No Relative grading

[40] → End
[10] → Mid

10 → Assignment ⇒ Poster (5th April)
40 → Project (A2)

5 | 10 | 25 → final entire report

Proposal presentation → Submit in report

Midterm → Base line experiment

[5,10] ⇒ Bonus for publication

midterm before Saturday
endterm before Saturday

Timelines

Project → Theory Project
→ Application → Data set curation

Pick a problem not solved by me before
(New problem)

Evolution, Current & Future

AI → Somebody as good as.

↓

What ever humans can do

AI → Rule-Based

↓ Eg ATM

↓ Data-Based (Machine learning)

↓ Next target

AGI → Artificial General Intelligence

↓ Humanoid

↓ Next target

SI → Super intelligence

Table

Data → ⇒ Easy
↳ Systematic
↳ Fast processing
O(1)

Data → ML → Table

& Rules

Use ML where data is not defined

4/1/2023

Machine Learning

Training data

Testing data

Train

Notes

(Memorisation)

Exam

Exactly same as (Not learning)

No Match (No learning)

(good)

(learning)

Question based on Notes but not same

(generalisation)

Not to correct model based on test set

Training data

Trained → Validation → Testing

These data should be in similar space

Machine L based on types of Data it has.

* Supervised (X, y) label

bright	sun	cloud	temp

data

* Un supervised
(Not labelled) (X)

* (x, y) you labelled
Semi Supervised

* Self Supervised
(x)

↓
Chat GPT

I love Indian food
I ___ India food
I love ___ food
I love Indian ___

* Reinforcement learning (x, y)

Bunch of steps, I could have done, but (game of chess)

Overall I have to win the game.

(I, y)

$x_1 \; - \; x_n$ $\rightarrow y$

Exam Testing

Regression

↓

Continous Value

eg Temp 22.67°C

Classification

↓

grouped

↓

Classis

eg High temp or low temp

Into multiple classes

multi labelled

Binary classification

Dog, cat, no animal, human

Page No
Date : /

Most Imp

INDUTIVE BIAS

Inherently Bias regardless of data

→ On various models

What is

Structured data

↓

which is in format on which we apply statistical principle

→ eg Table

→ tidy → ① Every column has to have one information

② Every row has to have one observation

③ Cell has to have one information

Support Vector Machine — eg of supervised learning

label

Email → Spam (1) ⇒ supervised, classified
→ Not-Spam (0)

→ SVM prefers 1 & -1

Convert non numeric into numeric value

ASCII ⇒ letter into numeric

but we need word convert to numeric

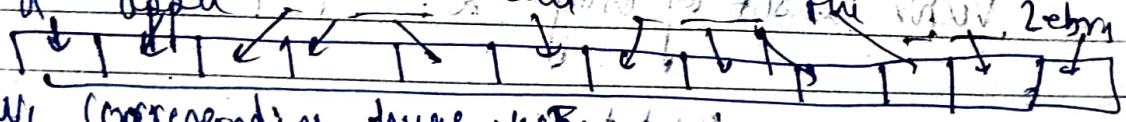
↓

[Bag of words]

↓

It takes entire english dictionary
(May be only popular words they found)

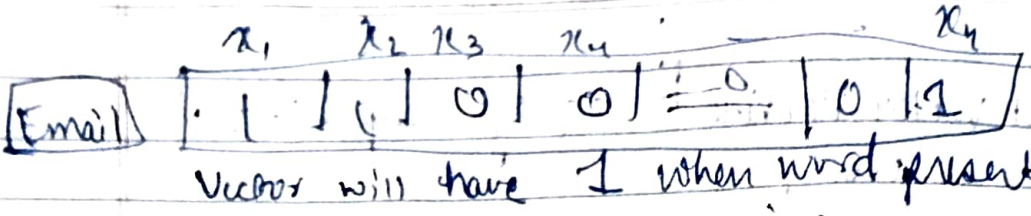
Eng Dict



have (corresponding huge vectors)

Bigger dataset vector problem ⇒ computational complex

Also word ⇒ not able to accomodate all words.



Vector will have 1 when word present in it.

⇓

Limitation

↳ Like, frequency of words will be lost

↳ Order of words doesn't matter, which is disadvantage

* One hot encoding *

Email converted into vector, now what to do?

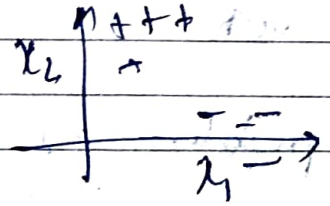
⇓

Can be represent in space

Inductive Bias

Algo expect a very specific type of pattern

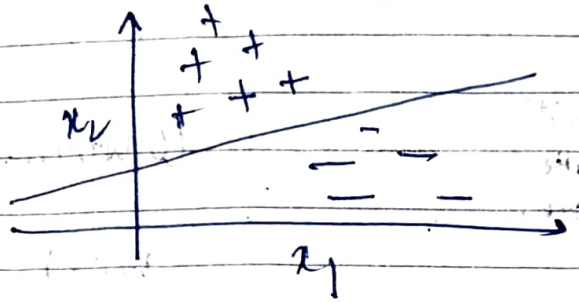
Vector consist of only 2. x_1 x_2



(+) & (-) → Corresponds email

↓ spam ↓ not spam

Goal is find line that cuts space into half



* Inductive bias is in SVM

↓

they have bias that either of thing it has to be left or right side

y ⇒ i am predicting

↓ y = f(x)

= sign(x)

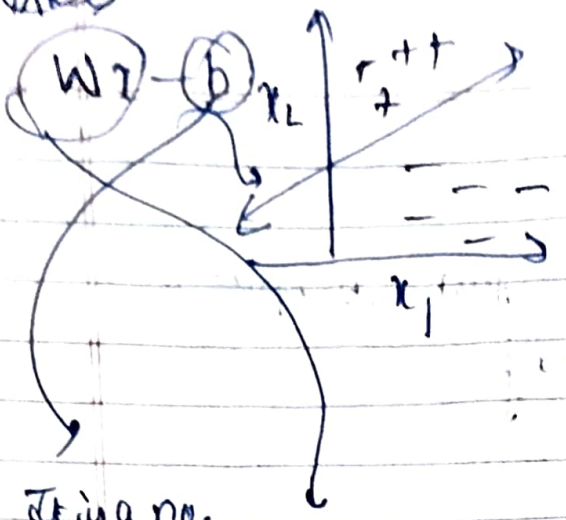
↓ function → - Integer ⇒ -1

→ + Int ⇒ +1

y = sign(W X − b)

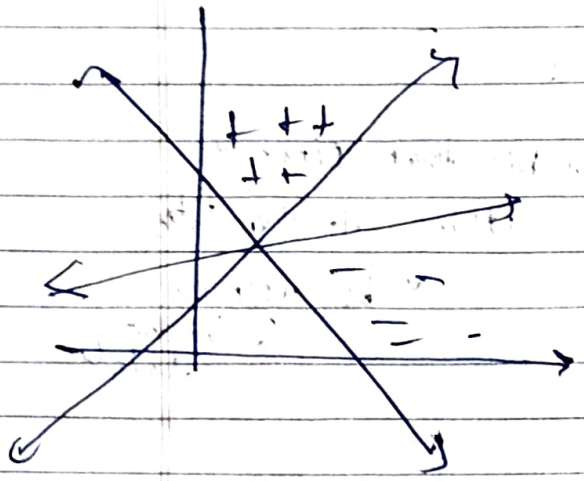
↓ ↓ ↓

Vector Vector bias



It is a no.

It will make a no.

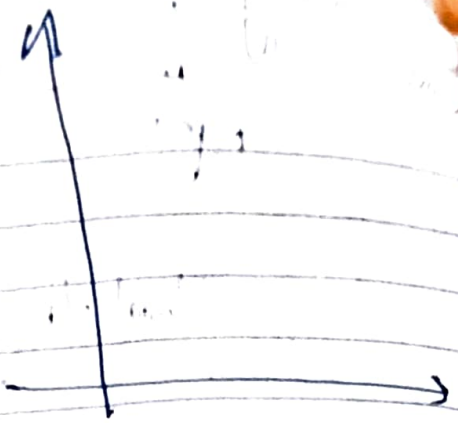


multiple such lines /many

Try to find best possible combination of w^* & b^*

$$y = \text{Sign}(w^* x - b^*)$$

is email



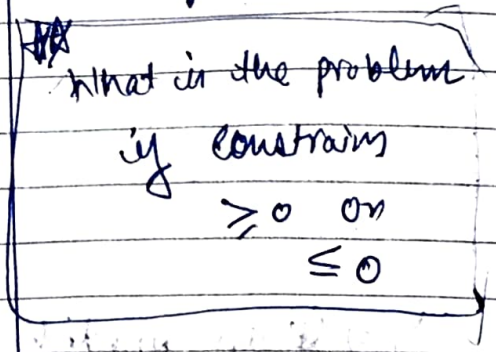
Constrains

$wx - b \geq 1, \quad y = +1$

$wx - b \leq -1, \quad y = -1$

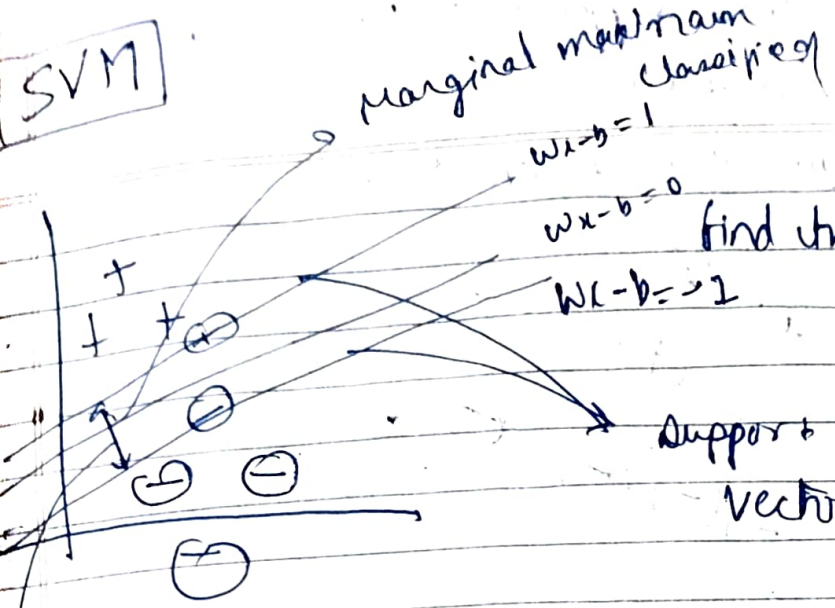
y sign = -1

What is the problem if constrains ≥ 0 or ≤ 0



Goal is to find best w & b

SVM



find two points which are closest to each other

Bad Centroids

By eqn there could be chance that there does not exist a line but in reality there exist a line

We take the closest points so nothing can ignore

- distance between the two lines = $\frac{2}{\|W\|}$, what is this?
- Maximize this so that accurately classify here.

⇓

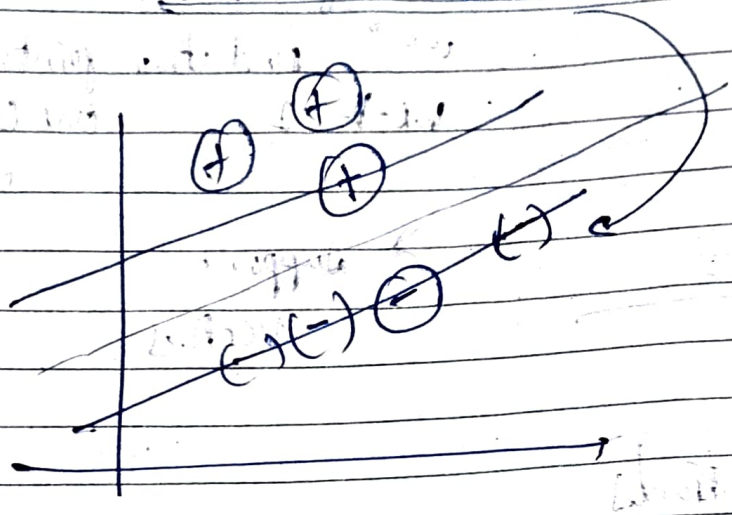
Minimize $\|W\|$

s.t. $Wx-b \geq 1$, $y=1$

$Wx-b \leq -1$, $y=-1$

At least two support vector has to be there

More than 2 can exists



Inductive Bias of the algorithm

↳ lines

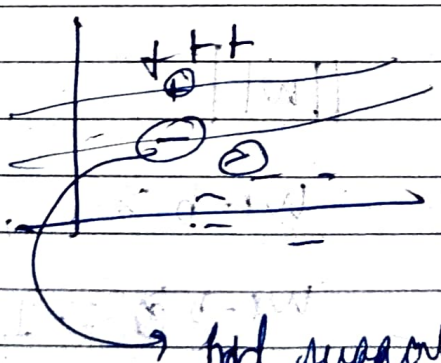
↳ ~~curves~~

Advantage of this algo ⇒ little complex

↳ w/t & b only depends upon the support vector.

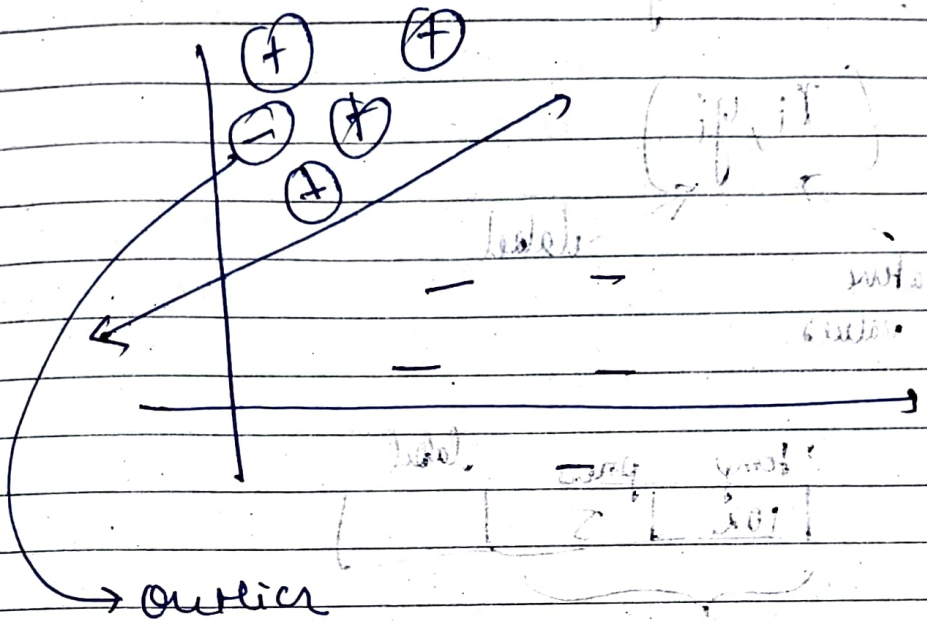
Model based learning

Bad support vectors are there then also

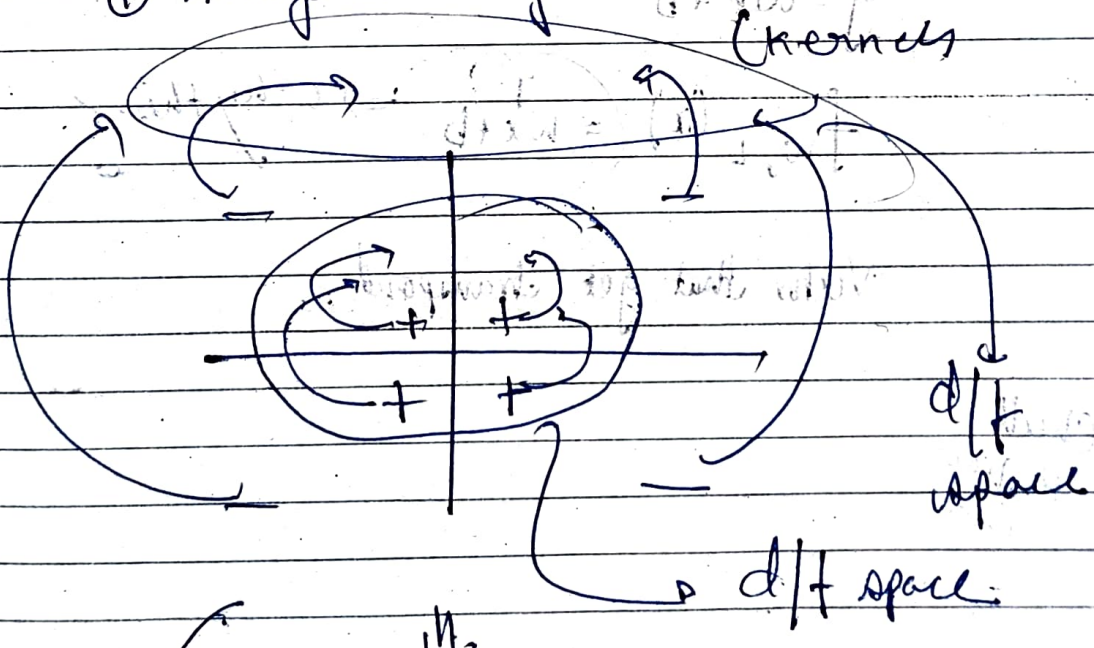


Page No.
Date : / /20

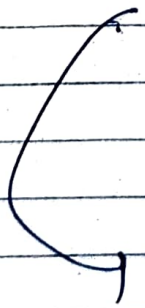
How to cure Outliers



① Transformation of non linear to linear (kernels)



Also this can't separate them



so square things up

(2) Penalty / error bar)

Page No.
Date: / /20

9/1/23

Linear Regression

(x_i, y_i)

x_i: feature values

y_i: label

temp	press	label
102	5	

x_1

$y = wx + b$

$f_{w,b}(x) = w^T x + b$ ⟹ why this?

Vector that get transpond.

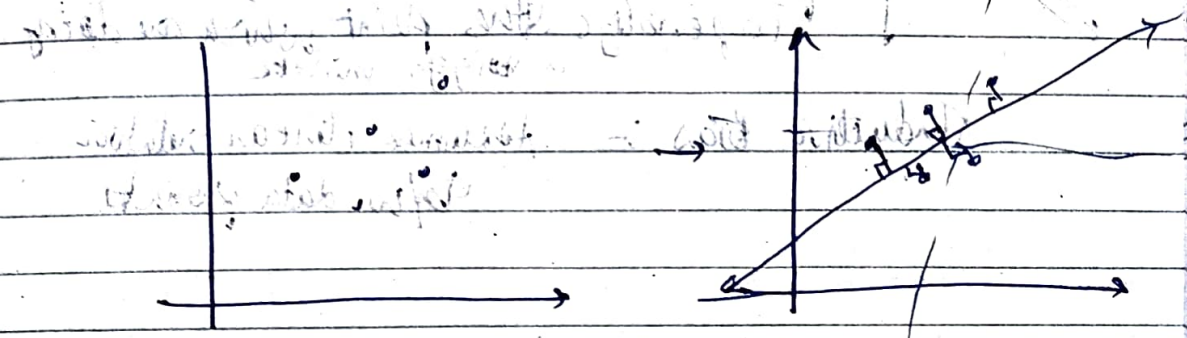
Check Richer news

Linear Regression

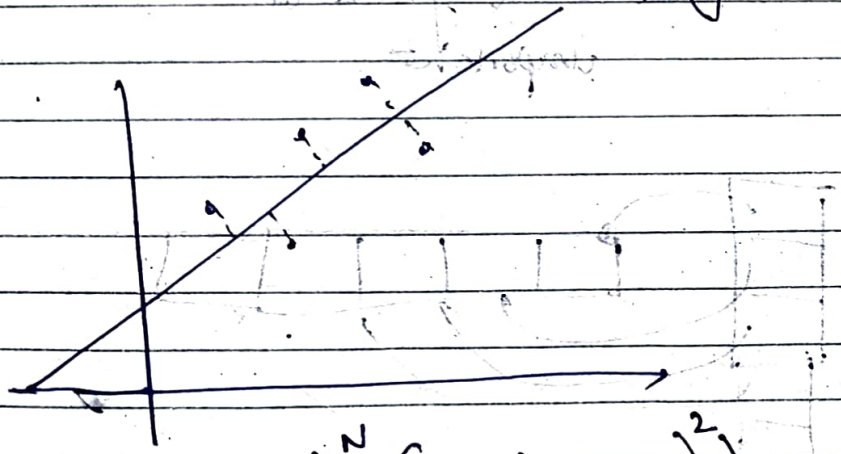
Page No.
Date: / /20

Q: used for Regression values

1 → find w & b st. most of the points fall on the line



find least square fit



$$\min\left(\sum_{i=1}^{N}\left(f_{w,b}(x_i) - y_i\right)^2\right)$$

cost function

$$\min\left(\frac{1}{N}\sum_{i=1}^{N}\left(f_{w,b}(x_i) - y_i\right)^2\right)$$

squared error

Empirical loss / risk

loss function

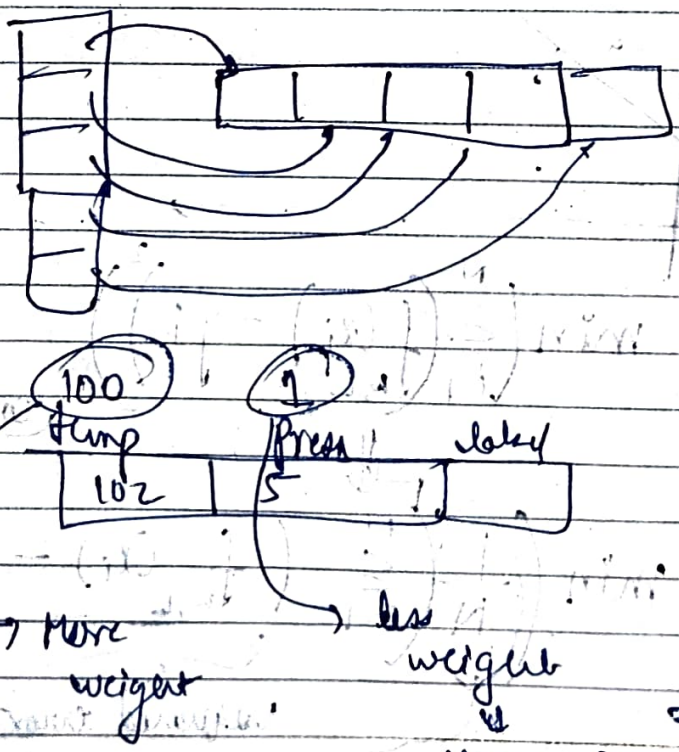
Analytical solution ⇒ the solved form of least square error.

⇒ gradient descent

Why square, why not mod?
To penalize the point which are doing a bigger mistake

Inductive bias :- Assume linear relation b/w data points

Why to chose linear regression?
⇓
tells which feature is important



What is the need of other algorithm

① less error → Empirical
② → Theoretical

Theoretical → Bounds on the error ? → Meaning ?

Empirical → to justify empirically → Standard dataset

How to prove that my algo is better?

⇓

Perform ML on standard dataset

How to make your own dataset standard?

Feature engineering

① Dataset should not contain features which are not useful

② Computer only understand Numbers

⇓

Convert to Number

⇓

Assign numbers

→ problem with this ⇒ ?

Continuous flow of information

• Computer thinks 1 is close to 2 & 3 is close to 2 but 1 to 3 is far which is wrong

One hot encoding

Page No.
Date: / / /20

Instead of giving no. represent them in vector of 3

R [0 0 1]
G [0 1 0]
B [1 0 0]

→ Disadvantage is the data become bigger

(3) Binning

[22.5, 20.1, 19.2] → normal
[32.2, 33.4, 35.5] → hot

⇒ Convert to one-hot encoding

(4) Normalisation [0, 1]

⇓

making max to the (1)

$$\text{Normalization} = \frac{x_i - \min}{\max - \min}$$

Page No.
Date: / /20

(5) Standardization

$\frac{x - (\mu)}{\sigma}$ → mean

⇒ Z-score Normalization

↳ deviation

if ~~[illegible]~~

* Why normalization is done? → To bold down them in same scale

Why standardization?

↳ Question it important

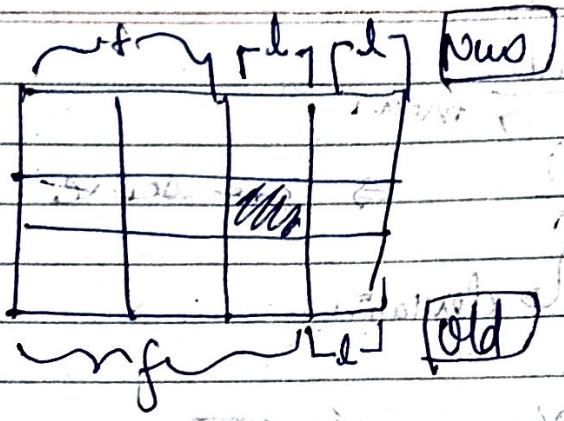
(6) Data imputation

There can be missing values

If you miss some value then what to do?

↳ fix it with average of this

Or just predict the missing data, put it as label



10/1/23

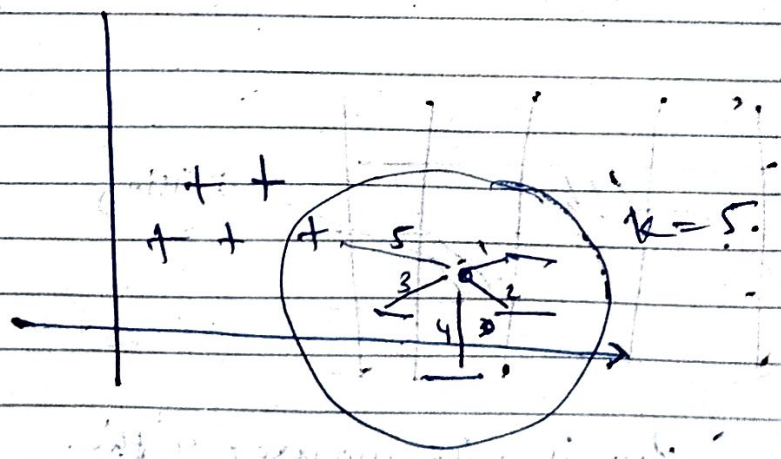
K-NN

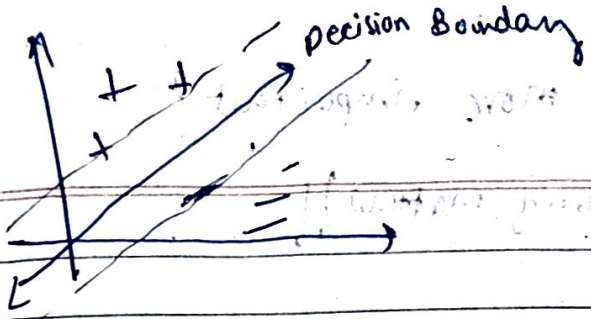
↳ K = Nearest Neighbour

Homophilli & Heterophilli

I Bihar the same way time they neighbour dekhna

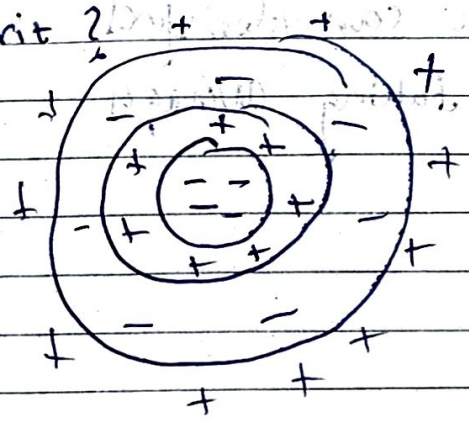
* Find K NN
* take Majority
* label with majority





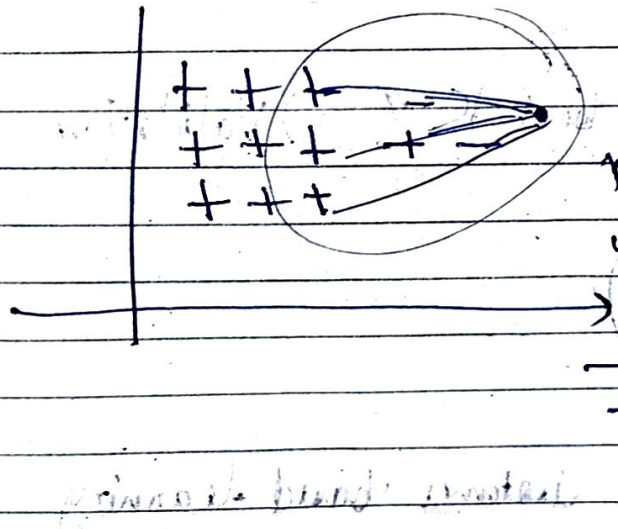
Page No.
Date: / /20

Can we have decision boundary which is discrit ?



What if K is larger

When larger K will help ? ⇒ outliers

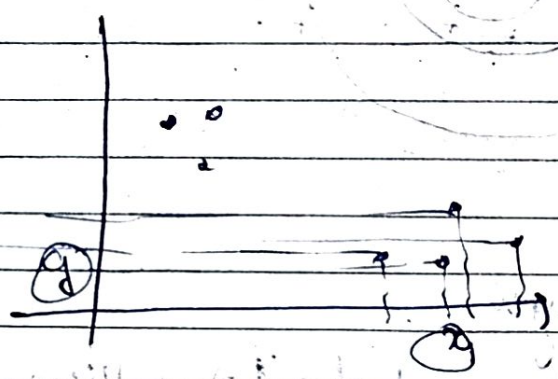


KNN is powerful because of non-linear boundary

Where small error is more important?

Minimizing matters!!

Regression we can also find by taking averages



$$y = \frac{y_1 + y_2 + y_3}{3}$$

What is the best K? ⇒ Validation

$$f_K(x)$$

model v/s instance based learning

Inductive Bias → Regardless of data we only consider k nearest neighbour

Page No.
Date: / / 20

Weighted K nearest neighbours

Radial Nearest neighbour

Problem with KNN

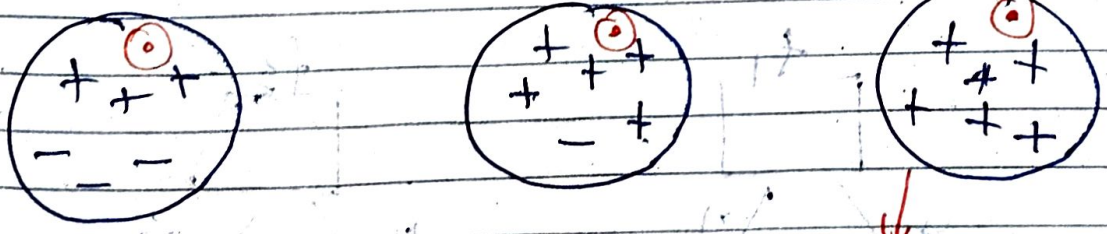
Data structure which can store close points together

To know the closest distance you have to go through every point.

Is KNN explainable ??

Exam Loss function of KNN ??

Decision Tree → Nonlinear model



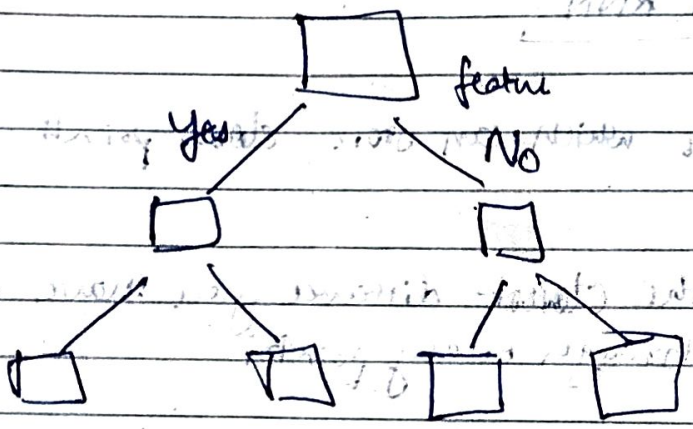
Page No.
Date: / /20

What features to look first??

↓

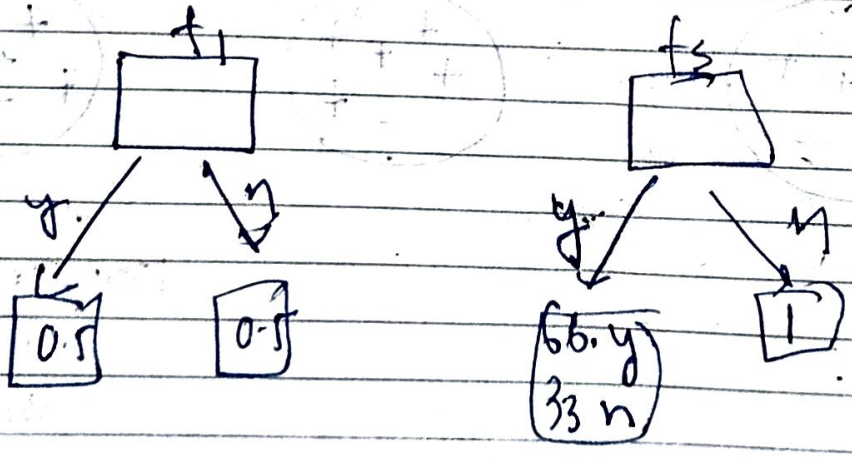
1st step to do

It is like dialogue



10

f_1	f_2	f_3
y	y	y
n	y	y
y	y	n
n	n	n



[Entropy] [Gini Impurity]

Page No.
Date: / /20

Advantage

(1) We don't have to ask all the features

Less question → good
Very high question → Memorize
Normal no. of question → learning

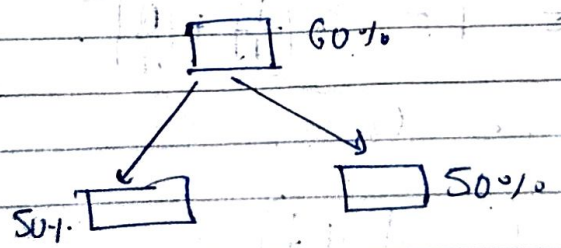
Inductive Bias of Decision tree

- Answer a question by asking alot of question.

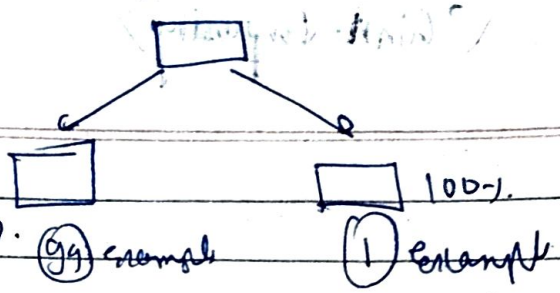
* Causality → for Bio people

* hyperparameters ⇒ check notes

15/1/24



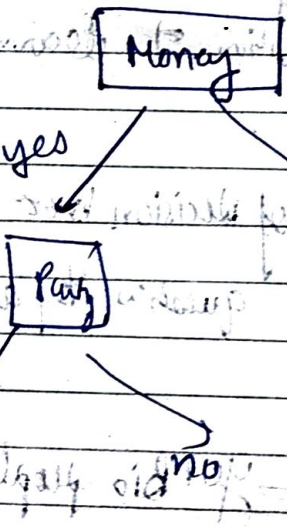
Accuracy should not reduce.



Gini Impurity

$$G(\ ) = 1 - \sum (P_i)^2 \rightarrow$$

0
||
Very
Pure
separation



	yes	no
Eg 1	100	0
Eg 2	50	50

$$G(1) = 1 - \sum \left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2$$

$$= 0$$

$$G(2) = 1 - \left(\frac{1}{2} + \frac{1}{2}\right)$$

$$= 1/2$$

No. of Branches depend on feature value

Page No a
Date / /20

Gi(Money) = Weight avg → Gi(Money = yes) [b], Gi(Money = no), Gi(Money = maybe) [c]

$= \frac{100}{150} * G_i(\text{Money} = \text{high})$

$+ \frac{50}{150} * G_i(\text{Money} = \text{low})$

$1 - \Sigma(p^2_{yes} + p^2_{no})$

↳ yes for party

⇒ If Money is high what possibility that it will go for party or not.

Currently we have 150 rows with Money & party Column

Gini impurity of your feature

$G_i(f_1)$
$G_i(f_2)$
⋮
$G_i(f_n)$

} weighted average

Which to choose → which has less gini impurity

PT	OI	TV	Return
+ve	low	high	up
-ve	high	low	Down
+ve	low	high	up
+ve	high	high	up
-ve	low	high	Down
+ve	low	low	Down
-ve	high	high	Down
+ve	low	low	Down
+ve	high	high	up
-ve	low	high	Down

$$G(f_{value_i}) = 1 - \sum p_j^2$$

p_j → value of label

i → value of feature

$$G(f_{value}) = \sum w_i \, G(f_{value_i})$$

$G(PT) = ?$

$G(OI) = ?$

$G(TV) = ?$

$$G(PT) = w_i \, G(f_{value_i})$$

$$= w_i \, G(PT_+) + w_i \, G(PT_-)$$

$$= \frac{No.\ of\ +ve}{10} G(PT_+) + \frac{No.\ of\ -ve}{10} G(PT_-)$$

$$= \frac{6}{10} G(PT_+) + \frac{4}{10} G(PT_-)$$

$$G(PT_+) = 1 - (P_{up}^2 + P_{down}^2)$$

$$= 1 - \left(\frac{16}{36} + \frac{4}{36}\right)$$

$$= 1 - \left(\frac{20}{36}\right) = 1 - \frac{5}{9} = \frac{4}{9}$$

$$G(PT_-) = 1 - (0^2 + 1^2)$$

$$= 0$$

$$= \frac{6}{10} \times \frac{4}{9} + \frac{4}{10} \times 0$$

$$G(PT) = \frac{24}{90} = \frac{8}{30} = \frac{4}{15}$$

$$G(OI) = \Sigma w_i \, G(OI_i)$$

$$G(OI_i) = 1 - \Sigma P_j^2$$

Calculate

~~G(0.5)~~

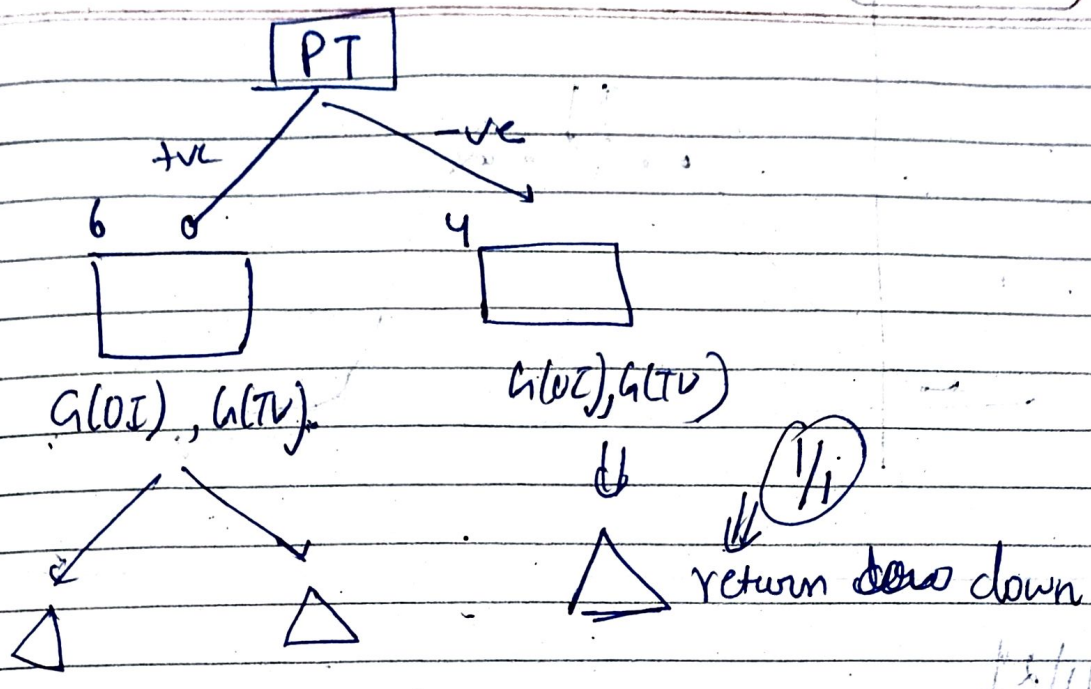
$G(0.5) =$

$G(TV) = \frac{24}{70}$

$G(0.5) = \frac{4}{10} \times \left(\frac{7}{8} + + \frac{6}{10} \right.$

$1 - \Sigma \frac{1}{8} = \frac{7}{8}$

Choose the lowest one



Only split when purity is 0

Right matches hyperparameter stop splitting it

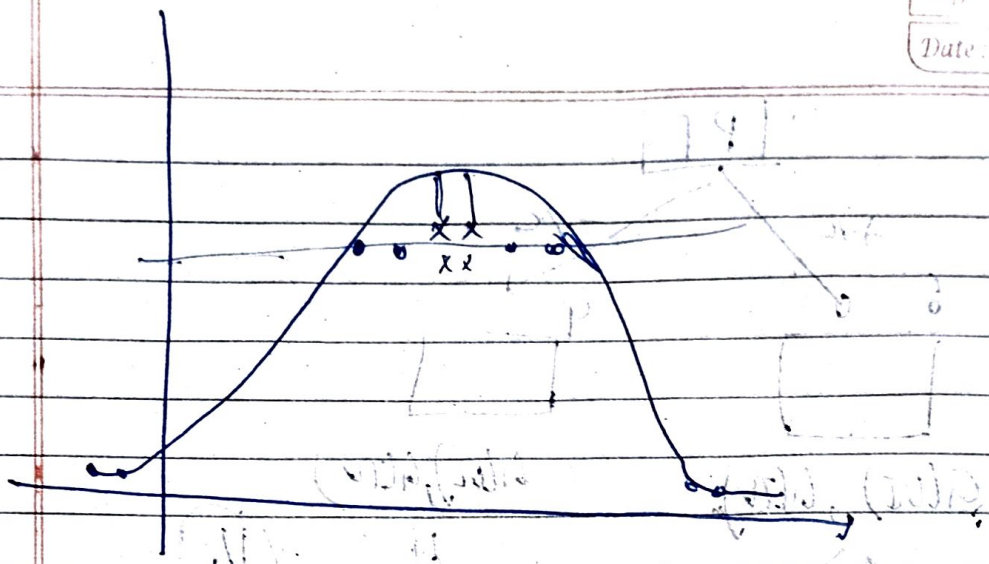
Inductive V/s transductive learning

↓

to look at which space

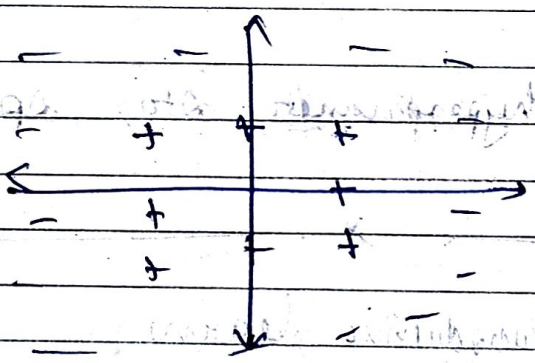
→ to test on specific one

No why look at the other features

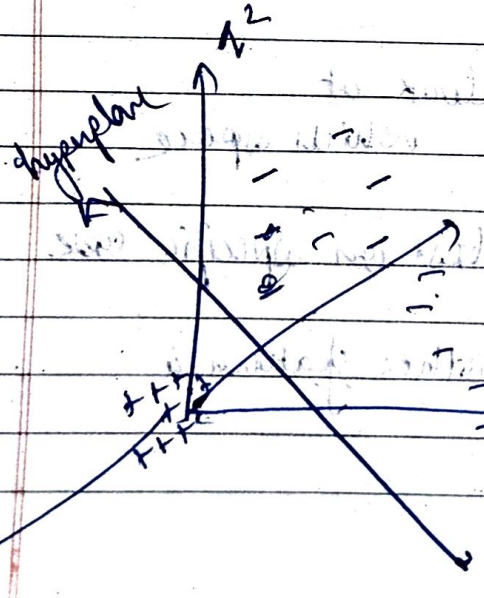


16/1/24

* Non linearly separable data



(p, q)



$(p^2, \sqrt{2}pq, q^2)$

21/1

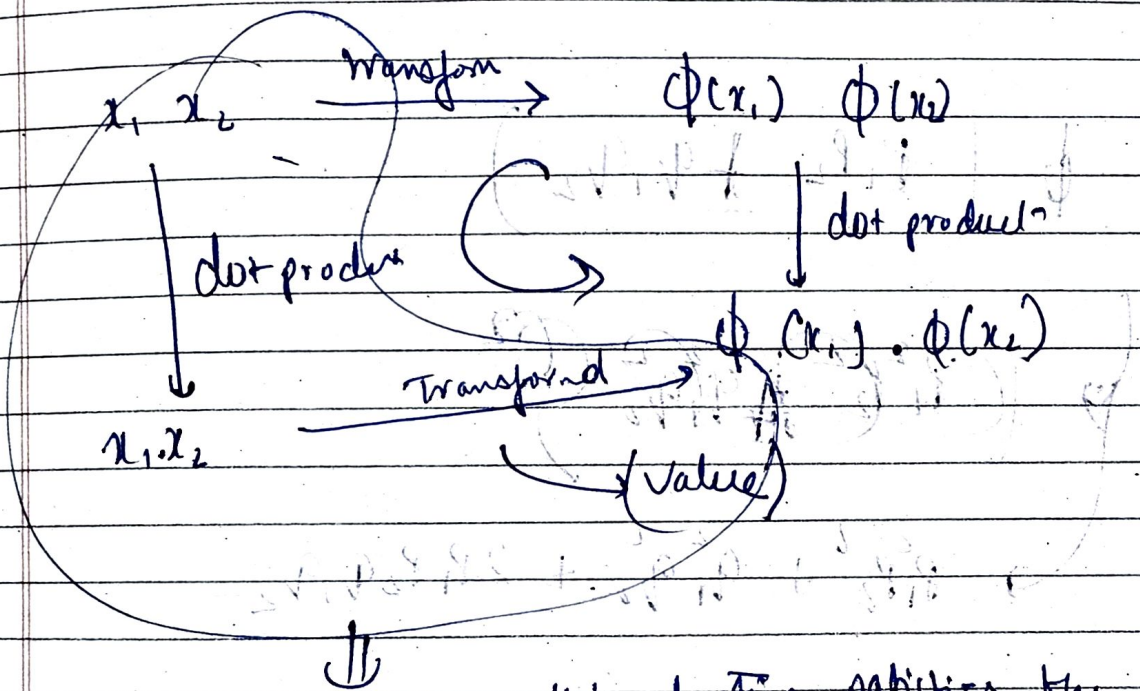
If something is not linear in some dimension then we can make it linear in higher dimension.

Ques: If something is non linear can it become linear in higher dimension. ?

Kernel trick

$x_1 \cdot x_2$ $\qquad$ $\phi(x_1) \cdot \phi(x_2)$

↳ Transform



In case this function satisfies the dot product

$$\phi(x_1 \cdot x_2) = \phi(x_1) \cdot \phi(x_2)$$

This is named

$p^2, \sqrt{2}pq, q^2$

$x_1 = (p_1, q_1) \xrightarrow[\text{transformer.}]{\text{quadratic}} (p_1^2, \sqrt{2}p_1q_1, q_1^2)$

$x_2 = (p_2, q_2) \longrightarrow (p_2^2, \sqrt{2}p_2q_2, q_2^2)$

$\left(p_1^2 p_2^2 + 2p_1q_1p_2q_2 + q_1^2q_2^2 \right)$

$\phi\left(p_1p_2 \pm q_1q_2 \right)$

$\left(p_1^2p_2^2 + q_1^2q_2^2 \right)^2$

$\rightarrow p_1^2p_2^2 + q_1^2q_2^2 + 2p_1p_2q_1q_2$

Page No.
Date: / /20

$x_1 (p_1, q_1) \longrightarrow p_1^2, \sqrt{2} p_1 q_1, q_1^2$

$x_2 (p_2, q_2) \longrightarrow p_2^2, \sqrt{2} p_2 q_2, q_2^2$

Ankit Richard PL

Why do we need other kernels.

Linear SVM

$$\min \frac{1}{2} \|w\|^2$$

$$\text{s.t.} \quad y_i (w x_i + b) \geq 1$$ → smjh nahi aaya!!

$$\|w\|^2 = w^T w = w \cdot w$$

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum \alpha_i [y_i (w x_i + b) - 1]$$

$$\therefore \rightarrow \sum \alpha_i y_i = 0$$

$$\alpha_i \geq 0.$$

Tricky way on right

$$\max \sum \alpha_i - \frac{1}{2} \sum_n \sum_i y_i \alpha_i (x_i \cdot x_k) y_k \alpha_k$$

$$\alpha_i \geq 0$$

Non linear SVM → Linear SVM Samgraanj

Non linearly separable data

Kernel Trick

$$W = \sum \alpha_i y_i x_i \quad b = ?$$

→ This is a support vector

- y should be -1 or 1
- x_i should exist for support vectors
- α_i should only be zero if it is not SVM

Page No.
Date: / /20

RBF ⇒ other kernel

$$K(x, x') = \exp\left(-\left(\frac{\|x - x'\|}{2\sigma^2}\right)\right)$$

→ sigma

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

→ variance across the axis

→ this particular function converts to ∞ dimension

Hard margin classifiers

Soft margins classifiers

* Hinge loss

$$\min \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m} \max(0, 1 - y_i(Wx_i + b))$$

hinge loss

C → constant but scalar

Page No.
Date: / /20

Question to ask:
- C → no. of mistakes
- $Wx_i + b$ → how they are related

Trying to reduce the mistake

How decision boundary looks-

17/1/24

Performance
↓
perform well independent of data

Regression
↓
→ MSE
→ Boston idea (1)

Classification
↓

Predicted \ Actual	+ve	-ve
+ve	TP	FP
-ve	FN	TN

TP ⇒ True Positive
FP ⇒ False

Confusion Matrix

	+ve	-ve
+ve	10	15
-ve	6	5

Classification

Performance Metrics

Accuracy

* Accuracy = $\frac{TP + TN}{All}$ → Problem with this if Data unbalanced

Accuracy is still high but doing not good in other cases

Do Biology people task Accuracy

* Precision = $\frac{TP}{TP + FP}$ = $\frac{1}{1 + \frac{1}{FP}}$ → Total positive predicted

mark + ve rest -ve

* Recall = $\frac{TP}{TP + FN}$ → Actual positive ⇒ mark everything -ve

Covid Test

How can I cheat Precision

→ find one which has full chance of positive and rest negative

Recall ↑ → Precision ↓

What if Precision ↑ → Recall ↓?

Page No.
Date: / /20

Make Born with !!

↓

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

↓

Better than F1 score (Poster idea 2)

Other matrix

* Imbalanced data

SMOTE : Synthetic minority oversampling technique

↓

very simple like KNN

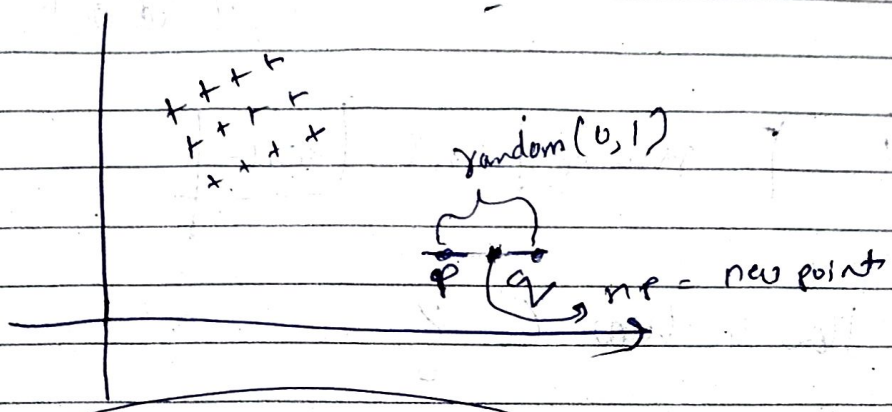
Poster Idea (3)

$1 + \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots$

$\frac{2 \times 1 \times 1}{1+1} = \frac{2(1)}{2(1+1)}$

Page No.
Date: / /

What is the issue with F1 score?



- Pick sample in pair
- $n_p = p + \kappa(0,1)(q-p)$

→ Pick samples

→ pick 1 in the KNN

Problem → random outlier

Undersampling

* Validation

→ parameter

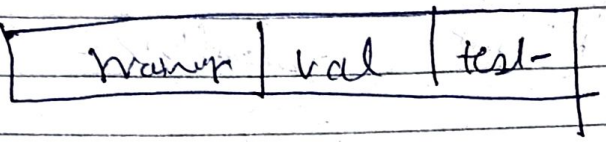
→ hyperparameter → height of decision tree

↳ C, in non linear SVM

kernel, ...

dataset

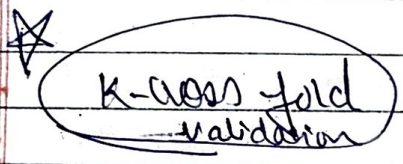
↳ hold out

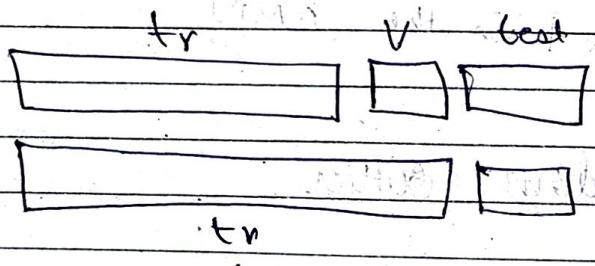


How to choose C ?

* grid search

Table when C value written
& kernel value is written.





K-fold
5-fold

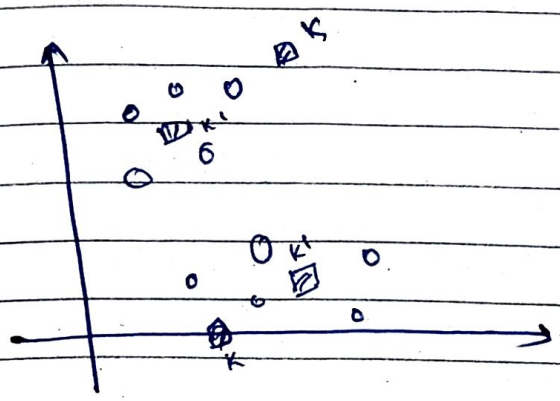
holdout

18/1/24

K fold

Page No.
Date: / /20

Unsupervised



→ ① Clustering

K Means

DBSCAN

(ϵ, n)

Randomly pick k centroids

Ensemble learn

Bias

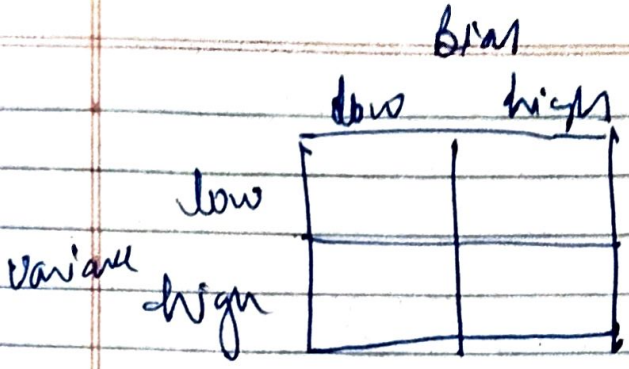
Variance → How much change in the data can change the model

Assumptions that we make lead to an error

def → SOT

weighs → DT

1) Bagging → D → D_1, D_2, D_3, D_4 } Bootstrapping



2) Boosting → $y = y_p + \text{residual}$

↑
Prediction

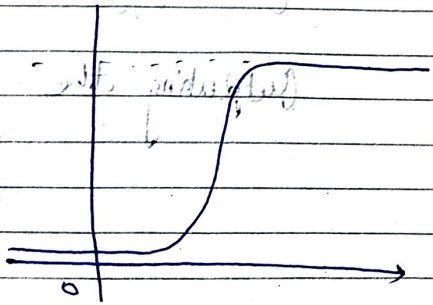
3) Grad Boost
Ada Boost
Xh Boost

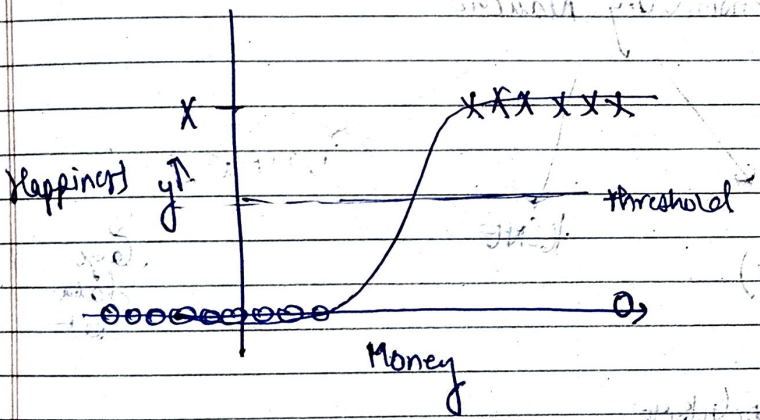
Page No.
Date: / /20

Classification (supervised)

Logistic Regression

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





$$f(x) = \frac{1}{e^{-(wx+b)}+1}$$

How to find error, find the least MSE

$$L_{w,b} = \prod f_{w,b}(x_i)^{y_i} \left(1 - f_{w,b}(x_i)\right)^{1-y_i}$$

Maximum likelihood

log (max likelihood)