01.22.24

"Intro to Statistical Learning"

Statistical learning: Theoretical foundation of ML:

used for 2 things:
→ Find a predict^n → supervised learning
→ Grouping → Unsupervised "

Supervised vs Unsupervised:

Supervised: Build a Model to predict some output from part data
eg: from wt to height.
Part data → y → Labels

Unsupervised: Inputs but no output (y) labels

Notations: y = output, response, dependent variable

x_i = Input, independent variable, predictors, features

n = number of observations (samples)

p = number of features / predictors

$\in$ = "is an element of"

$a \in B$

a is an element of set B

(RT.0.)

example:

	TV	Radio	Newspapers	Sales
n				

y = Sales
x_i = TV, radio, Newspaper (feature গুলো)
p = 3

Data কে matrix form এ লেখা :

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & & & \\ \vdots & & & \\ x_{n1} & \cdots & & x_{np} \end{bmatrix}$$

n = row sample or observation

Overview of notations with an example:

X = [TV] [Radio] [Newspaper]
x_1 x_2 x_3

[y Predicting] → [Given x]
Output input

Predicting / estimating the function :

we know X_i but no y

$$\hat{y} = \hat{f}(x)$$

y vs $\hat{y}$

y → True Label

$\hat{y}$ → estimated / predicted label

y vs $\hat{y}$ → funct^n need to be better

↓

so that error can be reduced

⇓

i) Reducible error

ii) Irreducible error

$[f(x) - \hat{f}(x)]^2$ → squaring for

→ getting + values

→ Amplifying the error so that model can perform better.

$$E[(y - \hat{y})^2] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{reducible error}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible error}}$$

estimated funct^n → so that error can be reduced / minimized during training

(P.T.O.)

After training :

Testing / inference

(i) Which predictors are the most important ?

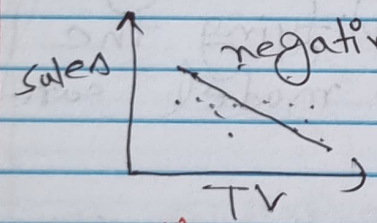
The model should predict

↓

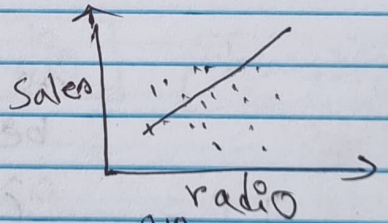
~~কোথায়~~ invest ~~করব~~,
TV / news

(ii) Is there a relationship between X and Y ?

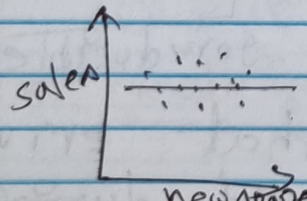
(iii) What kind of relationship exists between X and Y ?



negative - correlation



Positive correlation between sales and radio



→ no relationship / doesn't affect the sale anyway

Take some data
train a function
Test a "

How do we estimate f ?
We wanna find f such that
$Y \approx f(X)$

2 ways to estimate the function:
→ parametric method
→ Non- " "

Parametric Method:
2 step:

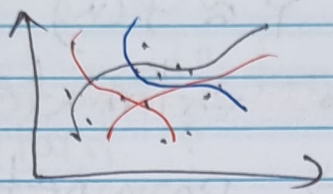
(i) Assume the function be of a certain form
eg: Linear fxn
$f(x) = \beta_0 + \beta_1 x_1$

x_1	y

(ii) Apply a procedure to fit / train the model.

Non-parametric method:
We don't make any explicit assumption
We try to find the funct^n itself from the data

→ No assumptn. of the fun form
→ Estimate the fun from the dataset.



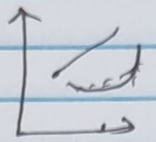
It requires lots of data.
eg: Deep - learning methods

01.24.24
Lec: 03

$y = f(x)$

parametric

$\hat{y} = \beta_1 x_1 + \beta_2 x_2$

$\beta_1 x_1 + \beta_2 x_2 \rightarrow$ 

Non-parametric → Data
↓
which funct^n can fit the data
uncountable model can fit # the data

XAI →
Explainable AI → computer vision model
↓
for classificat^n
eg: pic of cat, which parts are model taking

How do we measure the Accuracy of a Model;
qualitative : Labels, classes
eg: Default / Not-default

quantitative : Deals with numbers
Predict → GPA / numbers, sales, age.

Methods for qualitative → classification Method

Quantitative → Regression methods

Regression methods :

$y = f(x)$ ⇒ True funct^n

after training, $\hat{y} = \hat{f}(x)$ ⇒ estimated "

x_1	x_2	y
TV	radio	sales

$\hat{y}$ will give the estimate of sales

$y = 40$, $\hat{y} = 30$ } $y - \hat{y} = 40 - 30 = 10$

↓

over all the test data

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

Mean Squared Error

accumulate every error

↓

error will grow up

↓

so we take avg $\left[\frac{+++}{\text{total}}\right]$

Predicted response value is close to the true response value

** $\frac{1}{n}\sum_{i=1}^{n}\left[y - \hat{f}(x)\right]^2$ $\hat{y} = \hat{f}(x)$

MSE → for Regression Model (accurate)

Bias - variance Trade - off :

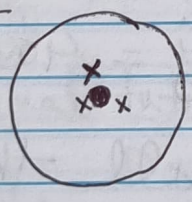
[Simple prob → complicated tool
compli → simple]

Bias → How good/accurate a model is
Accuracy, good

variance → How different a Model behaves with different data

Model 1

model hits →



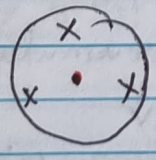
→ doesn't hit but close enough

↓

bais → ↑↑

variance → Low

Model 2



Bias → ↓

variance → ↑

Bias + variance] balance needed

↑ Bias
↑ variance } → good but ~~e~~ charge ~~করলে~~ different thing show করে model.

$$E\left[(y_0 - \hat{f}(x_0)\right] = \text{Expected test MSE}$$

↓
Accuracy of model
↓
we expect that

$$\underbrace{\text{Var}\left[\hat{f}(x_0)\right] + \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]}_{\text{reducible error}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible error}}$$

Classification :

$$y_0 = \hat{y}_0$$

If predict is = to true value,

$$y_0 \neq \hat{y}_0 \text{ (error)}$$

Human → eye
but in comp → indicator function

$$\frac{1}{n}\sum_{i=1}^{n} I(y_0 \neq \hat{y}_0)$$

↓
either true / false

x_1	x_2	Bun
		yes
		NO
		yes
		NO

or Ave $\boxed{I(y_0 \neq \hat{y}_0)}$

finding Indicator function

↓

variety of methods

(35 page / ISLP book)

Basic:

~~Base~~ - Bayes classifier : probability

$\boxed{y_0 \neq \hat{y}_0}$ → we want to minimize this

this classifier minimize this based on Probability.

2 - blue markers

5 - red "

blue মার্বেল probability → $\frac{2}{7}$

conditional probability

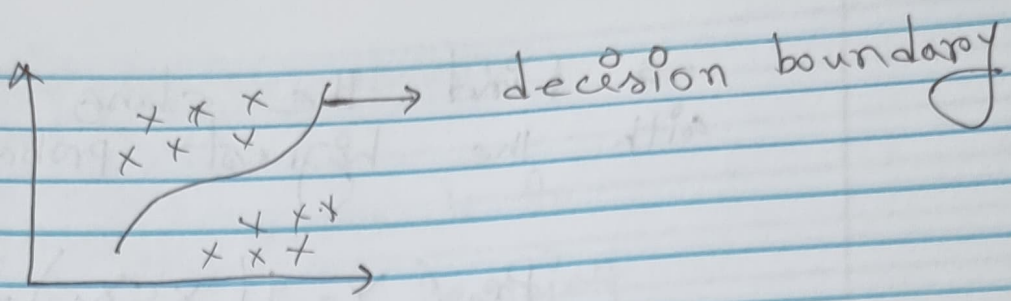
$\boxed{Pr(Y = j \mid X = x_0)}$ that largest

for 2 classes →

yes
no
maybe

$Pr(Y = j \mid X = x_0) > \frac{1}{2}$ or $\frac{1}{3}$

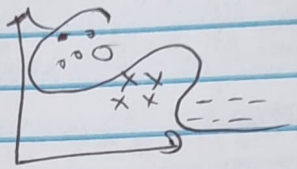
যদি class 1/2 এর বড় হয় তাহলে নিবে



using decision boundary → try to minimize Ave I

by separating the classes

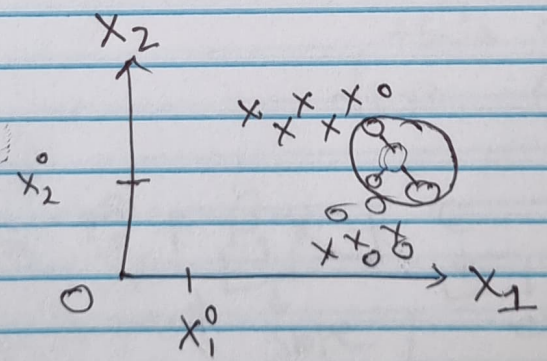
3- classes :



K - Nearest Neighbours (KNN):

Given k, X_0, (test data)

→ we wanna identify k- points in the training data closest to X_0, denoted as N_0



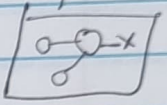
→ Find the class in N_0 (neighbors) with the highest probability.

$$Pr\left(\underset{(class)}{Y=j} \mid x \subset X_0\right) = \frac{1}{k}\sum_{i=N_0}^{k} I(y_0 = j)$$

$(\frac{1}{3} \times 3)$

for $X_1 = 0$

for 0, $= 1$
class

$\frac{1}{3} \times 2 =$ 

$\frac{1}{3} \times 1 = \frac{1}{3}$

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

$= \frac{3}{3} = 1$

$k = 3 \cdots 7$

$\frac{3}{5}$