Safwan Mahmood Assignment Theory - HW3

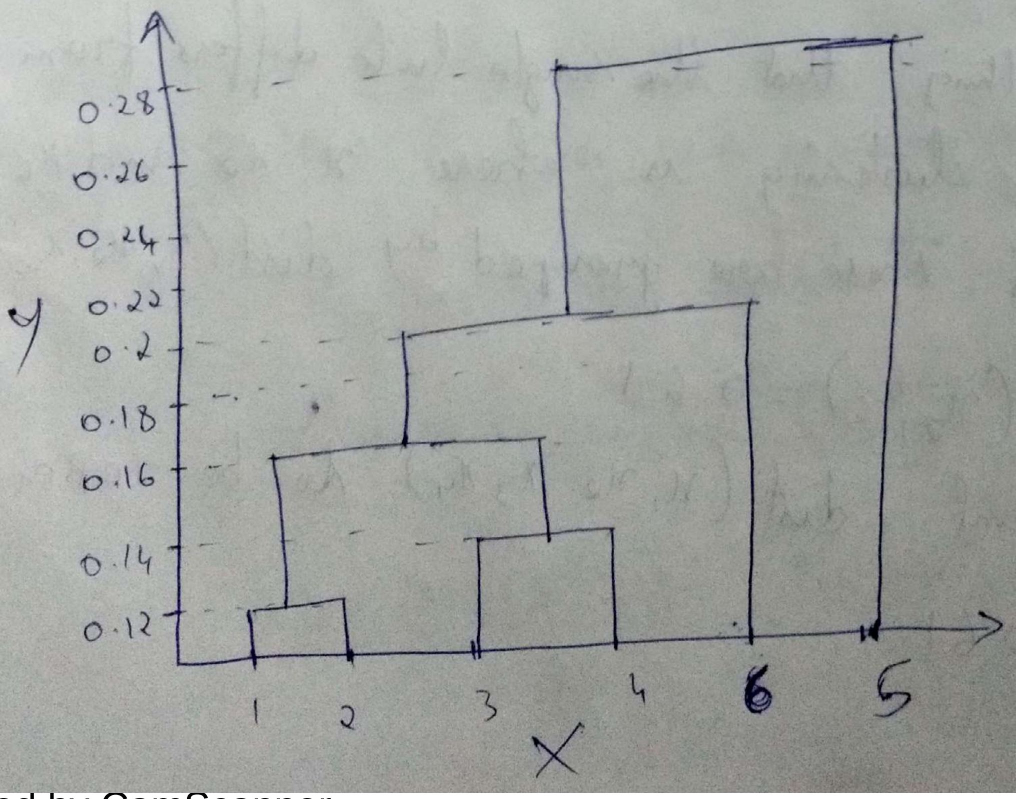
ES14BTECH11017

1) Heirarchial clustering

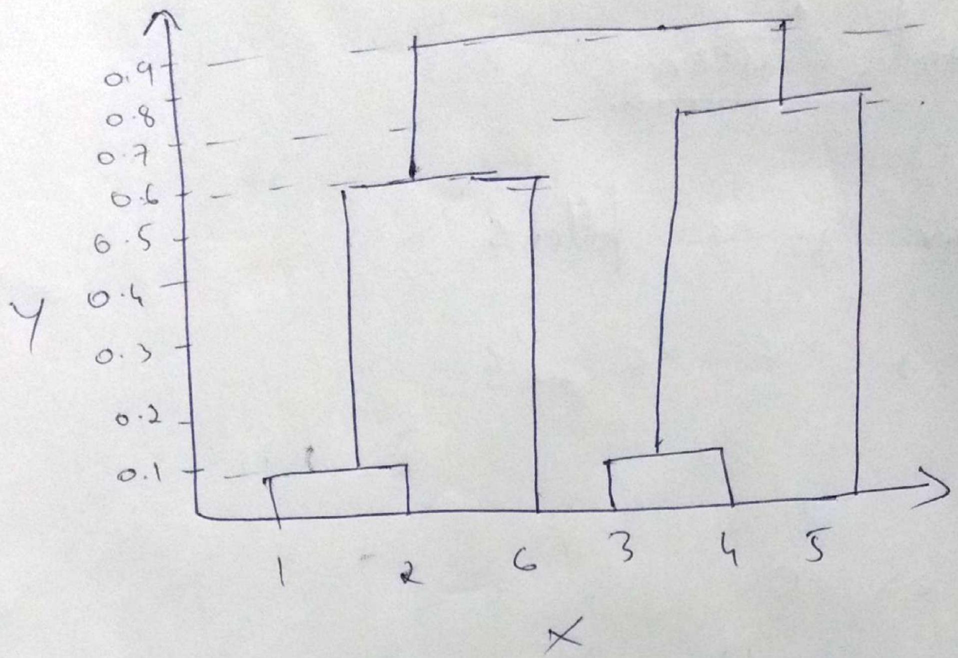
Distance matrix is as follows

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	0.12	0				
x_3	0.51	0.25	0			
x_4	0.84	0.16	0.14	0		
x_5	0.28	0.77	0.70	0.45	0	
x_6	0.34	0.61	0.93	0.2	0.67	0

a) Constructing Dendrogram:



b) Dendrogram for final result of hierarchial clustering with complete link



c) We notice that only 2 values change, so the answer remains the same, i.e it is same as previous answer above.

Now,
The first thing that the single link differs from complete link clustering is where x_1, x_2 and x_6 are grouped. These are grouped by dist (x_1, x_2, x_6)

$\Rightarrow$ dist $(x_2, x_6) = 0.61$

$\therefore$ We want dist $(x_1, x_2\ x_3 x_4)$ to be lesser than 0.61.

$\Rightarrow$ Then we want

$$\text{dist}(x_1 x_2 x_3 x_4, x_6) = \text{dist}(x_3, x_6)$$
$$= 0.93$$

to be lesser than this so that

$x_1 x_2 x_3 x_4$ and x_6 are grouped together.

After these changes both become identical.

2) Covariance Matrix $= C = E[(E[x]-x))(E[x]-x)^T]$

let $n = (\text{some } k) \times m$ (multiple)

$\Rightarrow C$ is

1) $E[a]^2 (M-1)^2 / M^2$ for (i,i).

2) $E[a]^2 \dfrac{(M-1)}{M^2}$ $\forall (x,i)$ or (i,x) $x < n$_samples.

3) $\underline{E[a]^2}/M$ otherwise

Taking Σ we get $\rightarrow$ 2)(a) Consider non-diagonal elements.

2 sample will contribute $= E[a]^2 \dfrac{(M-1)}{M^2}$

others will contribute (3) $\rightarrow \dfrac{E[a]^2}{M}$

$$\therefore \Rightarrow \left(\frac{M-2}{M^2}\right) E[a]^2 - \frac{2E[a]^2(M-1)}{M^2}$$

$$\Rightarrow -E[a]^2/M$$

Diagonal elements

one will contribute to (1) for given (i,i)

rest of them $\rightarrow$ other case

$$\Rightarrow \frac{M-1}{M^2} E[a]^2 - \frac{E[a]^2 (M-1)^2}{M^2}$$

$$= E[a]^2 - \frac{E[a]^2}{M}$$

$$\therefore C_{ij} = \begin{cases} -\frac{1}{M} E[a]^2, & i \neq j \\ \left(\frac{M-1}{M}\right) E[a]^2, & i = j \end{cases}$$

b) We can verify that

$$T = (1, 1, \ldots 1)$$

is given eigen vector of

$$C \times T = \left(\frac{M-1}{M}\right) E[a]^2 + \frac{(M-1) E[a]^2}{M^2}$$

$\rightarrow$ Let differce b/n diagonal and non-diagonal be δ.

$\therefore$ Rank $(C - \delta I) = 1$

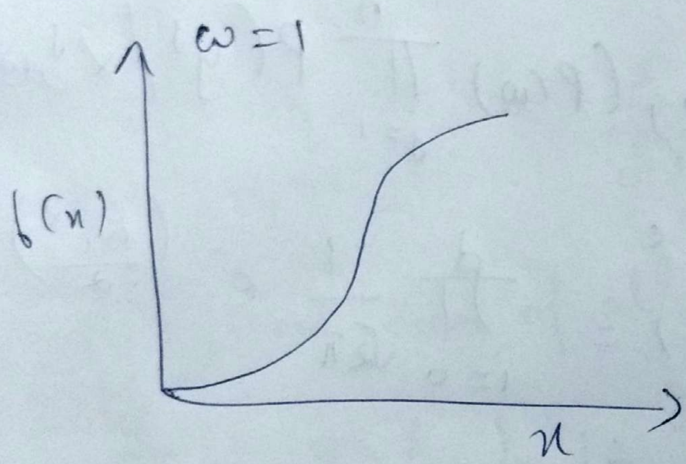
λ(geometrical) $=$ M - rank $(C - \delta I)$

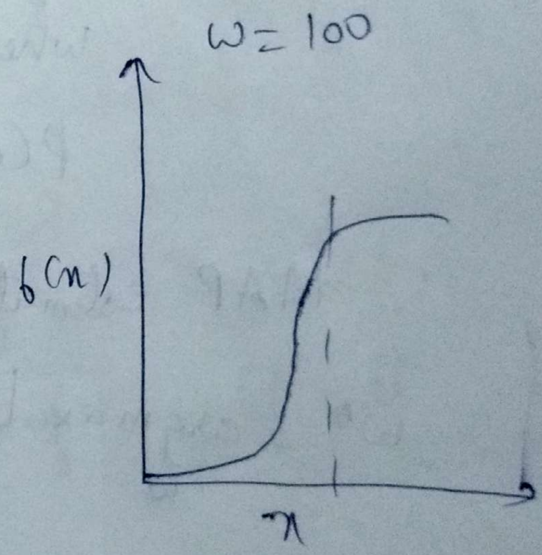
Hence geometric multiplicity of $\delta = M-1$,

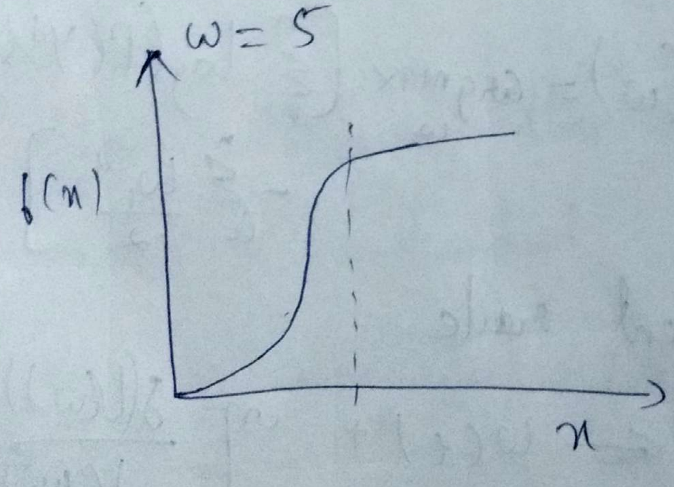
$\therefore$ Rest of $M-1$ eigen vectors have eigen values δ.

c) PCA is not suited as vectors S to be picked by NItk, can have only linear relation between them.

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







We can observe that as ω increases $f(x)$ gets steeper and the curve gets more steeper. It means model is completely sure of the class.

As heights get high, with large heights, small changes in x can lead to large changes in probability leading to misclassification and overfitting.

b) $w = [w_0, w_1, \ldots w_d]^T$

We take here log conditional posterior instead of log conditional likelihood.

i.e $L(w) = \log \left(P(w) \prod_{j=1}^{n} P(y^j / x^j, w) \right)$

where

$$P(w) = \prod_{i=0}^{d} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{w_i^2}{2}\right)}$$

$\therefore$ MAP estimate is

$$w^* = \arg\max_w L(w) = \arg\max_w \left[\sum_{j=1}^{n} \log(P(y^j/x_j, w)) - \sum_{i=}^{n} \frac{w_i^2}{2} \right]$$

$\therefore$ The gradient ascent rule

$$\Rightarrow \quad w_i(t+1) \leftarrow w(t) + \eta \left. \frac{\partial(L(w))}{\partial(w_i)} \right|_t$$

For log conditional posterior, it is

$$\frac{\partial(L(w))}{\partial(w_i)} = \frac{\partial}{\partial w_i} \log P(w) + \frac{\partial}{\partial w_i} \log \left(\prod_{j=1}^{n} (P(y^j/x^j, w) \right.$$

Here the second term is same as derived in before for unregularized case.

First term leads to extra factor of

$$\frac{\delta}{\delta w_i} \log(P(w)) = -w_i$$

Final update is as follows

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta\Big(-w_i^{(t)} + \sum_{j=1}^{n} X_i^j \big(Y^j - P(Y=1 \mid X^j; w^{(t)})\big)\Big)$$

(c) We know sum of all probabilities $= 1$

$$\therefore P(Y = Y_K / X) = 1 - \sum_{k=1}^{K-1} P(Y = Y_k / X)$$

Here introducing this set of weights will make it redundant.

$$\therefore P(Y = Y_K / X) = \frac{\left(e^{(w_{K0} + \sum_{i=1}^{d} w_{Ki} X_i)}\right)}{\left(1 + \sum_{k=1}^{K-1} e^{(w_{k0} + \sum_{i=1}^{d} w_{ki} X_i)}\right)}$$

d) Classification rule simply picks ~~the~~ label with highest probability

$$y = y_k \,, \quad k^* = \underset{k}{\operatorname{argmax}} \, P\left(y = y_k / X \right)$$