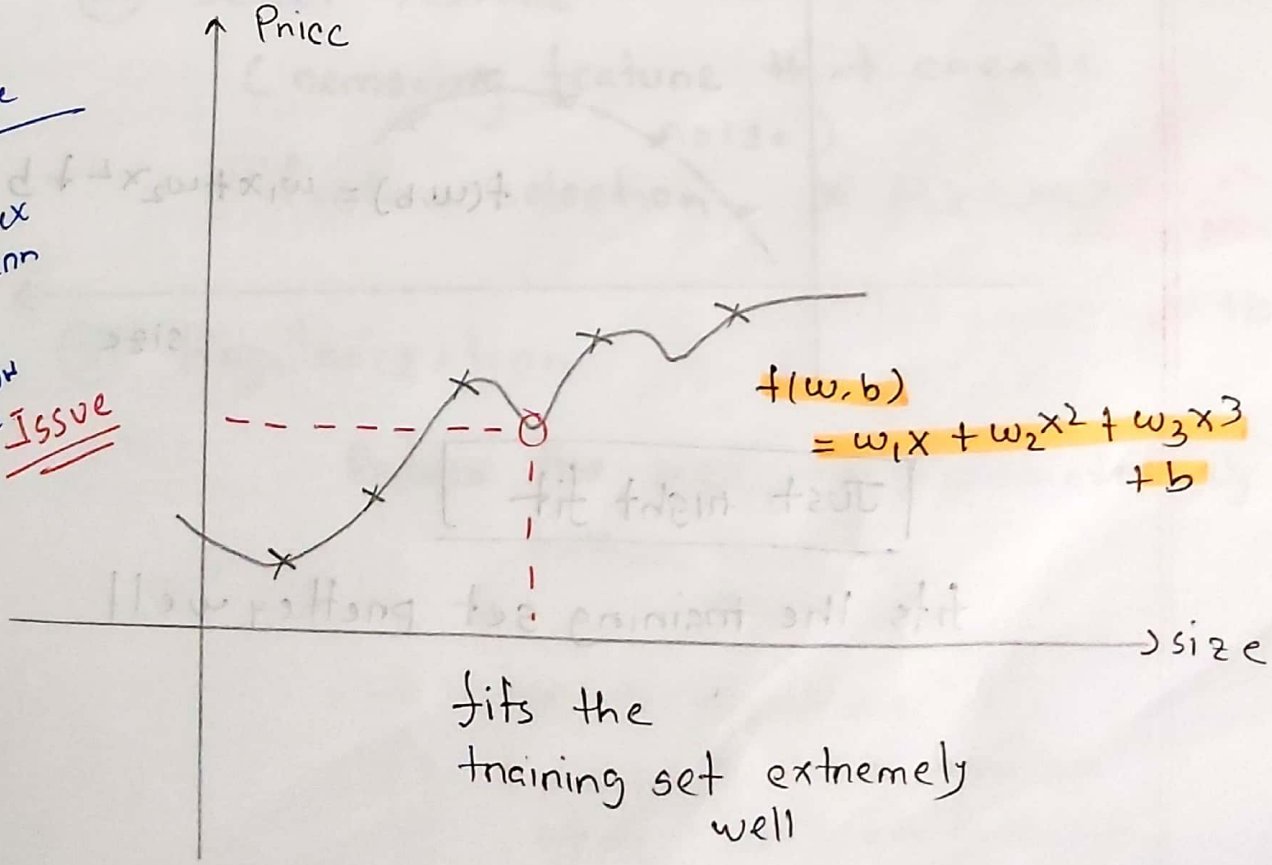
→ Model has too many parameters / feature

advantage

→ capture complex pattern

→ better prediction performance

Issue



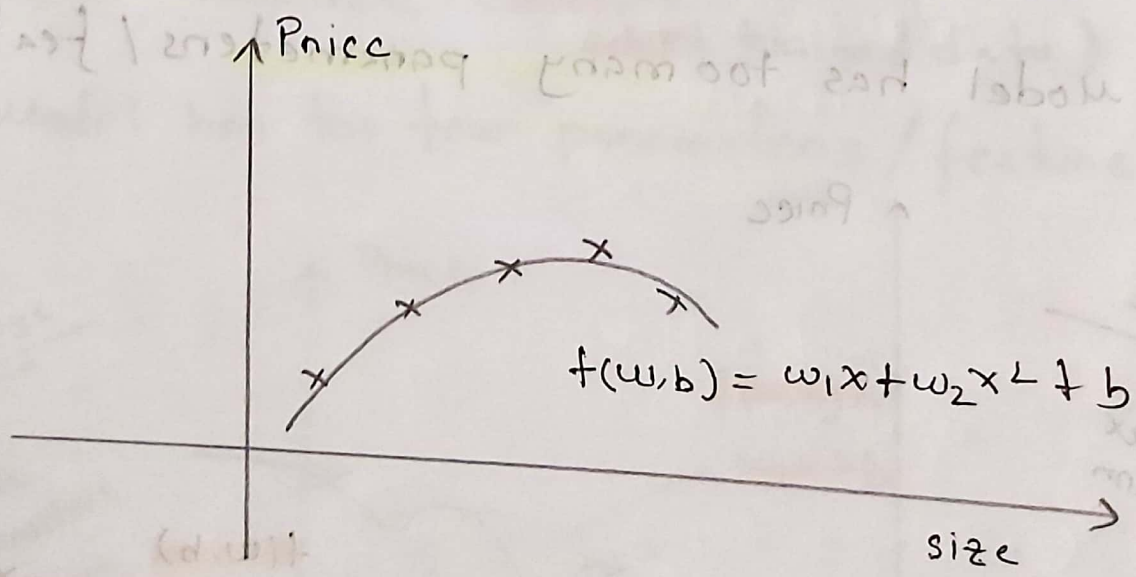
fits the training set extremely well

SWE-17

What is model Bias

Bias → represents how much the prediction of model differs from true values

Variance → measures how much the model's prediction fluctuate or vary when trained on different subsets of data



Just right fit

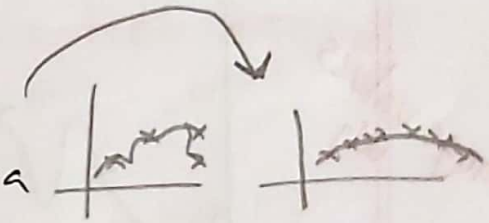
fits the training set pretty well

what is generalization?

refers to model ability to make accurate predictions on new, unseen data that has not been trained on.

SWE-17

How to avoid / solve overfiting?



1. Collect more trained data
2. Select feature to include / exclude
   (removing feature that create noise)
   feature selection

   ✗ $f() = w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 \ldots$

   ✓ $f() = w_1 x + w_2 x^2 + b$
3. Regularization

   Reduce the size of parameter w_j rather than elimination

   → takes all feature
   → but restrict them having large effect on dataset

SWE 17

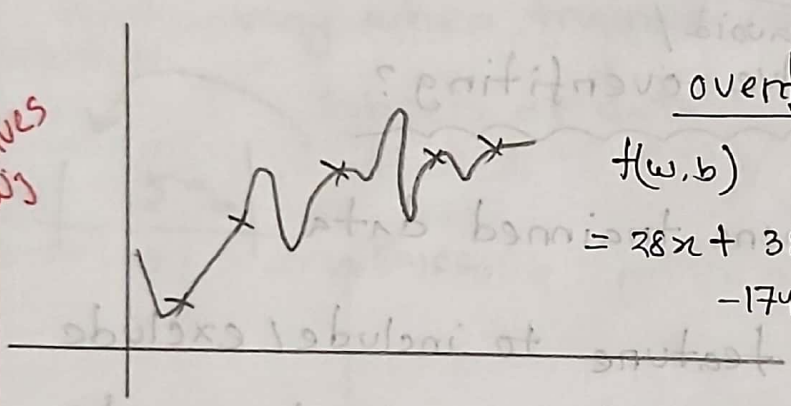
what is Regularization ?

CSF

Regularization

refers to regularize / reduce the magnitude of feature/parameter by keeping the same number of feature.

large values w_j



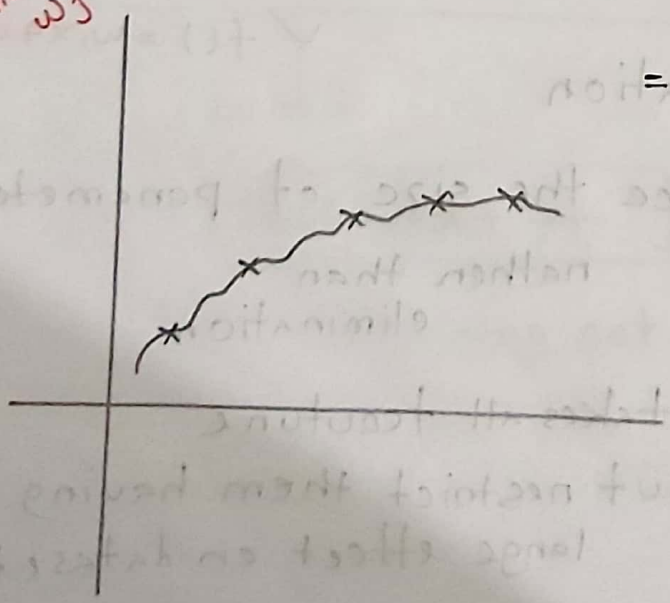
overfiting

$f(w,b)$

$= 28x + 385x^2 + 39x^3 - 174x^4 + 100$

small values w_j

after doing Regularization



$= 13x + (-.23x^2) + 0.000014x^3 - .0001x^4 + 10$

Linear ngression

Cost function of Regularization :-

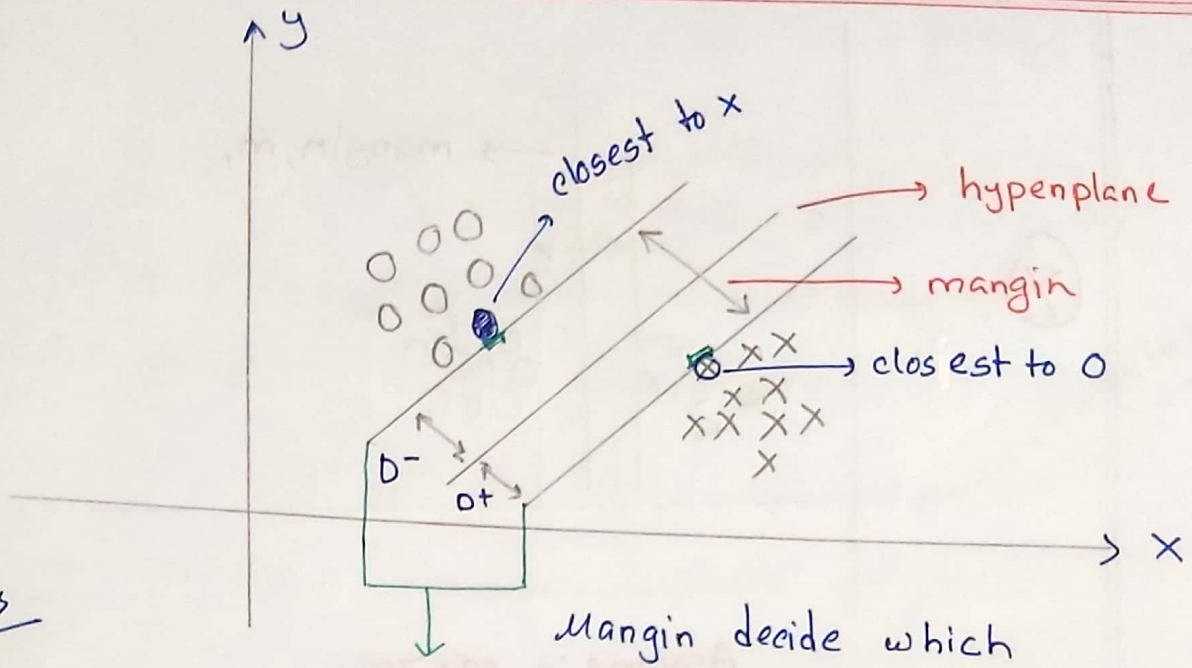
$$J(\vec{w}, b) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2 + \underbrace{\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2}_{\text{Regularization term}}\right]$$

Regularization parameter

Gradient descent

$$w_j = \underbrace{w_j\left(1 - \alpha\frac{\lambda}{m}\right)}_{\text{Shrinking } w_j} - \alpha\frac{1}{m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

Q. what are support Vectors?

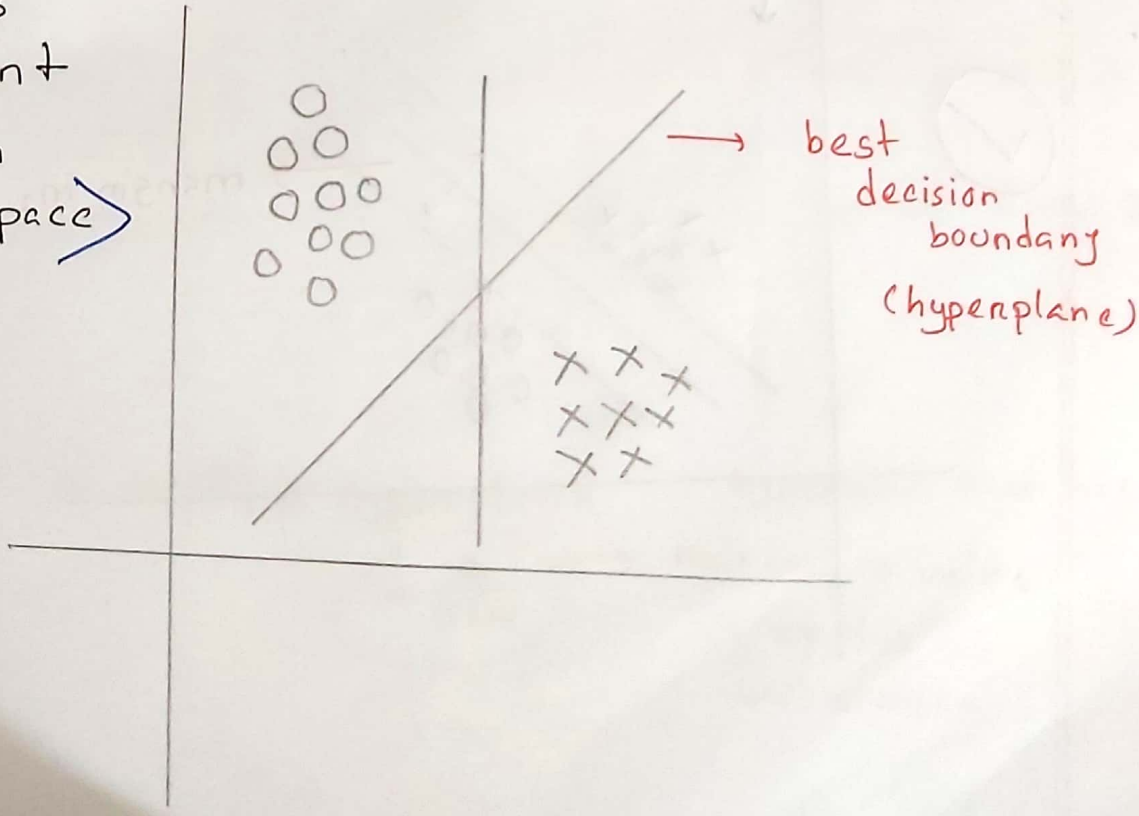


Support vectors

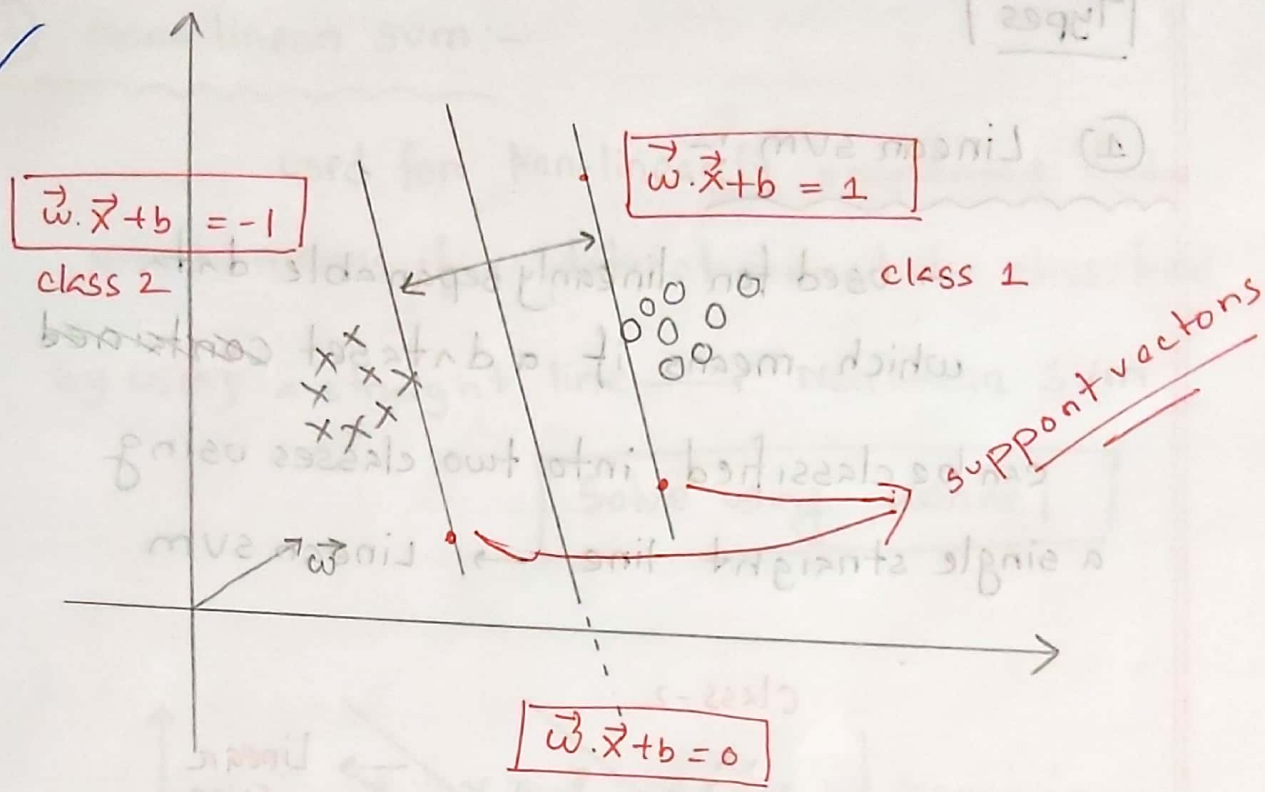
Mangin decide which hypERplane to choose?

Support vectors

SV are the data points that lie closest to the decision boundary (hyperplane) that seperates different classes in feature space



Q. How do support vector help to find the optimal mangin of model?



$(w \cdot x + b) \geq 1$; $f_{w,b} = 1$ $\forall x$ of class 1

$(w \cdot x + b) \leq -1$: $\forall x$ of class 2

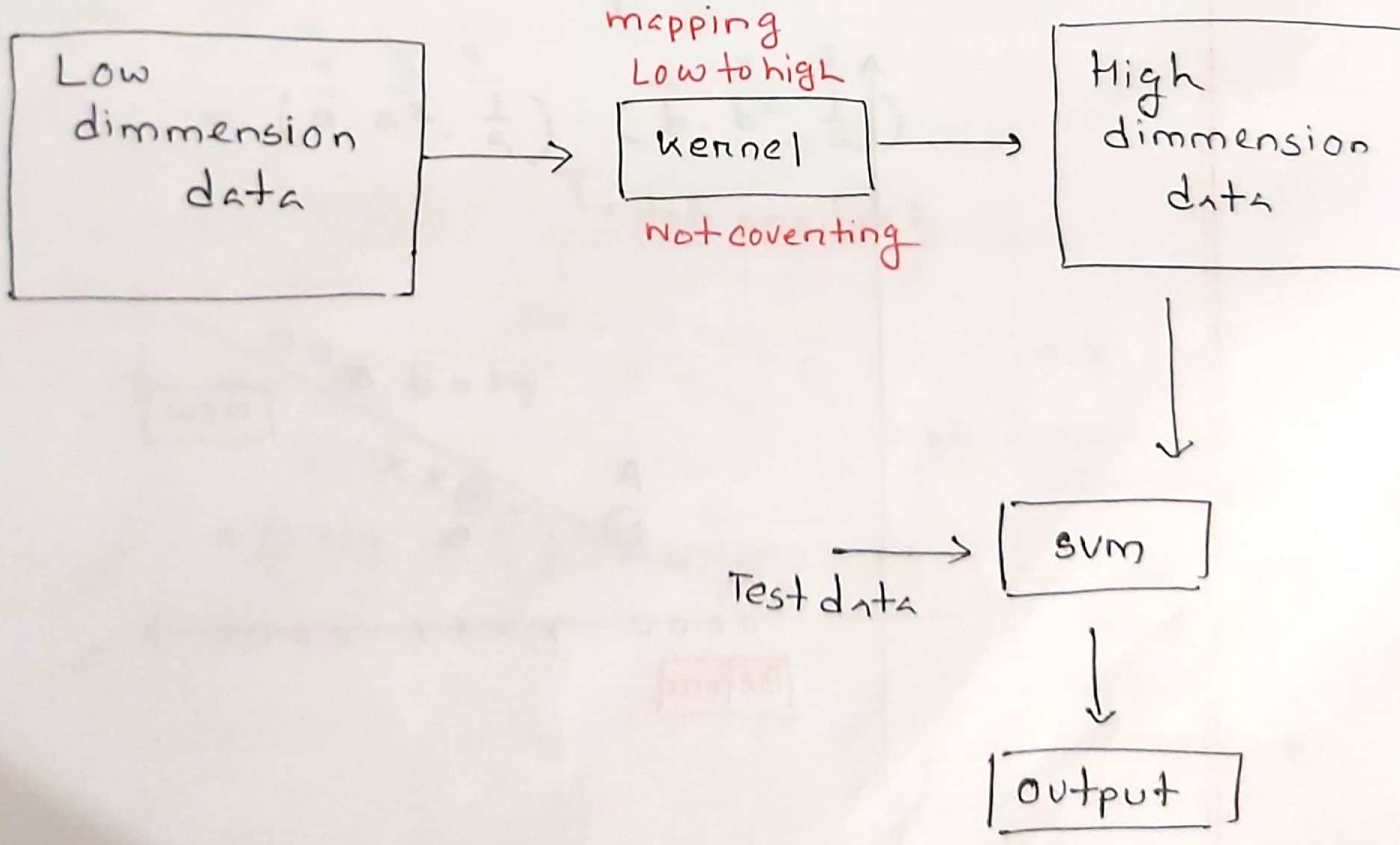
$\therefore$ if multiple hyperplanx $= \frac{2}{\|\vec{w}\|}$ → hypenplan that has maximum value of this will choosen

SWE-17

Kernel Trick

what is kernel trick

→ Kernel are a set of functions used to transform data from lower dimmension to higher dimmession and to manipulate data using dot product at higher dimmension.



SVDE 17

The Radial Kernel

what is RBF

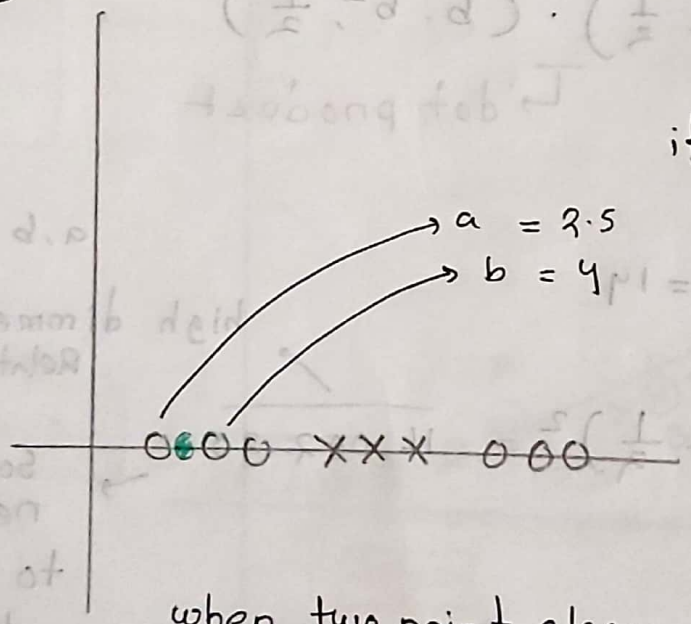
$$RBF = e^{-\gamma(a-b)^2} = e^{-\gamma(a-b)^2}$$

Radial Basis Function

a, b = two different observation in dataset

RBF is kernel trick in svm that transform data into higher dimensional space

γ = influence
= the amount of influence between two points



if $\gamma = 1$
RBF = 0.11

$\gamma = 2$
RBF = 0.01

when two point closer, RBF high

two point far each other, RBF low close to 0

SWF 17

How functions transforms the data into higher dimmension and reduce computation?

$$RBF = e^{-\gamma(a-b)^2}$$

① implicitly maps each data points into higher dimmension space

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

$$RBF \; ; \quad e^{-\frac{1}{2}(a-b)^2}$$

$$= e^{-\frac{1}{2}(a^2-2ab+b^2)}$$

$$= e^{-\frac{1}{2}(a^2+b^2)} \, e^{ab}$$

proof give

$$e^{ab} = 1 + \frac{1}{1!}ab + \frac{1}{2!}(ab)^2 + \frac{1}{3!}(ab)^3 + \cdots + \frac{1}{\infty!}(ab)^{-\infty}$$

$$\boxed{\text{Infinite dimmension}}$$

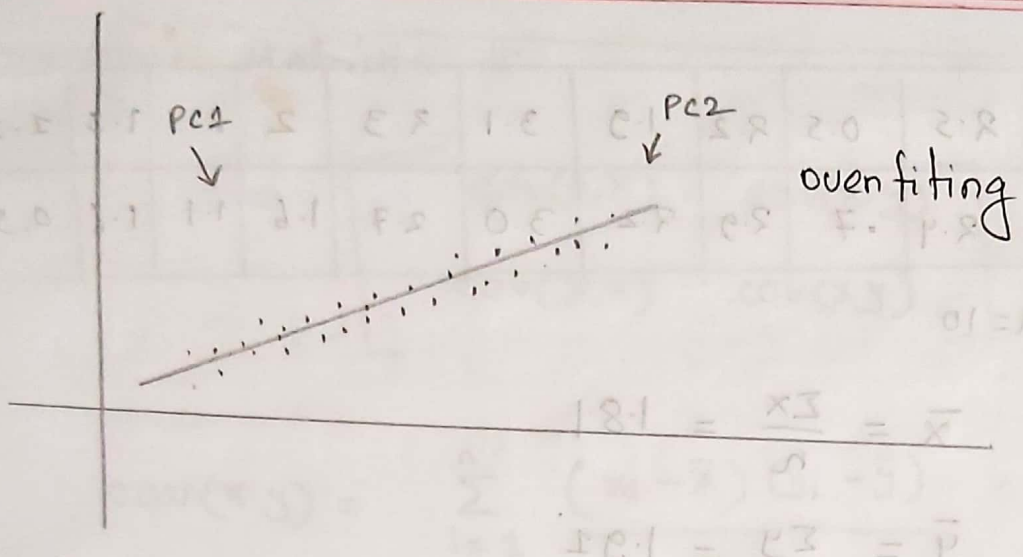
SWE-17

Principle Component Analysis

→ solves overfitting problem

→ It is a statistical technique used for simplifying and summarizing the complexity in high dimensional data while retaining its essential feature.

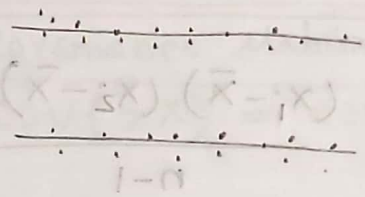
→ helps to indentify and emphasize most important patterns and relationship within a dataset reducing it to a smaller set of variable called principle component



Principle component

PC 1

PC2



SWE-17

what are limitations of PCA ?

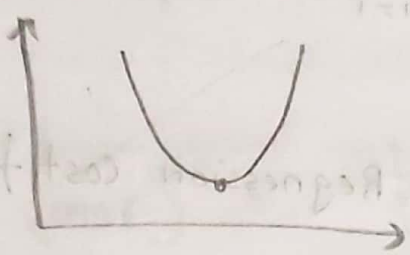
1. Loco interpretability of principle component
2. Trade off between the important information loss and dimmensionality
3. sensitive to scale of feature
4. not efficient for non linear

SF-17

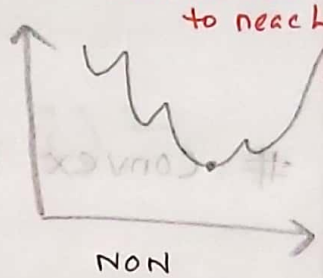
TT-1

(*) Does gradient descent require a convex cost function to coverage?

Yes



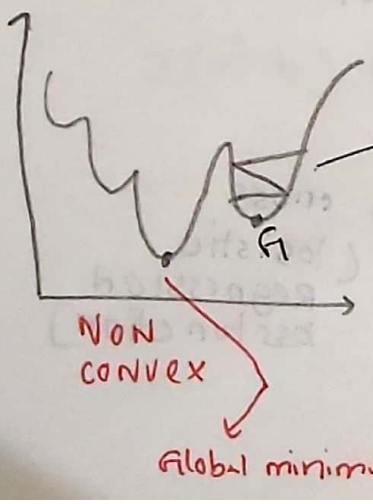
convex



NON CONVEX

why convexity is important gradient descent

1. help to reach global minimum of cost function
2. shape of cost function
3. Learing rate or step size of gradient descent



Here gradient descent stuck at local minimum G and failed to reach global minimum due to NONCONVEX

NONconvex Logistic Regnesseion cost function

$$J(\vec{w},b) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(f_{w,b}(x^{(i)}) - (y^{(i)}) \right)^2$$

convex Logistic Regnession cost function

$$J(\vec{w},b) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{L\left(f_{w,b}(\vec{x}^{(i)}), y^{(i)} \right)}_{\text{Loss function}}$$

SWE-17
CSE

—o—

(*) what is loss function?

loss function is a measurement of finding how well a given machine learing model fits into specific data set

Few example of loss function

1. MSE
2. SVM
3. Binary cross entrop (losistic Regnession loss function)

SWE - 17

(*) can we use mean squared Error for calculating gradient discent of Logisitic Regression to converge to global optima?

mean squared error (MSE) $= \frac{1}{2}(y - \hat{y})^2$

↓

if we create cost of function of Logistic Regression using MSE

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2$$

which is a nonconvex

To converge to global optima we need a convex cost function

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{L\left(f_{w,b}(\vec{x}^{(i)}) - y^{(i)} \right)}_{\text{Loss function}}$$

$$= \begin{cases} -\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) \\ -\log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)})\right) \end{cases}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

SWF-17

F1-score

F1 is harmonic mean of precision and Recall.

$$F_1\text{-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

How f1-score use to select the best model?

F1 score Range (0 — 1) best possible score

Good f1 score indicates good and balance between both precision and Recall

Example :- A model class A 95% instance, B 5% instance model can predict A 95% high accuracy but to B model is very poor.

using F1 we check both FP & FN which helps predict better model

(*) K means Automatically adjust the number of clusters

→ False user need to specify

(*) what is decision boundary?

→ It is conceptual dividing line or surface that seperates different classes or groups in classification problem.

(*) what is model bias?

→ refers to the presence of systematic errors in a model that can cause it to consistently make incorrect prediction.

CS CamScanner

(*) what is confusion matrix ?

→ A confusion matrix is an N * N matrix used for evaluating the performance of a classification model where N is the total number of target class. where the matrix compares the actual target values with those predicted by machine learning model.

(*) confusion matrix (spam on not spam)

Target / Actual	Prediction: SPAM (positive)	Prediction: NOTSPAM (Negative)
SPAM positive	TP 1+1+1	~~FP~~ FN 1+1
Notspam Negative	~~FN~~ FP 1	TN 1+1+1+1

$TP = 3$

$TN = 4$

$FP = 1$

$FN = 2$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

$$= \frac{7}{9}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$= \frac{3}{4}$$

$$\text{sensitivity} = \frac{TP}{TP + FN} = \frac{3}{5}$$

$$\text{specificity} = \frac{TP}{TP + FP} = \frac{3}{4}$$

(*) what is sensitivity?

→ it is a metric that measures the ability of a model to connectly identify positive instance from all actual possible instance in the dataset.

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN}$$

(*) what is feed forward Network?

→ Also known as multilayer perceptron is a type of neural Network where information flows in one direction from input to output layer - without forming cycle/loops

CS CamScanner

(*) Gradient descent formula for polynomial Regression?

$\rightarrow \quad h(x) = a_1 x_1 + a_2 x_2^2 + a_3 x_2$

$$J(a_1, a_2, a_3) = \frac{1}{2m} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)} \right)^2$$

Gradient descent

$$a_1 = a_1 - \alpha \frac{d}{da_1} J(a_1, a_2, a_3)$$

$$a_2 = a_3 - \alpha \frac{d}{da_2} J(a_1, a_2, a_3)$$

$$a_3 = a_3 - \alpha \frac{d}{da_3} J(a_1, a_2, a_3)$$

CS CamScanner

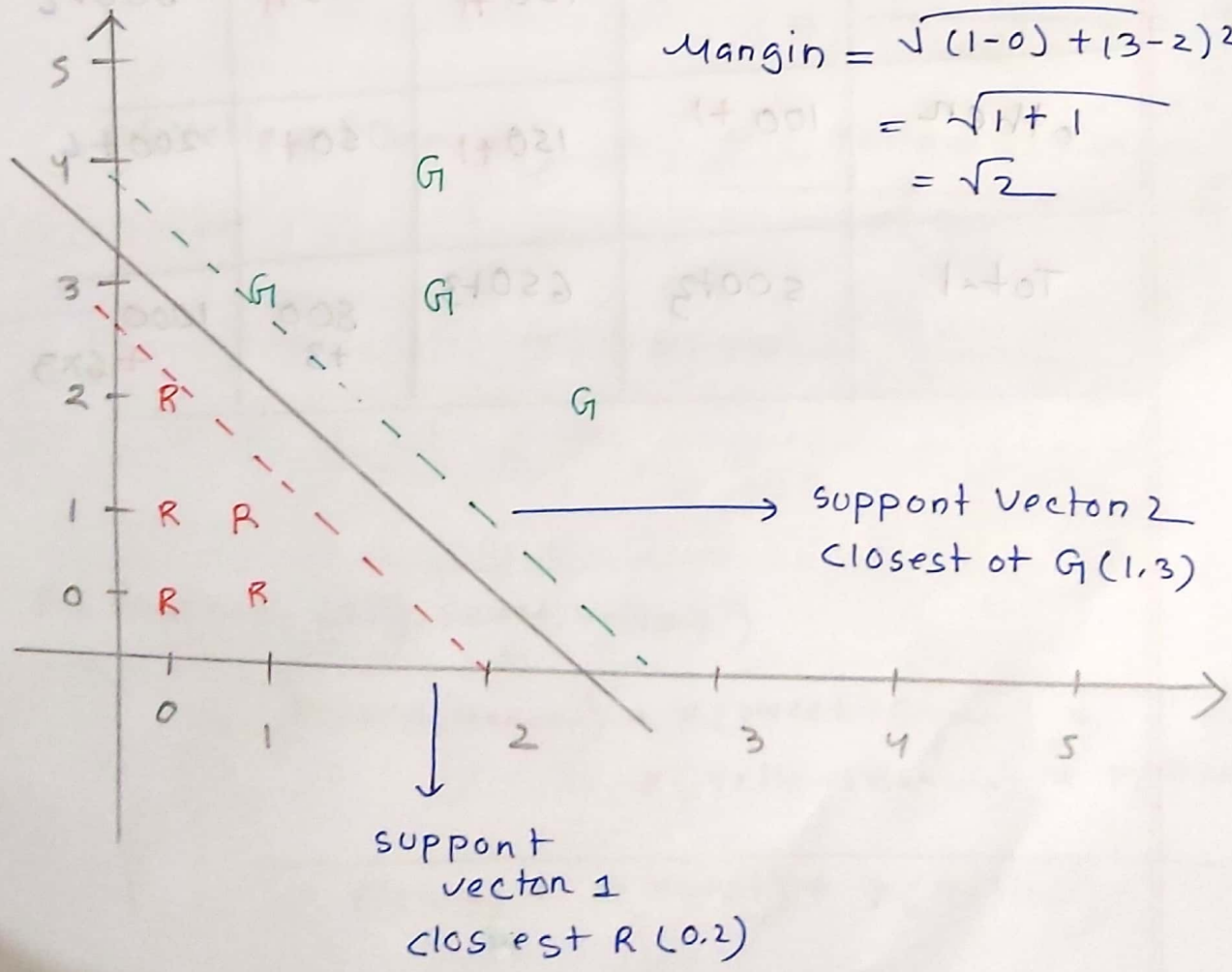
$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

what is MSE → measure the average squared difference between predicted($\hat{y}_i$) and true value (y_i)

(b)

(0, 0, Red) (0, 1) Red (0, 2, Red) (1, 0 Red) (1, 1 Red)
(1, 3, Green) (2, 3 Green) (2, 4 Green) (3, 2 Green)

$$\text{Margin} = \sqrt{(1-0) + (3-2)^2}$$
$$= \sqrt{1+1}$$
$$= \sqrt{2}$$



(c) the zero frequency issue

adding all values +1

Type	Long	Sweet	Yellow	Total
Bannana	400+1	350+1	450+1	500+6
onange	0 +1	150+1	300+1	300+6
other	100+1	150+1	50+1	200+6
Total	500+3	650+3	800 +3	1000 +6x3

$$P(\text{long} \mid \text{Banana}) = \frac{P(\text{Long} \cap \text{Banana})}{P(\text{Banana})}$$

$$= \frac{401}{506} = .79$$

$$P(\text{sweet} \mid \text{Banana}) = \frac{P(\text{sweet} \cap \text{Banana})}{P(\text{Banana})}$$

$$= \frac{351}{506} = .69$$

$$P(\text{Yellow} \mid \text{Banana}) = \frac{P(\text{Yellow} \cap \text{Banana})}{P(\text{Banana})}$$

$$= \frac{451}{506}$$

$$= .89$$

$$P(\text{Banana} \mid \text{Long, sweet, Yellow})$$

$$= \frac{P(\text{Long} \mid \text{Banana}) * P(\text{sweet} \mid \text{Banana}) * P(\text{Yellow} \mid \text{Banana}) * P(\text{Banana})}{P(\text{Long}) * P(\text{sweet}) * P(\text{Yellow})}$$

CamScanner

P(Banana | Long, sweet, yellow)

$$= \frac{.79 * .65 * .89 * (506/1018)}{\frac{503}{1018} * \frac{653}{1018} * \frac{803}{1018}}$$

$$= .96$$

P (orange | Long, sweet, yellow)

$$= \frac{P(\text{Long} \mid \text{orange})\ P(\text{sweet} \mid \text{orange})\ P(\text{yellow} \mid \text{orange})\ P(\text{orange})}{P(\text{Long})\ P(\text{yellow})\ P(\text{sweet})}$$

$$= \frac{\frac{1}{306} * \frac{151}{306} \times \frac{301}{306} \times \frac{306}{1018}}{\frac{503}{1018} * \frac{803}{1018} * \frac{653}{1018}}$$

$$= 1.90 \times 10^{-3}$$

$P(\text{other} \mid \text{long, sweet, yellow})$

$$= \frac{P(\text{long} \mid \text{other})\, P(\text{sweet} \mid \text{other})\, P(\text{yellow} \mid \text{other})\, P(\text{other})}{P(\text{long})\, P(\text{sweet})\, P(\text{yellow})}$$

$$= \frac{\frac{101}{206} \times \frac{151}{206} \times \frac{51}{206} \times \frac{206}{1018}}{\frac{503}{1018} \times \frac{653}{1018} \times \frac{803}{1018}}$$

$= 0.072$

$\therefore$ $P(\text{Banana} \mid \text{Long, sweet, yellow})$

has the highest probability

$\therefore$ The fruit is banana

(d) $A : B = 1 : 20$

$$P(A) = \frac{1}{21} \qquad P(B) = \frac{20}{21}$$

Given

$$P(UN/A) = 12\% = 0.12$$

$$P(UN/B) = 30\% = 0.3$$

a person got chance in UN is from A region

$$P(A/UN) = \frac{P(UN/A) * P(A)}{P(UN)}$$

$$= \frac{P(UN/A) * P(A)}{P(UN/A)P(A) + P(UN/B)P(B)}$$

$$= \frac{0.12 \times \frac{1}{21}}{0.12 \times \frac{1}{21} + 0.3 \times \frac{20}{21}}$$

$$= 0.02$$

(*) what is Minmax scaling?

→ is a techique used in data preponcessing to scale the values of numenical featune to a specific nange.

original value of feature

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

(*) what is activation function?

→ is a mathmatical openation applied to the output of each neuron/node in given layen which detenmines the output of a node given a set of weight + bias

(*) when the featun e space is lange, overfitting is mone likely

True

what is generative leaning algorithm?

→ refers to an algorithm / model that is designed to generate new samples/ data instance that resssemble a ~~giving~~ given training dataset (goal is to learn underlying pattern / structures in training data to create new similar samples)

CamScanner

66

{3, 4, 6, 12, 18, 22, 23, 32, 33, 35}

A B C D E F G H I J

K = 2

9 — K_1

22 — K_2

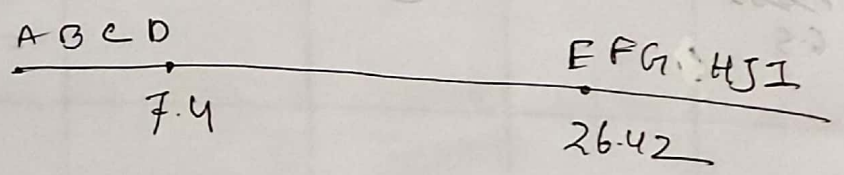
	K_1	K_2	K_1	K_2	K_1 centenoid	K_2 centroid
A = 3	|9−3| = 6 ✓	|22−3| = 19	{A}	{ }	$\frac{9+3}{2} = 6$	22
B = 4	|6−4| = 2	|22−2| = 20	{A, B}	{ }	$\frac{9+3+4}{3} = 5.34$	22
C = 6	|5.34−6| = .66	|22−6| = 16	{A, B, C}	{ }	$\frac{9+3+4+6}{4} = 5.5$	22
D = 12	|5.5−12| = 6.5	|22−12| = 10	{A, B, C, D}	{ }	$\frac{9+3+4+6+12}{5} = 7.4$	22
E = 18	|7.4−18| = 10.6	|22−18| = 4	{A, B, C, D}	{E}	7.4	$\frac{22+18}{2} = 20$

CamScanner

F=22	|7.4-22| =14.6	|20-22| 2	{A, B, C, D}	{E, F}	7.4	$\frac{18+20+22}{3}$ = 20.67
G=23	|7.4-23| =15.6	|20.67-23| = 2.67	{A, B, C, D}	{E, F, G}	7.4	$\frac{18+22+22+23}{4}$ = 21.25
H=32	|7.4-32|	|32-21.25| = 10.75	A, B, C D	E, F, G, H	7.4	$\frac{18+22+22+23+32}{5}$ = 23.4
I=33	|7.4-33|	⇒ |23.4-32| =8.6	A, B, C D	E, F, G, H, I	7.4	25
J=35	|7.4-35|	|25-35| =16	A, B, C, D	E, F, G, H, I, J	7.4	26.42

$\therefore K_1 = \{A, B, C, D\}$

$\therefore K_2 = \{\cancel{A}, E, F, G, H, I, J\}$



(e) Given

$$P(\text{alarm}) = \frac{40}{100}$$

$$P(\text{bucket sent}) = \frac{30}{100}$$

$$P(\text{alarm} \cap \text{bucket sent}) = \frac{20}{100}$$

$$\therefore \quad P(\text{bucket} \mid \text{alarm}) = \frac{P(\text{alarm} \cap \text{bucket})}{P(\text{alarm})}$$

$$= \frac{\frac{20}{100}}{\frac{40}{100}}$$

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

(*) what is cross validation? why is it used?

CV is a statistical tecnique used in machine leaning to assess the performance and ~~generi~~ generalization ability of a predictive model by providing a unbiased esitamate of how well a model will perform on independent dataset

purpose :-

1. detect overfitting
2. Remove Bias
3. Avoid creating wrong Prediction
4. partition dataset into subsets

CamScanner

why do we need dimmension reduction ?

→ It is a process of reducing the number of random variables by obtaining a set of principle variable

<u>purpose</u>

help

① Reduce the timp & space Required

② Remove Irrelevant feature from data

③ increase accuracy of model

④ easier to visualize data

⑤ Improves the interpretations of parameter

<u>what are drawback ?</u>

① may lost some data which impacts howwell algorithm works

CS CamScanner

② a lot of processing power needed

③ may lead overfitting

④ interpretting transformed chanacteristic might be challenging

⑤ Independent Variables beome harden to comprehened

⑥ assumes the data is linear

⑦ Sensitive to outliners

What is propagation?

Forward propagation

refers to the process of passing the input data through neural network to generate predictions or outputs

prediction based on current weight + bias

no adjustment

Backward propagation

refers to the process of updating the model's parameters during training phase by minimizing error/loss

adjusting model parameters to make better prediction and improves ~~dem~~ performance

CamScanner