Unit - 2: Decision Theory.

Bayesian Decision Theory:

Bayes Theorem:

→ named after british mathematician Thomas Bayes, is a mathematical formula for determining Conditional probability.

→ where it is used? ⇒ Trying to find answer.

ie, there are only 2 possible events for given question:

A: It will rain tomorrow.
B: It will not rain tomorrow.

} 50% chance to rain tomorrow if it raining today.

→ Probability theory is framework for making decisions under uncertainity.

→ In classification, baye's rule is used to calculate the probabilities of classes

→ Thus, baysian decision thory, is based on the existence of prior distributions of parameters.

Example:

* Tossing a coin is a random process because we cannot predict at any toss whether the outcome will be head or tail – thats why we toss coin, or buy lottery tickets.
* we can only talk about the probability that the outcome of next toss will be heads or tails.
* extra piece of knowledge that we don't have access to are named unobservable variables.
* In coin tossing example, only observable variable is the outcome of the toss.

$$x = f(z)$$

observable ← x, z → unobservable

where $f()$ is deterministic function that defines the outcome from the unobservable piece of knowledge.

* Let us say $x=1$ denotes that outcome of toss is head
  $x=0$ " " " toss is tails.

Such 'x' are bernoulli distributed where the parameter of distribution p_0 is the probability that outcome is head:

$$P(x=1) = p_0$$
$$P(x=0) = 1 - P(x=1)$$
$$= 1 - p_0$$

Assume we want to predict the outcome of next toss. If we know p_0,

prediction will be heads if $p_0 > 0.5$,
otherwise tails

Eg: x^t is random variable &

$x^t = 1$ if outcome of toss 't' is head.
$x^t = 0$ " " tail.

Sample = {H, H, H, T, H, T, T, H, H}

$x = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$

$$p_0 = \frac{\sum_{t=1}^{N} x^t}{N} = \frac{6}{9} = 0.66$$

$\therefore p_0 = \text{head}$ //

$p_0 =$ No. of toss with outcome head / total No. of toss

Baye's formula:

— also called as baye's rule or baye's law.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$$

$P(A|B)$ → posterior
$P(B|A)$ → likelihood
$P(B)$ → evidence
$P(A)$ → prior

Eg: Bill is 35 years old & earns 40000$/yr. He has a very fair credit rating. Will he buy a computer?

i.e, $$P(H|X) = \frac{P(X|H) * P(H)}{P(X)}$$

where,

X = bill is 35yr old with fair credit rating & income of 40000$/yr.

H = Hypothesis that bill will buy the computer.

P(H|X) = prob. that bill will buy computer given that his age, income & credit rating. (posterior)

P(H) = prob. that bill will buy computer (regardless of knowning his age, income etc.) (Prior)

P(X|H) = prob. that someone is 35yr old, has fair credit rating, earns 40,000$/yr & has bought the computer [Likelihood]

P(X) = prob. that bill is 35yr old, has fair credit rating & earns 40000/yr [Evidence]

Bayesian classifiers:

- based on baye's theorem.
- are statistical classifiers, used to find the prob. that a given tuple belong to particular class.
- exhibits high accuracy & speed when applied to large DB.

$P(X|C_i)$ - is prob. of seeing x as I/P when it is known to belong to class C_i.

The posterior probability of class C_i is:

$$P(C_i|X) = \frac{P(X|C_i) P(C_i)}{P(X)}$$

$$= \frac{P(X|C_i) P(C_i)}{\sum_{K=1}^{K} P(X|C_K) P(C_K)}$$

↓ chain rule

Baye's classifier: Prerequisites:

Random variable:

- is a function that maps a possible set of outcomes to some values

Eg: Tossing a coin & getting head H as 1 & " tail T as 0

where 0 & 1 are random variable.

prior (or) state of Nature:

- is how likely each class is going to occur
- priors are known before training process.

class conditional probability or likelihood:

- prob. of how likely a feature x occurs given that it belong to particular class.
- denoted by $P(x|A)$ where x is particular feature.
- is the quantity that we have to evaluate while training the data.

Evidence:

- prob. of occurrence of particular feature. $P(x)$.
- Can be calculated using chain rule;

$$P(x) = \sum_i P(x|w_i) P(w_i)$$

posterior probability:

- prob. of occurrence of class A when certain features are given.
- It is what we aim at computing in test phase in which we have testing I/P or features.

Naive Bayes classification:

- is a program which predicts a class value, given a set of attributes.
- use baye's rule to derive conditional probabilities for class variable.
- finally output the class with highest probability.

- The dataset is divided into two parts:
  - ↳ feature matrix
  - ↳ response vector.
- feature matrix, contains all the vectors (rows) of dataset in which each vector consists of value of dependent feature
- response vector, contains the value of class variable (prediction or output) for each row of feature matrix.

Example:

1) Given DB: Training Data from Allelectronics Customer DB:-
[check whether buys_computer = Yes or No based on given condition]

ID	age	Income	Student	Credit_rating	class: buys_computer
1	Youth	high	no	fair	no
2	"	"	no	excellent	"
3	middle-age	"	"	fair	yes
4	Senior	medium	"	"	"
5	"	low	yes	"	"
6	"	"	"	excellent	no
7	middle-age	"	"	"	yes
8	youth	medium	no	fair	no
9	"	low	yes	"	yes
10	Senior	medium	"	"	"
11	youth	"	"	excellent	"
12	middle-age	"	no	"	"
13	middle-age	high	yes	fair	"
14	Senior	medium	no	excellent	no

X = (age = youth, Income = medium, Student = yes, Credit_rating = fair)

Now,

P (buys-comp = yes) = 9/14 → total (yes+no) (9 yes)
= 0.643

P (buys_comp = no) = 5/14 (5 no)
= 0.357

To compute $P(x|C_i)$ for $i = 1, 2, \ldots$

$P(\text{age} = \text{youth} | \text{buys\_comp} = \text{yes}) = 2/9 = \boxed{0.222}$

$P(\text{age} = \text{youth} | \text{buys\_comp} = \text{no}) = 3/5 = \boxed{0.600}$

$P(\text{income} = \text{medium} | \text{buys\_comp} = \text{yes}) = 4/9 = \boxed{0.444}$

$P(\text{income} = \text{medium} | \text{buys\_comp} = \text{no}) = 2/5 = \boxed{0.400}$

$P(\text{student} = \text{yes} | \text{buys\_comp} = \text{yes}) = 6/9 = \boxed{0.667}$

$P(\text{student} = \text{yes} | \text{buys\_comp} = \text{no}) = 1/5 = \boxed{0.200}$

$P(\text{credit\_rating} = \text{fair} | \text{buys\_comp} = \text{yes}) = 6/9 = \boxed{0.667}$

$P(\text{credit\_rating} = \text{fair} | \text{buys\_comp} = \text{no}) = 2/5 = \boxed{0.400}$

Using these,

$P(x | \text{buys\_comp} = \text{yes}) = P(\text{age} = \text{youth} | \text{buys\_comp} = \text{yes}) \times P(\text{income} = \text{medium} | \text{buys\_comp} = \text{yes}) \times P(\text{student} = \text{yes} | \text{buys\_comp} = \text{yes}) \times P(\text{credit\_rating} = \text{fair} | \text{buys\_comp} = \text{yes})$

$= 0.222 \times 0.444 \times 0.667 \times 0.667$

$= \boxed{0.044}$

Similarly,

$P(x | \text{buys\_comp} = \text{no}) = 0.600 \times 0.400 \times 0.200 \times 0.400$

$= \boxed{0.019}$

To find the class C_i, maximize $P(x|C_i) \cdot P(C_i)$,

$P(x | \text{buys\_comp} = \text{yes}) \cdot P(\text{buys\_comp} = \text{yes}) = 0.044 \times 0.643$

$= \boxed{0.028}$

$P(x | \text{buys\_comp} = \text{no}) \cdot P(\text{buys\_comp} = \text{no}) = 0.019 \times 0.357$

$= \boxed{0.007}$

Among both [yes] is has a more probability

$\therefore$ the naive bayesian classifier predicts

$\boxed{\text{buys\_computer} = \text{yes}}$ for tuple x.

2) Training Example : Play Tennis

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	"	"	"	Strong	"
D3	Overcast	"	"	Weak	Yes
D4	Rain	Mild	"	"	"
D5	"	Cool	Normal	"	"
D6	"	"	"	Strong	No
D7	Overcast	"	"	"	Yes
D8	Sunny	Mild	High	Weak	No
D9	"	Cool	Normal	"	Yes
D10	Rain	Mild	"	"	"
D11	Sunny	"	"	Strong	"
D12	Overcast	"	High	"	"
D13	"	Hot	Normal	Weak	"
D14	Rain	Mild	High	Strong	No

Given Instance,

X = (Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

Now,

$P(\text{Play} = \text{Yes}) = 9/14 = \boxed{0.643}$

$P(\text{Play} = \text{No}) = 5/14 = \boxed{0.357}$

To compute $P(X|C_i)$ for $i = 1, 2, \ldots$

$P(\text{Outlook} = \text{Sunny} \mid \text{Play} = \text{Yes}) = 2/9 = \boxed{0.222}$

$P(\text{"} \mid \text{Play} = \text{No}) = 3/5 = \boxed{0.600}$

$P(\text{Temp} = \text{Cool} \mid \text{Play} = \text{Yes}) = 3/9 = \boxed{0.333}$

$P(\text{Temp} = \text{cool} \mid \text{Play} = \text{No}) = 1/5 = \boxed{0.200}$

$P(\text{Humidity} = \text{High} \mid \text{Play} = \text{Yes}) = 3/9 = \boxed{0.333}$

$P(\text{"} \mid \text{Play} = \text{No}) = 4/5 = \boxed{0.800}$

$P(\text{Wind} = \text{Strong} \mid \text{Play} = \text{Yes}) = 3/9 = \boxed{0.333}$

$P(\text{"} \mid \text{Play} = \text{No}) = 3/5 = \boxed{0.600}$

Testing phase:

$$P(\text{Yes}|x) = [P(\text{Sunny}|\text{Yes}) \times P(\text{cool}|\text{Yes}) \times P(\text{High}|\text{Yes}) \times P(\text{Strong}|\text{Yes})] * P(\text{Play} = \text{yes})$$

$$= [0.222 \times 0.333 \times 0.333 \times 0.333] * 0.643$$

$$= 0.008 * 0.643$$

$$= \boxed{0.005}$$

Similarly for,

$$P(\text{No}|x) = [0.600 \times 0.200 \times 0.800 \times 0.600] * 0.357$$

$$= 0.072 * 0.357$$

$$= \boxed{0.025}$$

$\therefore$ $P(\text{Yes}|x) < P(\text{No}|x)$.

naive bayes classifier predicts **play_tennis = 'No'**.

Advantage:

→ fast for training & prediction.
→ easily interpretable.
→ can handle both continuous & discrete data.

Disadvantage:

→ data scarcity & chances of loss of accuracy.
→ Zero frequency.
i.e, if the category of any variable is not seen in training dataset then model assigns a zero prob. to that category & prediction cannot be made.

Applications:

→ Text classification.
→ Sentiment Analysis
→ Recommendation System.

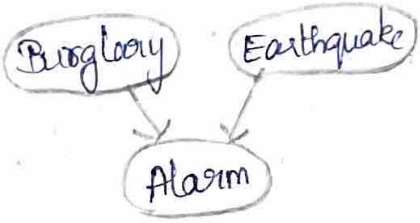
Classification:

Bayesian Network:

- capture joint prob of events represented by model.
- bayesian belief n/w describe the joint prob distribution for set of variables.
- Naive bayes is restricted form of bayesian n/w where nodes should have no parents & ~~edges~~ node corresponding to attribute variable have no edges b/w them.
- BN is made up of two parts:
  1. directed acyclic graph
  2. Set of parameters.

⇒ Directed Acyclic graph:

- nodes are random variables. (Can be discrete or continuous)
- arrows connect pair of nodes. (X is parent of Y if there is an arrow from X to Y)



⇒ Set of parameters:

- parameters are prob. in these conditional prob. distribution.
- each node X_i has conditional prob. distribution $P(X_i \mid Parents(X_i))$ that quantifies effect of parents on node

burglary

B	P(B)
F	0.999
T	0.001

} = 1

earthquake

E	P(E)
F	0.998
T	0.002

} = 1

Alarm

B	E	A	P(A|B,E)
F	F	F	0.999
F	F	T	0.001
F	T	F	0.71
F	T	T	0.29
F	F	F	0.06
T	F	T	0.94
T	T	F	0.05
T	T	T	0.95

$P(A \mid B, E)$ — parents

} = 1 } = 1 } = 1 } = 1

Conditional probability Tables.

Bayesian Belief network:

- "A bayesian n/w is a probabilistic graphical model which represents a set of variables & their conditional dependencies using directed acyclic graph".
- Also called as bayes n/w, belief n/w, decision n/w or bayesian model.

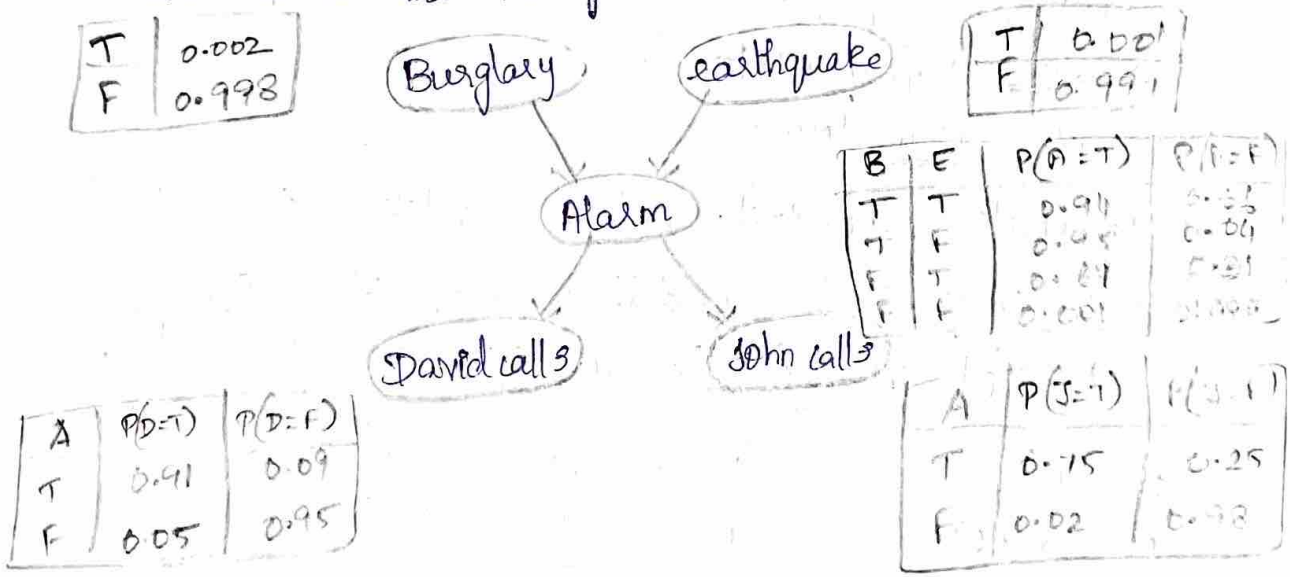
Joint probability Distribution:

* If we have variables of $x_1, x_2, \ldots x_n$, then prob. of diff. combination of $x_1, x_2, \ldots x_n$ is known as joint prob. distribution.

$$P[x_1, x_2, \ldots x_n] = P[x_1 | x_2, x_3 \ldots x_n] \cdot P[x_2, x_3 \ldots x_n]$$
$$= P[x_1 | x_2, x_3 \ldots x_n] \cdot P[x_2 | x_3, x_4 \ldots x_n] \cdots P[x_{n-1} | x_n] \cdot P[x_n]$$

Example:

(1) Calculate the prob. that alarm has sounded, but there is neither burglary nor earthquake occurred, & david, John both called the Harry.



List of all events in this n/w:

- Burglary (B)
- earthquake (E)
- Alarm (A)
- David calls (D)
- John calls (J)

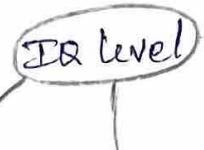
$P(J, D, A, \neg B, \neg E) = ?$

$$P(J, D, A, \neg B, \neg E) = P(J|A) * P(D|A) * P(A|\neg B \wedge \neg E) * P(\neg B) * P(\neg E)$$
$$= 0.75 * 0.91 * 0.001 * 0.998 * 0.999$$
$$= \underline{0.00068045}$$

---

(2)

e^0	e^1
0.7	0.3



i^0	i^1
0.8	0.2

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

	s^0	s^1
i^0	0.75	0.25
i^1	0.4	0.6

	a^0	a^1
m^0	0.6	0.4
m^1	0.9	0.1

i. Calculate the prob. that inspite of exam level being difficult, the student having a low IQ & low Aptitude score, manages to pass the exam & secure admission.

$P[A=1, M=1, I=0, E=1, S=0] = ?$

$= P(a^1|m^1) \times P(m^1|i^0 e^1) \times P(i^0) \times P(e^1) \times P(s^0|i^0)$

$= 0.1 * 0.1 * 0.8 * 0.3 * 0.75$

$= \underline{0.0018}$

ii. Calculate the prob that the student has a high IQ level & Aptitude score, the exam being easy yet fails to pass & doesn't secure admission.

$P[A=0, M=0, I=1, E=0, S=1] = ?$

$= P(a^0|m^0) \times P(m^0|i^1 e^0) \times P(i^1) \times P(e^0) \times P(s^1|i^1)$

$= 0.6 * 0.5 * 0.2 * 0.7 * 0.6$

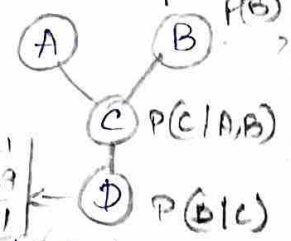
$= \underline{0.0252}$

---

3) Suppose there are four boolean random variables:

0	0.2
1	0.8

P(A)



P(B)

0	0.6
1	0.4

	D=0	D=1
C=0	0.1	0.9
C=1	0.9	0.1

A	B	C=0	C=1
0	0	0.2	0.8
0	1	0.6	0.4
1	0	0.3	0.7
1	1	0.5	0.5

Joint prob. distribution of random variable is given by:

$$P(A, B, C, D) = P(A) \cdot P(B) \cdot P(C|A, B) \cdot P(D|C)$$

i. Find $P(A=0, B=1, C=0, D=1)$

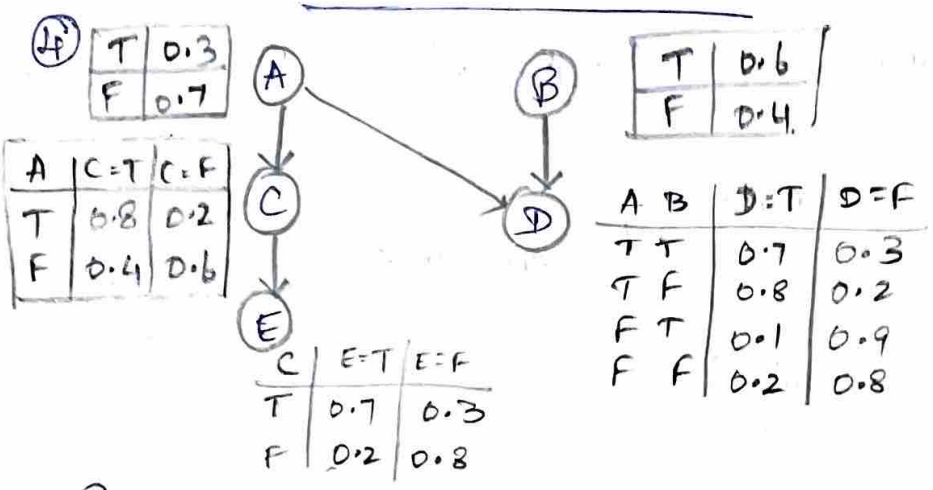
$= P(A=0) * P(B=1) * P(C=0 | A=0, B=1) * P(D=1 | C=0)$

$= 0.2 * 0.4 * 0.6 * 0.9$

$= \underline{0.0432}$

---

ii)



T	0.3
F	0.7

T	0.6
F	0.4

A	C=T	C=F
T	0.8	0.2
F	0.4	0.6

A	B	D=T	D=F
T	T	0.7	0.3
T	F	0.8	0.2
F	T	0.1	0.9
F	F	0.2	0.8

C	E=T	E=F
T	0.7	0.3
F	0.2	0.8

Given: $P(A=T) = 0.3$, $P(C=T | A=T) = 0.8$

$P(B=T) = 0.6$, $P(C=T | A=F) = 0.4$

$P(D=T | A=T, B=T) = 0.7$

$P(D=T | A=T, B=F) = 0.8$

$P(D=T | A=F, B=T) = 0.1$

$P(D=T | A=F, B=F) = 0.2$

$P(E=T | C=T) = 0.7$

$P(E=T | C=F) = 0.2$

i, $P(D=T) = ?$

$= P(D=T, A=T, B=T) + P(D=T, A=T, B=F) + P(D=T, A=F, B=T) + P(D=T, A=F, B=F)$

$= P(D=T | A=T, B=T) \cdot P(A=T, B=T) + P(D=T | A=T, B=F) \cdot P(A=T, B=F) + P(D=T | A=F, B=T) \cdot P(A=F, B=T) + P(D=T | A=F, B=F) \cdot P(A=F, B=F)$

$= P(D=T | A=T, B=T) \cdot P(A=T) P(B=T) + P(D=T | A=T, B=F) \cdot P(A=T) P(B=F) + P(D=T | A=F, B=T) \cdot P(A=F) \cdot P(B=T) + P(D=T | A=F, B=F) \cdot P(A=F) \cdot P(B=$

$= (0.7 \times 0.3 \times 0.6) + (0.8 \times 0.3 \times 0.4) + (0.1 \times 0.7 \times 0.6) + (0.2 \times 0.7 \times 0.4)$

$= \underline{0.32}$

ii, $P(A=T | C=T) = ?$ (0.8, 0.3)

$= \dfrac{P(C=T | A=T) \cdot P(A=T)}{P(C=T)}$

now, $P(C=T) = P(C=T, A=T) + P(C=T, A=F)$

$= P(C=T | A=T) \cdot P(A=T) + P(C=T | A=F) \cdot P(A=F)$

$= (0.8 \times 0.3) + (0.4 \times 0.7)$

$= 0.52$

$$\text{So } \frac{P(C=T \mid A=T) \cdot P(A=T)}{P(C=T)} = \frac{0.8 \times 0.3}{0.52}$$

$$= \underline{0.46}$$

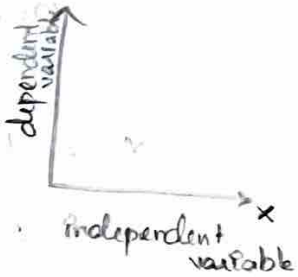
Regression:

- is a statistical modeling to show the relationship b/w two variables with linear regression.
- is form of predictive modeling technique which investigate the relationship b/w a dependent & independent variable.

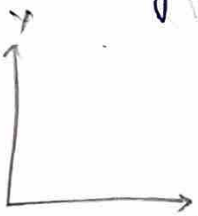
Simple Linear Regression:

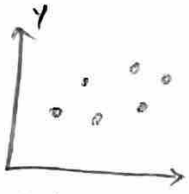
- data is modeled using straight line.
- $\boxed{y = b_1 x + b_0}$ is a straight line. where

  x = independent variable.
  y = dependent variable.
  b_1 = slope
  b_0 = y-intercept.

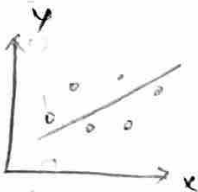


- every value of 'x' there is a corresponding value of 'y' in it if it is a continuous.

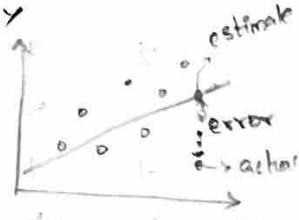




plot data points

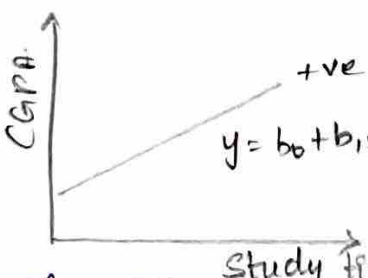


draw regression line

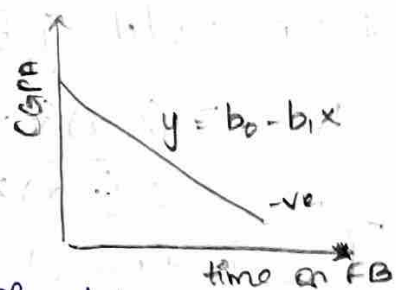


plot new predicted point to reduce error

Positive & Negative slope:



- If data point in x axis increases then increase in y axis value is called a positive slope.



- If data point in x increasing & dependent variable y decreasing is called negative slope.

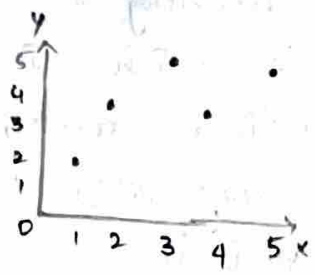
Example:

1) Given:

independent X	dependent Y
1	2
2	4
3	5
4	4
5	5

Find the regression line & R^2 co-efficient.

Step 1: plot data



S2: calculate slope.

Straight line formula: $\boxed{y = b_0 + b_1 x}$

where slope $b_1 = \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}$

and $\bar{x}$ = mean of x

$\bar{y}$ = mean of y

$\bar{x} = \dfrac{1+2+3+4+5}{5} = \boxed{3}$

$\bar{y} = \dfrac{2+4+5+4+5}{5} = \boxed{4}$

X	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
$\bar{x}=3$	$\bar{y}=4$			$\Sigma = 10$	$\Sigma = 6$

Linear model: $\boxed{\hat{y} = b_0 + b_1 x}$

$b_1 = \dfrac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2} = \dfrac{6}{10} = \boxed{0.6}$

$\therefore \boxed{\hat{y} = b_0 + 0.6x} \longrightarrow (1)$

Now, at mean $(\bar{x}, \bar{y})$, $(1) \Rightarrow b_0 = \bar{y} - 0.6\bar{x}$

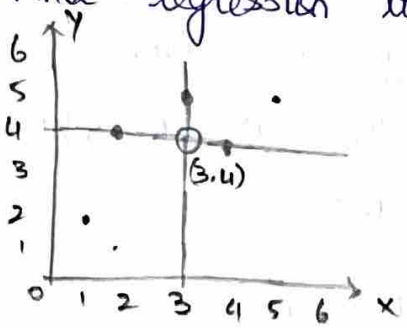
$= 4 - 0.6(3)$

$\boxed{b_0 = 2.2}$

Sub b_0 in equ ①.

$\therefore$ $\underbrace{\bar{y} = 0.6x}_{b_1} + \underbrace{2.2}_{b_0}$ is a linear model equation.

Q3: Find regression line.



At mean (3,4)

given $x = \{1, 2, 3, 4, 5\}$

$\hat{y} = 0.6x + 2.2$

Sub x in y equ:

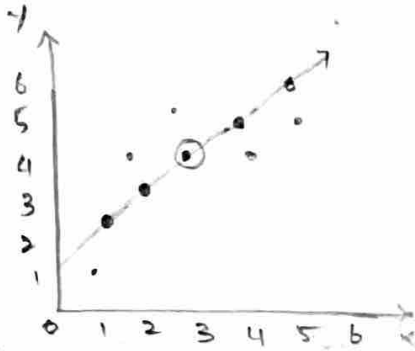
$y = 0.6(1) + 2.2 = 2.8$

$y = 0.6(2) + 2.2 = 3.4$

$y = 0.6(3) + 2.2 = 4$

$y = 0.6(4) + (2.2) = 4.6$

$y = 0.6(5) + 2.2 = 5.2$



$\therefore$ Regression line points are:

(1, 2.8) (2, 3.4) (3, 4) (4, 4.6) (5, 5.2)

Mean Squared error:

* task is find the distance b/w actual & predicted value & reduce the error or reduce the distance.
* line with least error is line of regression.
* R^2 value is a statistical measure of how close the data to be fitted in regression line

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

where,

- $\hat{y}$ = predicted value of y
- $\bar{y}$ = mean value of y.

$\therefore$ from the given points,

$$R^2 = 1 - \frac{(2-2.8)^2 + (4-3.4)^2 + (5-4)^2 + (4-4.6)^2 + (5-5.2)^2}{(2-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (5-4)^2}$$

$$= 1 - \frac{2.4}{6} \Rightarrow 1 - 0.4$$

$$R^2 = 0.6$$

Condition:

If $R^2 \simeq 0.3$ ⇒ data points are far away from regression line

If $R^2 \simeq 0.7$ ⇒ data points are closer to the regression line.

Hence for the given values the $R^2 = 0.6$ which is closer to regression line.

---

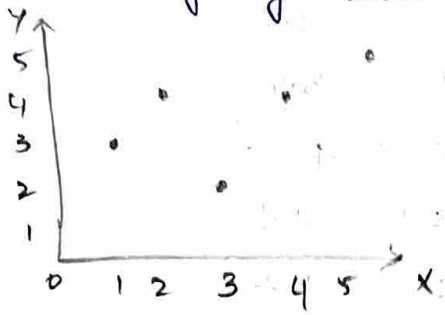
2) Given:

X	1	2	3	4	5
Y	3	4	2	4	5

a) find least square regression line $y = ax + b$

b) Estimate value of y when $x = 10$

Sl:



S2:

$$\bar{x} = \frac{1+2+3+4+5}{5} = \boxed{3}$$

$$\bar{y} = \frac{3+4+2+4+5}{5} = \boxed{3.6}$$

X	Y	$x-\bar{x}$	$Y-\bar{Y}$	$(x-\bar{x})^2$	$(x-\bar{x})(Y-\bar{Y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
$\bar{x}=3$	$\bar{Y}=3.6$			$\Sigma = 10$	$\Sigma = 4$

$\therefore \quad \bar{y} = ax + b$

$$a = \frac{\Sigma (x-\bar{x})(Y-\bar{Y})}{\Sigma (x-\bar{x})^2} = \frac{4}{10} = \boxed{0.4}$$

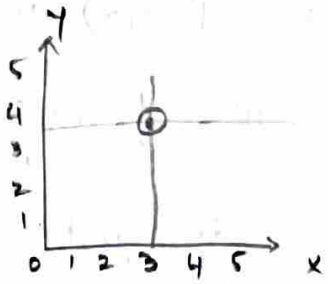
$\therefore \quad \boxed{y = 0.4x + b} \rightarrow (1)$

Now, at mean (3, 3.6)

① equ ⇒ $b = y - 0.4x$

$= 3.6 - 0.4(3)$

$\boxed{b = 2.4}$



Thus, $\boxed{y = 0.4x + 2.4}$ is a linear model

Q3: find regression line,

$x = \{1, 2, 3, 4, 5\}$

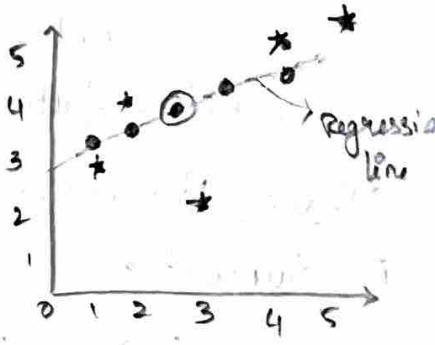
$y = 0.4(1) + 2.4 \Rightarrow 2.8$

$y = 0.4(2) + 2.4 \Rightarrow 3.2$

$y = 0.4(3) + 2.4 \Rightarrow 3.6$

$y = 0.4(4) + 2.4 \Rightarrow 4.0$

$y = 0.4(5) + 2.4 \Rightarrow 4.4$



∴ points for Regression line are (1, 2.8), (2, 3.2), (3, 3.6), (4, 4.0), (5, 4.4)

(b) when $x = 10$,

$y = 0.4(10) + 2.4$

$\boxed{y = 6.40}$ //

Association Mining:

- finding regularities in data.

i.e what products were often purchased together.

* Frequent patterns are patterns that appear frequently in a data set

↳ Set of items, such as milk & bread, that appear frequently together in a transaction dataset is a frequent itemset

⇒ Association Rules:

- is an implication of the form $\boxed{X \to Y}$ where 'X' is an antecedent & 'Y' is consequent of rule.
- used to find dependency b/w two items X & Y.

Eg: $P(Y|X)$ → is prob. that somebody who buys X also buys Y.

i.e. 90% of transaction that purchase bread & butter also purchase milk.

$X \rightarrow Y$

"If" part = antecedent [before → part i.e. X]

"then" part = Consequent [after → part i.e. Y]

Here antecedent & consequent are disjoint. (i.e, have no items in common)

Antecedent = bread & butter

Consequent = milk

Confidence factor = 90%.

Three Measures:

① Support:

$$\text{support}(X,Y) = P(X,Y) = \frac{\text{No. of cust. who bought X and Y}}{\text{No. of customers}}$$

② Confidence:

$$\text{Conf}(X \rightarrow Y) = P(Y|X) = \frac{P(X,Y)}{P(X)} \leftarrow \text{support}$$

$$\therefore \text{Conf}(X \rightarrow Y) = \frac{\text{No. of customers who bought X \& Y}}{\text{No. of customers who bought X.}}$$

③ Lift (or) interest:

$$\text{Lift}(X \rightarrow Y) = \frac{P(X,Y)}{P(X) \cdot P(Y)} = \frac{P(Y|X)}{P(X)}$$

Market basket Analysis:

- is the earliest form of frequent pattern mining & association rules.

↳ Apriori Alg.:

- proposed by Agrawal et al in 1996.

2 Steps:

1. finding frequent itemset. i.e, those which have enough support
2. Converting them to rules with enough Confidence.

Concepts:

- An Item : is an item/article in basket.
- I : the set of all items sold in store.
- transaction : items purchased in basket, it may have TID (transaction ID)
- transactional dataset : set of all transactions.

Example:

1) find all frequent itemset using Apriori with min. support of 2 and confidence of 50% for following transactions:

TID	List of items ID's
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

1- Item set:

1-item set	freq.
I1	6
I2	7
I3	6
I4	2
I5	2

⟹ take only items with freq ≥ 2 i.e, min-supp = 2

Freq. 1-item set	freq.
I1	6
I2	7
I3	6
I4	2
I5	2

2-Item set:

2-item set	freq.
I1, I2	4
I1, I3	4
I1, I4	1
I1, I5	2
I2, I3	4
I2, I4	2
I2, I5	2
I3, I4	0
I3, I5	1
I4, I5	0

⟹

Freq. 2-item set	freq.
I1, I2	4
I1, I3	4
I1, I5	2
I2, I3	4
I2, I4	2
I2, I5	2

3-Item set:

3-Item sets	freq.
I1, I2, I3	2
I1, I2, I4	1
I1, I2, I5	2
I1, I3, I4	0
I1, I3, I5	1
I1, I4, I5	0
I2, I3, I4	0
I2, I3, I5	1
I2, I4, I5	0
I3, I4, I5	0

→

freq. 3-Item sets	freq.
I1, I2, I3	2
I1, I2, I5	2

4-Item Set:

4-Item set	freq.
I1, I2, I3, I5	1

⇒ Not possible

Now,

freq. 3-Item set :

$I = \{1, 2, 3\}$ & $\{1, 2, 5\}$

Non-empty subsets 'S' are :

- {1} {2} {3} {1,2} {1,3} {2,3}
- {1} {2} {5} {1,2} {1,5} {2,5}

How to form Association Rule?

for every non-empty subset 'S' of I,

$S \rightarrow (I - S)$

If $Supp(I) / Supp(S) \geq min\_conf(50\%)$

(1) for Item Set I : {1,2,3} Subsets: {1} {2} {3} {1,2} {1,3} {2,3}

$S \rightarrow (I - S)$

Rule 1: {1} → {2,3}

Supp = 2/9 ← (1,2,3) occur 2 times; ← total Transaction

= 22.22 %

$Conf = \frac{Supp(1,2,3)}{Supp(1)} = \frac{2/9}{6/9} = 2/6 = 33.34\% < 50\%$

∴ Invalid Rule.

Rule 2 : {2} → {1,3}

Supp = 2/9 = 22.22%

Conf = $\frac{\text{Supp}(1,2,3)}{\text{Supp}(2)} = \frac{2}{7} = \boxed{28.57\%} < 50\%$ ✗

∴ Invalid Rule.

Rule 3 : {3} → {1,2}

Supp = 2/9 = 22.22%

Conf = $\frac{\text{Supp}(1,2,3)}{\text{Supp}(3)} = \frac{2}{6} = \boxed{33.3\%} < 50\%$ ✗

∴ Invalid Rule.

Rule 4 : {1,2} → {3}

Supp = 2/9 = 22.22%.

Conf = $\frac{\text{Supp}(1,2,3)}{\text{Supp}(1,2)} = \frac{2}{4} = \boxed{50\%} \geq 50\%$ ✓

∴ Valid Rule.

Rule 5 : {1,3} → {2}

Supp = 2/9 = 22.22%.

Conf = $\frac{\text{Supp}(1,2,3)}{\text{Supp}(1,3)} = \frac{2}{4} = \boxed{50\%} \geq 50\%$ ✓

∴ Valid Rule.

Rule 6 : {2,3} → {1}

Supp = 2/9 = 22.22%.

Conf = $\frac{\text{Supp}(1,2,3)}{\text{Supp}(2,3)} = \frac{2}{4} = \boxed{50\%} \geq 50\%$ ✓

∴ Valid Rule.

② for Itemset 2 : {1,2,5}, Subsets : {{1} {2} {3} {1,2} {1,5} {2,5}}

Rule 1 : {1} → {2,5}

Supp = 2/9 = 22.22%

Conf = Supp(1,2,5)/Supp(1) ⇒ 2/6 = $\boxed{33.34\%} < 50\%$.

∴ Invalid Rule

Rule 2 : {2} → {1,5}

Supp = 2/9 = 22.22%.

Conf = Supp(1,2,5)/Supp(2) ⇒ 2/7 ≈ $\boxed{28.57\%} < 50\%$ ✗

∴ Invalid Rule.

Rule 3: {5} → {1,2}

Supp = 2/9 = 22.22%

Conf = Supp(1,2,5)/Supp(5) = 2/2 ⇒ 100% > 50% ✓

∴ Valid Rule.

Rule 4: {1,2} → {5}

Supp = 2/9 = 22.22%

Conf = Supp(1,2,5)/Supp(1,2) = 2/4 = 50% ≥ 50% ✓

∴ Valid Rule.

Rule 5: {1,3} → {2}

Supp = 2/9 = 22.22%

Conf = Supp(1,2,3)/Supp(1,3) = 2/4 = 50% ≥ 50% ✓

∴ Valid Rule.

Rule 6: {2,3} → {1}

Supp = 2/9 = 22.22%

Conf = Supp(1,2,3)/Supp(2,3) = 2/3 = 66.7% > 50% ✓

∴ Valid Rule.

Hence all the valid Rules are the association rule for the given transaction.

---

(2) Find the frequent itemset for following Transaction DB with min. Supp = 50%

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Since total Tid = 4 & min support is

∴ min. Supp = 2

i.e, half of total Tid

Also since min. conf. is not given, we take all possible association rules.

1-Item Set:

1-Itemset	freq.
A	2
B	3
C	3
D	1
E	3

⇒

freq. 1-Itemset	freq.
A	2
B	3
C	3
E	3

2-item sets:

2 item set	freq.
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

⇒

freq. 2-item set	freq.
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

3-item set:

3-item set	freq.
A, C, E	1
A, B, E	1
B, C, E	2

⇒

freq. 3 item set	freq.
B, C, E	2

Now possible Association Rule for {B, C, E} are:

I = {B, C, E}   S = {B} {C} {E} {B,C} {B, E} {C, E}

∴ $S \rightarrow I - S$

i.e, $B \rightarrow \{C, E\}$; $C \rightarrow \{B, E\}$; $E \rightarrow \{B, C\}$

$\{B, C\} \rightarrow E$; $\{B, E\} \rightarrow C$; $\{C, E\} \rightarrow B$

Losses And Risks:

- caused due to decisions that are not equally good or costly

Eg: A financial institution when making a decision for loan applicant should take into account the potential gain & loss as well.

Actions : α_i - as decision to assign the i/p to class C_i

Loss : λ_{ik} - is loss incurred for taking action α_i when the i/p actually belong to C_k.

∴ Expected risk for taking acting α_i is:

$$R(\alpha_i | x) = \sum_{k=1}^{K} \lambda_{ik} P(C_k | x)$$

Decision: Choose α_i with minimal expected risk.

i.e, choose α_i, $R(\alpha_i | x) = \min_k R(\alpha_k | x)$

⇒ **Loss & Risk : 0/1 Loss :**

Correct decision : 0 loss

Incorrect decision : 1 loss.

Let us define 'k' actions α_i where $i = 1, 2, \ldots k$.

Special case of 0/1 loss is :

$$\lambda_{ik} = \begin{cases} 0, & \text{if } i = k \\ 1, & \text{if } i \neq k. \end{cases}$$ (Correct decision has no loss)

Now, risk for taking action α_i is :

$$R(\alpha_i | x) = \sum_{k=1}^{k} \lambda_{ik} P(C_k | x)$$

$$= \sum_{k \neq i} 1 \cdot P(C_k | x)$$ • if no loss then risk is 0 since $\lambda_{ik} = 0$ • if k ≠ i then $\lambda_{ik} = 1$

$$R(\alpha_i | x) = 1 - P(C_i | x)$$

Decision : To minimize the risk, take the most probable class $P(C|x)$

⇒ **Loss & Risk : Reject :**

- In some applications, wrong decisions namely misclassification may have very high cost.

Eg : manual decision is made if automatic system has low certainty of its decision.

Here, consider an extra option : $(k+1)^{st}$ class i.e. reject

$$\therefore \text{loss function } \lambda_{ik} = \begin{cases} 0, & \text{if } i = k \\ \lambda, & \text{if } i = k+1 \\ 1, & \text{if otherwise} \end{cases} \quad 0 < \lambda < 1$$

Risk of reject is,

$$R(\alpha_{k+1} | x) = \sum_{k=1}^{k} \lambda P(C_k | x)$$

$$= \lambda$$

Risk of misclassification is,

$$R(\alpha_i | x) = \sum_{k \neq i} P(C_k | x)$$

$$R(\alpha_i | x) = 1 - P(C_i | x)$$

Decision: Choose C_i which min. $R(\alpha_i|x)$, $1 \le i \le k$.

o/p: Choose C_i, if $P(C_i|x) > P(C_k|x)$ for all $k \neq i$ & $P(C_i|x) > 1-\lambda$

reject, otherwise.

Hence, If $\boxed{\lambda = 0}$ → we always reject
i.e, correct Classification.

If $\boxed{\lambda \ge 1}$ → never reject
ie, costlier than error

Discriminant Functions: (function of several variables used to assign items into 1 or more group

- classification can be seen as implementing a set of discriminant functions, $g_i(x)$, $i = 1, 2, \dots k$.

Choose C_i if $\boxed{g_i(x) = \max g_k(x)}$

$$g_i(x) = \begin{cases} -R(\alpha_i|x), & \text{max. corresponds to min. risk.} \\ P(C_i|x), & \text{using 0/1 loss function.} \\ p(x|C_i)\,P(C_i), & \text{ignore } p(x) \end{cases}$$

This divides the feature space into 'k' decision regions of $R_1, R_2 \dots R_k$ where

$$R_i = \{x \mid g_i(x) = \max g_k(x)\}$$

When there is 2 classes, define a single discriminant as:

$$\boxed{g(x) = g_1(x) - g_2(x)}$$

Choose $\begin{cases} C_1, & \text{if } g(x) > 0 \\ C_2, & \text{otherwise} \end{cases}$

When $\boxed{K = 2}$ ⇒ it is dichotomizer classification system.

& $\boxed{K \ge 3}$ ⇒ it is polychotomizer.

Parametric Methods: (how we estimate prob from given dataset)

- Statistic is any value that is calculated from given sample.
- In statistical inference, we make decision using the information provided by sample.
- first approach is parametric where we assume that sample is drawn from some distribution that obeys a known model, for eg. Gaussian
- advantage of parametric approach is that model is defined to small No. of parameters.

  Eg: mean, variance - sufficient statistics of distribution
- so we estimate parameters of distribution from sample. use these estimates in assumed model, get an estimated distribution → which we use to make decision.
- method used to estimate parameters of distribution is "maximum likelihood estimation".

Parametric Estimation:

★ Density estimation is general case of estimating $p(x)$.

↳ for classification:

estimated densities are class densities $p(x|C_i)$

prior $[p(C_i)]$, posteriors, $[p(C_i|x)]$

↳ for regression:

estimated density is $p(y|x)$

where 'x' is 1-dimensional & hence the densities are univariate.

for an independent & identically distributed sample,

$$X = [x^t]_{t=1}^N$$

x^t is instance drawn from some known prob. density family.

$$x^t \sim p(x|\theta)$$

want to find θ that makes sampling x^t from $p(x|\theta)$ as likely as possible.

Maximum Likelihood Estimation:

- Since x^t are independent, likelihood of parameter θ given sample X is the product of likelihood of individual points.

$$l(\theta|X) \equiv p(X|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

- Search for θ that maximizes the likelihood which is denoted by $l(\theta|x)$.
- we use log of likelihood where it takes max. value.

Log likelihood is defined as:

log() → converts product to sum

$$L(\theta|x) \equiv \log l(\theta|x) = \sum_{t=1}^{N} \log p(x^t|\theta)$$

∴ maximum likelihood estimator (MLE):

$$\theta^* = \arg\max_{\theta} L(\theta|x)$$

* If we have 2-class problem, distribution used is → Bernoulli
* where there are $K > 2$ classes, generalization used is → Multinomial
* most frequently used density for modeling class-conditional i/p densities → Gaussian (normal) density.

① Bernoulli Density:

- event occurs & bernoulli random variable x takes value 1 with prob. p & non-occurrence of event has prob. $1-p$ & denoted by x taking value 0.

i.e. $\boxed{P(x) = p^x (1-p)^{1-x}}$, $x \in (0, 1)$

Expected value & variance are:

$$E(x) = \sum_x x\, p(x) = 1 \cdot p + 0 \cdot (1-p)$$

$$\boxed{E(x) = p}$$

$$\text{Var}(x) = \sum_x [x - E(x)]^2 p(x)$$

$$\boxed{\text{Var}(x) = p(1-p)}$$

p is only parameter.

We need to calculate the estimator, $\hat{p}$:

log likelihood is:

$$L(p|x) = \log \prod_{t=1}^{N} p^{(x^t)} (1-p)^{(1-x^t)}$$

$$= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log (1-p)$$

$\therefore$ after solving for $\frac{dL}{dp} = 0$,

estimator: $\boxed{\hat{p} = \frac{\sum_t x^t}{N}}$

② Multinomial Density:

- Instead of two states, the outcome of random event is one of 'k' mutually exclusive & exhaustive states
- each of which has prob. of occurring p_i with $\boxed{\sum_{i=1}^{k} p_i = 1}$
- x_i is 1 if outcome is state 'i' or 0 otherwise.

$$\boxed{P(x_1, x_2, \cdots x_k) = \prod_{i=1}^{k} p_i^{x_i}}$$

$\therefore$ MLE of p_i is

$$\boxed{\hat{p}_i = \frac{\sum_t x_i^t}{N}}$$

⑤ Gaussian (Normal) Distribution:

X is gaussian distributed with mean $E[X] = \mu$ and $\text{var}(X) = \sigma^2$, denoted as $N(\mu, \sigma^2)$.

density function, $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$, $-\infty < x < \infty$

log likelihood is,

$$L(\mu, \sigma | x) = \frac{-N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t - \mu)^2}{2\sigma^2}$$

After taking partial derivatives,

$$m = \frac{\sum_t x^t}{N}$$ (m: mean) and $$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$ (s²: s.d) //

Dimensionality Reduction:

- is the transformation of data from high-dimensional space into low-dimensional space, so that low dimensional representation retains some meaningful properties of original data.

Techniques:

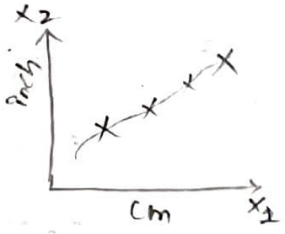
1. Principal Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA).

Dimension Reduction:

- is a process of converting a dataset having vast dimensions into dataset with lesser dimensions.
- ensure that converted dataset conveys similar info.

Eg: Consider graph with 2 dimensions x_1 & x_2

- x_1 → represents measurement of object in cm
- x_2 → " " " inches



Using dimension reduction techniques:

- Convert dimension of data from 2 (x_1 & x_2) to 1 (z_1)

Benefits:

- → Compress the data - which reduce storage space.
- → eliminate redundant feature.
- → Improve model performance.
- → reduce time required for computation.

⇒ PCA:

- transforms the variables into new set of variables called as principal components.
- there can be only two principal components for two dimensional data set.

PCA Alg:

S1 - Get data
S2 - Compute mean vector (μ)
S3 - Substract mean from given data.
S4 - Calculate covariance matrix.
S5 - Calculate eigen vectors & eigen values of covariance matrix
S6 - Choose components & form feature vector.
S7 - Derive new data set

Problem:

1) Given:

Data = {2, 3, 4, 5, 6, 7 ; 1, 5, 3, 6, 7, 8}

(or)

Data = (2,1) (3,5) (4,3) (5,6) (6,7) (7,8)

(or)

$$\text{Data} = \begin{cases} \underline{\text{Class 1}} : X = 2,3,4 \; ; \; Y = 1,5,3 \\ \underline{\text{Class 2}} : X = 5,6,7 \; ; \; Y = 6,7,8 \end{cases}$$

Compute the principle components using PCA Alg.

S1: Given feature vectors are:

$x_1 = (2,1)$ $x_2 = (3,5)$ $x_3 = (4,3)$ $x_4 = (5,6)$ $x_5 = (6,7)$ $x_6 = (7,$

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

S2: Calculate mean vector (μ)

$$\mu = \frac{2+3+4+5+6+7}{6}, \frac{1+5+3+6+7+8}{6}$$

$$= (4.5, 5)$$

$\therefore$ Mean Vector (μ) $= \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$

S3: Substract mean vector (μ) from given feature vectors.

$x_1 - \mu = (2-4.5, 1-5) \Rightarrow (-2.5, -4)$

$x_2 - \mu = (3-4.5, 5-5) \Rightarrow (-1.5, 0)$

$x_3 - \mu = (4-4.5, 3-5) \Rightarrow (-0.5, -2)$

$x_4 - \mu = (5-4.5, 6-5) \Rightarrow (0.5, 1)$

$x_5 - \mu = (6-4.5, 7-5) \Rightarrow (1.5, 2)$

$x_6 - \mu = (7-4.5, 8-5) \Rightarrow (2.5, 3)$

$$\begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

S4: Calculate Covariance matrix.

$$\text{Covariance matrix} = \frac{\sum (x_i - \mu)(x_i - \mu)^t}{n}$$

$$m_1 = (x_1 - \mu)(x_1 - \mu)^t = \begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} 2.5 & -4 \end{bmatrix} = \begin{bmatrix} 6.25 & 10 \\ 10 & 16 \end{bmatrix}$$

$$m_2 = (x_2 - \mu)(x_2 - \mu)^t = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -1.5 & 0 \end{bmatrix} = \begin{bmatrix} 2.25 & 0 \\ 0 & 0 \end{bmatrix}$$

$$m_3 = (x_3 - \mu)(x_3 - \mu)^t = \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} -0.5 & -2 \end{bmatrix} = \begin{bmatrix} 0.25 & 1 \\ 1 & 4 \end{bmatrix}$$

$$m_4 = (x_4 - \mu)(x_4 - \mu)^t = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$m_5 = (x_5 - \mu)(x_5 - \mu)^t = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 1.5 & 2 \end{bmatrix} = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$$

$$m_6 = (x_6 - \mu)(x_6 - \mu)^t = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix} \begin{bmatrix} 2.5 & 3 \end{bmatrix} = \begin{bmatrix} 6.25 & 7.5 \\ 7.5 & 9 \end{bmatrix}$$

now,

Covariance matrix $= \frac{m_1 + m_2 + m_3 + m_4 + m_5 + m_6}{6}$

$= \frac{1}{6}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$ - after adding all values of $m_1, m_2 \ldots$

$$M = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

Q5: Calculate eigen value & eigen vector

'λ' is eigen value for matrix 'M'.

$$|M - \lambda I| = 0$$

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\begin{vmatrix} 2.92-\lambda & 3.67 \\ 3.67 & 5.67-\lambda \end{vmatrix} = 0$$

$\therefore (2.92-\lambda)(5.67-\lambda) - (3.67)(3.67) = 0$

$16.56 - 2.92\lambda - 5.67\lambda + \lambda^2 - 13.47 = 0$

$$\boxed{\lambda^2 - 8.59\lambda + 3.09 = 0}$$

$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.

Solving this quadratic equation, $\lambda = 8.22, 0.38$.

$\therefore \boxed{\lambda_1 = 8.22}$ & $\boxed{\lambda_2 = 0.38}$ are two eigen values.

Q6: Take the highest eigen value.

The eigen vector with highest eigen value is the principle component of given data.

Equation to find eigen vector: $\boxed{MX = \lambda X}$

when M = covariance matrix.

X = eigen vector

λ = eigen value.

$$\therefore \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 8.22 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Solving these,

$$2.92x_1 + 3.67x_2 = 8.22x_1$$
$$3.67x_1 + 5.67x_2 = 8.22x_2$$

on simplification.

$$5.3x_1 = 3.67x_2 \longrightarrow (1)$$
$$3.67x_1 = 2.55x_2 \longrightarrow (2)$$

from (1) & (2):

$$x_1 = 0.69x_2 \Rightarrow \frac{x_1}{x_2} = \frac{0.69}{1}$$

$$\therefore \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.69 \\ 1 \end{bmatrix}$$

Now normalize this value,

$$e_1 = \begin{bmatrix} 0.69/\sqrt{0.69^2+1^2} \\ 1/\sqrt{0.69^2+1^2} \end{bmatrix} = \begin{bmatrix} 0.568 \\ 0.823 \end{bmatrix}$$

Q7: Derive new dataset

Principle component PC1	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}

$$P_{11} = e_1^T \begin{bmatrix} x_1 - \bar{x} \\ y_1 - \bar{y} \end{bmatrix} = \begin{bmatrix} 0.568 & 0.823 \end{bmatrix} \begin{bmatrix} 2-4.5 \\ 1-5 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0.568 & 0.823 \end{bmatrix} \begin{bmatrix} -2.5 \\ -4 \end{bmatrix}$$

$$= -1.42 - 3.292$$

$$\boxed{P_{11} = -4.712}$$

$$P_{12} = e_1^T \begin{bmatrix} x_2 - \bar{x} \\ y_2 - \bar{y} \end{bmatrix} \Rightarrow \begin{bmatrix} 0.568 & 0.823 \end{bmatrix} \begin{bmatrix} 3-4.5 \\ 5-5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.568 & 0.823 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix}$$

$$\boxed{P_{12} = -0.852}$$

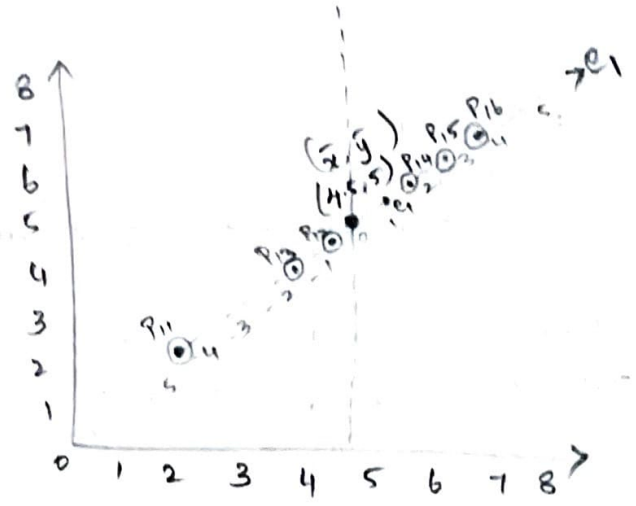
$$\boxed{P_{13} = -1.93}$$

$$\boxed{P_{14} = 1.136}$$

$$\boxed{P_{15} = 2.498}$$

$$\boxed{P_{16} = 3.889}$$

	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}
∴ PC	-4.712	-0.852	-1.93	1.136	2.498	3.859



project the PC into new subspace

Linear Discriminant Analysis: (LDA)

- was developed in the year 1936 by Ronald A. Fisher & was named Linear Discriminant or Fisher's Discriminant Analysis.
- originally was described as two-class technique later generalized by C.R Rao as multiple discriminant analysis.
- LDA is supervised classification technique.
- used as pre-processing step in ML.

Representation of LDA models:

① Consists of statistical properties of data, calculated for each class.

② for single I/p variable (x) there is mean & variance of variable for each class.

③ for multiple variables, this is same properties calculated over multivariate gaussian namely mean & co-variance matrix.

④ these statistical properties are estimated from data & plug into LDA equation to make predictions.

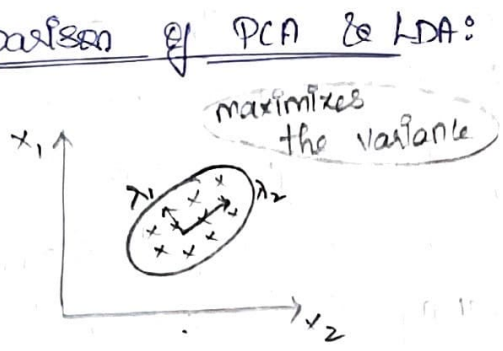
Applications:

- Medical
- Customer identification.
- facial Recognition.

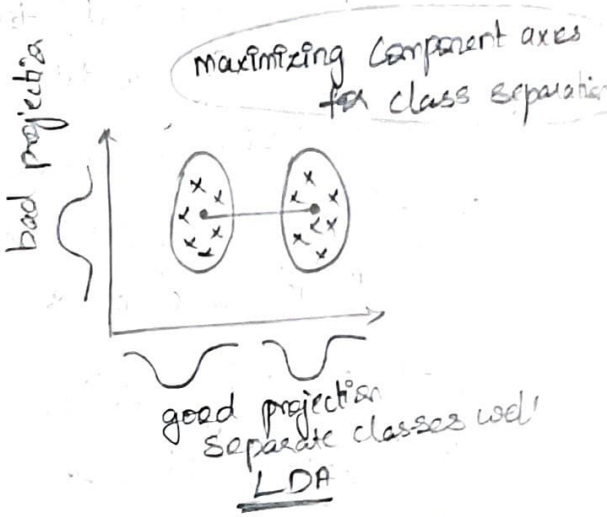
Extensions to LDA:

→ QDA - Quadratic Discriminant Analysis.
→ Flexible Discriminant Analysis. (FDA)
→ Regularized Discriminat Analysis (RDA).

Comparison of PCA & LDA:



PCA



LDA

* PCA ignores class labels & focuses on finding the principle components that maximizes variance in given data. Hence PCA is unsupervised Algorithm.

* LDA is Supervised alg. that intend to find the linear discriminants that represents those axes which maximize separation b/w different classes.

* LDA perform better multi-class classification task than PCA
* PCA performs better when sample size is small.

⇒ PCA is followed by LDA.