Unit - III

Some basic ~~data~~ statistics :-

Avg :- median, mode

Expectation :-

Variance :- $Var(\{x_j\}) = \sigma^2(\{x_j\})$

$$= E((\{x_i\} - \mu)^2 \cdot \sum_{i=1}^{N} (x_i - \mu)^2$$

$E \rightarrow$ expected value

$\mu \rightarrow$ mean

$\sigma =$ standard deviation

Relation of S.D and variation is square root of S.D

* procedure of two variables are varying w.r.t x_i and y_i is co-variance.

$$cov(\{x_i\}, \{y_i\}) = E(\{x_i - \mu\}^2) E(\{y_i - \nu\}^2)$$

$\mu \rightarrow$ mean of x_i

$\nu \rightarrow$ mean of y_i

$\Rightarrow$ when ~~the~~ ~~covariance~~ variance is independent the covariance $= 0$.

$\Rightarrow$ If one is $\uparrow$ing other should ~~also~~ is $\uparrow$ing vice-versa then it is +ve

$\Rightarrow$ If one is $\downarrow$ing the other shld $\uparrow$ing v/v then it is -ve.

Covariance matrix:

$$\Sigma = \begin{pmatrix} E(x_1 - \mu_1)(x_1 - \mu_1) \ldots & E(x_1 - \mu_2)(x_1 - \mu_2) \ldots \\ E(x_2 - \mu_2)(x_1 - \mu_1) & E(x_1 - \mu_1)(x_3 - \mu_3) \\ & E(x_1 - \mu_1)(x_n - \mu_n) \end{pmatrix}$$

~~covariance~~ Here diagonal is '0' and it will be symmetric matrix.

⇒ we use this matrix to find the covariance b/w the set of values.

* Mahalamboes distance:

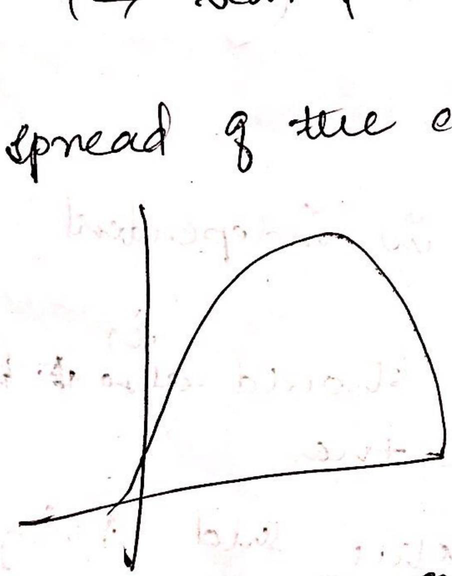
$$D_m(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$x \to$ column vector of data with mean μ and Σ^{-1}

it is inverse of co-variance matrix

$T \to$ transpose

spread of the data: Gussian distribution



G.D of single ~~distribution~~ dimension is given as:

$$P(x) = \frac{1}{\sigma 2\pi} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

This is for higher dimensions.

- when G.D is plotted it is spherical for one dimension.

- when it is more than one dimension then it is identical

- According to the elements of diagonal it is given as

imp 2m

* Bias and Variance:

If we design a model it doesn't work because of two reasons

1) Accurate (Bias)
2) Prediction (Variance)

it is dilema becoz it has to compromise any one for Accuracy (or) prediction.

Error function of Bias and Variance:

$$E[(y - \hat{f}(x))^2] = \sigma^2 + \left[f(x) - \frac{1}{K}\sum_{l=0}^{n} f(x_n)\right]^2 + \frac{\sigma^2}{K}.$$

$K \rightarrow$ no. of samples

$x_n \rightarrow$ function wherein it requires bias and function.

Unit-III Chapter-II

Bayesian learning:-

— most imp M.L algorithm.
— this kind of learning cannot be ignored in M.L
— helps in D.T.A and in MDL (favoring of shortest term) ↳ (Decision tree)

features:-
— Incrementally learning (hypothesis is not rejected either it is used to be true or not)
— make probabilistic ~~decs~~ prediction
— New instance (allow to predict new instances)

Bayes
imp
MPH → most probable hypothesis
MAP → max Aposterior hypothesis

* when the cases is $P(h_1) = P(h_2) = P(h_3)$ of hypotheses having same probabilities then $P(D|h)$ = likelihood of D given h

A hypothesis that maximises $P(D|h)$ is known as maximum likelihood of hypothesis and it is given as,

$$\boxed{h_{ML} = \underset{h \in H}{\mathrm{argmax}}\ P(D|h)}$$

cancer prediction

$\oplus$ $\ominus$

$p(\text{cancer}) = 0.008$ $p(\sim\text{Cancer}) = 0.992$

$p(\oplus|\text{cancer}) = 0.98$ $p(\ominus|\text{cancer}) = 0.02$

$p(\oplus|\sim\text{cancer}) = 0.03$ $p(\ominus|\sim\text{Cancer}) = 0.97$.

$$p(h|D) = \frac{p(D|h)\, p(h)}{p(D)}$$

$p(\oplus|\text{cancer})\, p(\text{cancer}) = (0.98)(0.008) = 0.00784$

$p(\oplus|\sim\text{cancer})\, p(\sim\text{cancer}) = (0.03)(0.992)$

2m

* Bayes optimal classifier :

$h_1 = 0.4$; $h_2 = 0.3$; $h_3 = 0.3$.

$$P(V_j|D) = \sum_{h_i \in H} P(V_j|H_i)\, P(h_i|D)$$

$V_j \rightarrow$ ~~the values~~

B.O.C is defined in terms of argmax i.e.,

$$\boxed{\underset{V_j \in Y}{\text{argmax}} \sum_{h_i \in H} P(V_j|h_i)\, P(h_i|D)}$$

conditions

$P(h_1|D) = 0.4$ $P(\oplus|h_1) = 1$ $P(\ominus|h_1) = 0$

$P(h_2|D) = 0.3$ $P(\ominus|h_2) = 1$ $P(\oplus|h_2) = 0$.

$P(h_3|D) = 0.3$ $P(\ominus|h_3) = 1$ $(P\oplus|h_3) = 0$.

$$\frac{nc+mp}{n+m}$$

p → prior estimate of probability.

If an attribute has k possible values then

$p = 1/k$

m → constant called equivalent sample size which determines how heavily wt. p relative to the observe data.

Graphical Models + probabilistic graphical model.

~~pro~~ Two components in graph model:

→ node

→ link (which connects one node to another)

$A \rightarrow B$

node → random variable

link → A cause B (shows the relation) (reading the graph).

$\boxed{P(a,b) = P(b|a)\,P(a)}$

two types of nodes

→ observed nodes

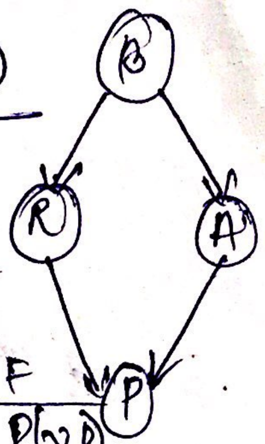
→ latent / hidden nodes

- Graphical model is an powerfull tool to understand the algorithms of M.L

Bayesian n/w + represents in the form of tables which comprises of Directed Acyclic graph and conditional probability.

Eg for B.N + Exam panic

P(b)	P(~b)
0.5	0.5



	T	F
B	P(r)	P(~r)
T	0.3	0.7
F	0.8	0.2

	T	F
B	P(a)	P(~a)
T	0.1	0.9
F	0.5	0.5

		T	F
R	A	P(p)	P(~p)
T	T	0	1
T	F	0.8	0.2
F	T	0.6	0.4
F	F	1	0

$$P(p) = \sum_{b,r,a} P(b,r,a,p)$$

$$P(b) \times P(r|b) \times P(a|b) \times P(p|r,a).$$

(Taking value row-wise)

$$\Rightarrow P(r|b) \times P(a|b) \times P(b|r,a)$$

$$\Rightarrow 0.3\ (0.1) \times 0 + 0.3 \times 0.9 \times 0.8$$

$$+ 0.7 \times 0.1 \times 0.6 \times 0.7 \times 0.9 \times 1$$