MTH443A: End semester examination
Full Marks: 100

[1] Let Y_0, Y_1, Y_2 be independent and identically distributed $N(0,1)$ random variables and define $X_i = Y_0 + Y_i; i = 1,2.$ (a) Find the proportion of total variation in $\underset{\sim}{X} = (X_1, X_2)^T$ explained by the first principal component derived from the covariance matrix of $\underset{\sim}{X}$. (b) Find the relationship between the first principal component and mean random variable based on X_1 and X_2. (c) If Z_1 denote the first principal component obtained in (a), find $Corr(Z_1, Y_0)$.

6 marks

[2] Let $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_{n_j}$ be n_j p-dimensional feature vectors present in the J^{th} cluster after performing a k-means clustering and define $Q = \sum_{i=1}^{n_j} \sum_{j=1}^{n_j} (\underset{\sim}{x}_i - \underset{\sim}{x}_j)(\underset{\sim}{x}_i - \underset{\sim}{x}_j)^T$. (a) Find the relationship between Q and the sample variance covariance matrix (with divisor n_j) for the J^{th} cluster computed from $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_{n_j}$.
(b) Does there exist any relationship between Q and within-cluster sum of squares for cluster J?

8 marks

[3] Consider the following supermarket transactions data with 6 records

Transaction ID	Items in shopping cart
T_1	Tuna, Tomato, Cheese
T_2	Yoghurt, Tomato, Sausage
T_3	Tuna, Yoghurt, Tomato, Sausage, Cheese
T_4	Yoghurt, Sausage, frozen falafel
T_5	Tuna, Yoghurt, Tomato, Sausage
T_6	frozen falafel, Yoghurt, Tomato, Cheese

Apply apriori algorithm to derive all the association rules satisfying minimum support level of 60% and minimum confidence level 80%; mention the steps where anti-monotone property of apriori algorithm is used in deriving the rule set. List the derived association rules with their corresponding support and confidence.

12 marks

[4] Let $(X_1, \ldots, X_n)$ be a random sample from a population having a mixture density

$$p(x) = \sum_{j=1}^{2} \pi_j \, p(x|\theta_j); \text{ where } 0 < \pi_1, \pi_2 < 1, \pi_1 + \pi_2 = 1;$$

$$p(x|\theta_j) = \begin{cases} \theta_j \, e^{-x\,\theta_j}, & x > 0 \\ 0, & \text{otherwise} \end{cases}, \theta_j > 0, i = 1,2.$$

$\pi_1, \pi_2, \theta_1, \theta_2$ are unknown parameters. Formulate the maximum likelihood method for the parameter estimation under E-M algorithm framework and derive the E-M algorithm update equations for π_js and θ_js for density estimation.

12 marks

[5] Let π_1 and π_2 be 2 p-dimensional populations, $\pi_i \equiv N_p\left(\underline{\mu_i}, \Sigma\right), i = 1,2; \underline{\mu_i} \in \mathbb{R}^p, \Sigma > 0.$

Let $\Delta^2 = \left(\underline{\mu_1} - \underline{\mu_2}\right)' \Sigma^{-1} \left(\underline{\mu_1} - \underline{\mu_2}\right)$ denote the square of Mahalanobis distance between π_1 and π_2 and $J_B = -\log_e\left(\int \ldots \int \left(f(\underline{x}|\pi_1)\, f(\underline{x}|\pi_2)\right)^{1/2} \prod_{i=1}^{p} dx_i\right)$ denote the Bhattacharya distance between π_1 and π_2.

(a) Find the relationship between the J_B and Δ.
(b) Suppose the prior probabilities of the 2 populations are $p(\pi_1) = 1/4$ and $p(\pi_2) = 3/4$. Derive the classification rule corresponding to Bayes classifier.

(c) Suppose the prior probabilities of the 2 populations are $p(\pi_1) = 1/4$ and $p(\pi_2) = 3/4$. Prove or disprove the statement: "The total probability of misclassification (TPM) of FLDF based classifier is given by $\Phi(-\Delta)$, where $\Phi(\,.\,)$ denotes the distribution function of a standard normal distribution".

$$\left[\text{p.d.f. of } \underset{\sim}{X} \sim N_p\left(\underset{\sim}{\mu}, \Sigma\right) \text{ for } \Sigma > 0 \text{ is } f\left(\underset{\sim}{x}\right) = \left((2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}\right)^{-1} e^{-\frac{1}{2}\left(\left(\underset{\sim}{x}-\mu\right)^T \Sigma^{-1}\left(\underset{\sim}{x}-\mu\right)\right)}\right]$$

21 (7+7+7) marks

[6] Consider a 3-class (π_1, π_2 and π_3) classification problem where the class conditional densities are given by: $f_1(x|\pi_1) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$ $f_2(x|\pi_2) = \begin{cases} 2\,e^{-2x}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$ and

$$f_3(x|\pi_3) = \begin{cases} 3\,e^{-3x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The prior probabilities are such that $p(\pi_1) = p(\pi_2) = p(\pi_3)$.

(a) Find the TPM minimizing classification partition.

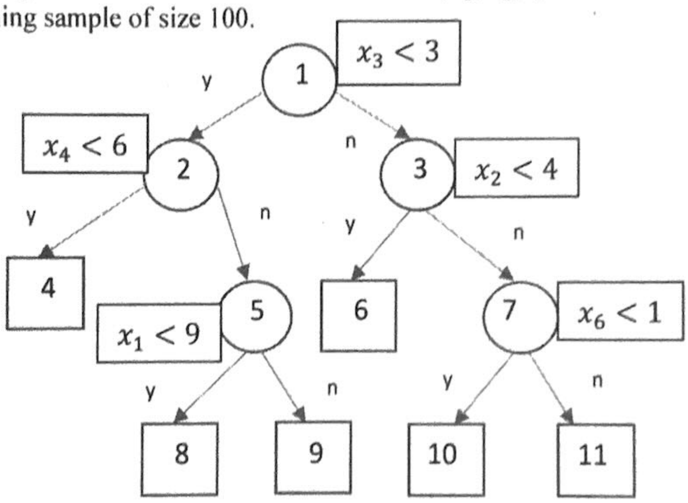
(b) Find $P(\pi_2|\pi_1)$.

(c) Find the TPM of the rule obtained in (a).

($\log_e 2 = 0.693, \log_e 3 = 1.099, \log_e 5 = 1.609, \log_e 7 = 1.946$)

16 (5+3+8) marks

[5] Consider the following classification tree T for a 2-class (π_1, π_2) problem with 6-dimensional feature space, obtained from a learning sample of size 100.



For the constructed tree,

t	1	2	3	4	5	6	7	8	9	10	11
$N(t)$	100	60	40	10	50	16	24	20	30	10	14
$N_2(t)$	60	40	20	0	40	16	4	15	25	1	3

$N(t)$: # of training patterns reaching node t and

$N_2(t)$: # of training patterns with label π_2 reaching node t.

(a) Classify the feature vector $(9,1,0,8,12,2)$ using the above tree.

(b) Find a measure of tree impurity.

(c) Find the strength of the internal node 3.

(d) Find the Gini Index of node 5.

(e) Find a measure of change in impurity function due to split at node 2.

(f) Under a weakest link pruning approach, obtain the first pruned subtree, T_1.

$$\left[\begin{array}{l}\text{Use misclassification error rate at node } t \text{ as it's impurity measure wherever required;} \\ \text{i.e. } Imp(t) = \frac{\sum_{i:\, \underset{\sim}{x}_i \in U(t)} I(y_i \neq \pi_{j(t)})}{N(t)}; \text{ where } j(t) = \underset{i}{\operatorname{argmax}}\, p(\pi_i|t).\end{array}\right]$$

25 (2+4+4+3+4+8) Marks

(1)

(a) $X_1 = Y_0 + Y_1$ $\quad\Sigma_X = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

$X_2 = Y_0 + Y_2$

$$|\Sigma_X - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0 \Rightarrow (2-\lambda)^2 = 1$$

$$\lambda_1 = 3, \ \lambda_2 = 1 \quad 2$$

$\Rightarrow 1^{st}$ PC explains : $\frac{3}{4}$ prop of total variance

(b) $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \underline{x} = 3 \underset{\sim}{x}$

$$\begin{aligned} 2x_1 + x_2 &= 3x_1 \\ x_1 + 2x_2 &= 3x_2 \end{aligned} \Rightarrow x_1 = x_2 \Rightarrow \underset{\sim}{e}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

PC 1 : $\frac{1}{\sqrt{2}}(X_1 + X_2) = \sqrt{2}\ \bar{X}$ w.r.t 2

$$Z_1 = \frac{1}{\sqrt{2}}(2Y_0 + Y_1 + Y_2)$$

(c) $Cov(Z_1, Y_0) = \sqrt{2}$; $V(Z_1) = 3$; $V(Y_0) = 1$

$$\Rightarrow Corr^n(Z_1, Y_0) = \frac{\sqrt{2}}{\sqrt{3}} = \sqrt{\frac{2}{3}} \quad 2$$

~~$\mathcal{S}$~~

(2)

(a) $\mathcal{G} = \sum_{i=1}^{n_J} \sum_{j=1}^{n_J} (\underline{x}_i - \underline{x}_j)(\underline{x}_i - \underline{x}_j)'$

i.e. $\mathcal{G} = 2n_J \left(\sum \underline{x}_i \underline{x}_i' - n_J \bar{\underset{\sim}{x}} \bar{\underset{\sim}{x}}' \right)$

$$S_{n_J} = \frac{1}{n_J} \sum (\underline{x}_i - \bar{\underset{\sim}{x}})(\underline{x}_i - \bar{\underset{\sim}{x}})'$$

$$= \frac{1}{n_J} \left(\sum \underline{x}_i \underline{x}_i' - n_J \bar{\underset{\sim}{x}} \bar{\underset{\sim}{x}}' \right)$$

4

$$\Rightarrow \quad \mathcal{G} = 2 n_J^2 S_{n_J}$$

Within-cluster SS

$$\sum_{i=1}^{n_J} |\underline{x}_i - \bar{\underset{\sim}{x}}|^2$$

$$\frac{\mathcal{G}}{2n_J} = \sum (\underline{x}_i - \bar{\underset{\sim}{x}})(\underline{x}_i - \bar{\underset{\sim}{x}})'$$

$$\frac{1}{2n_J} \operatorname{tr} \mathcal{G} = \sum_{i=1}^{n_J} |\underline{x}_i - \bar{\underset{\sim}{x}}|^2$$

$\Rightarrow \operatorname{tr} \mathcal{G} = 2n_J$ (within cluster SS for cluster J)

4

(3) Let A = Tuna

B = Tomato

C = Cheese

D = Yoghurt

E = Sausage

F = frozen falafel

Step 1 : generation of frequent itemsets (min sup 60%)

C_1

Item	Count
A	3 — ✗ (infreq)
B	5
C	3 — ✗ (infreq)
D	5
E	4
F	2 — ✗ (infreq)

→

L_1

Item	Count
B	5
D	5
E	4

→

C_2

Item	Count
B D	4
B E	3 — ✗
D E	4

using apriori algo

L_2

item	count
B D	4
D E	4

5

No further frequent itemset of higher order can be obtained (apriori algorithm)

Frequent itemsets are thus BD & BE

ARM rule generation : step 2 (min conf 80%)

From BD : $B \Rightarrow D \rightarrow \text{conf} = \frac{4}{5}$ ✓

$D \Rightarrow B \rightarrow \text{conf} = \frac{4}{5}$ ✓

$2\frac{1}{2}$

From DE : $D \Rightarrow E \rightarrow \text{conf} = \frac{4}{5}$ ✓

$E \Rightarrow D \rightarrow \text{conf} = \frac{4}{4}$ ✓

$2\frac{1}{2}$

Final ARM rules

B ⇒ D (support = 80%, conf = 80%)

D ⇒ B (supp = 80%, conf = 80%) 2

D ⇒ E (supp = 80%, conf = 80%)

E ⇒ D (supp = 80%. conf = 100%)

i.e. Tomato ⇒ Yoghurt (80, 80)

Yoghurt ⇒ Tomato (80, 80)

Yoghurt ⇒ Sausage (80, 80)

Sausage ⇒ Yoghurt (80, 100)

(4) $f(x) = \sum_{j=1}^{2} \pi_j f(x \mid \theta_j)$

$$f(x \mid \theta_j) = \begin{cases} \theta_j e^{-\theta_j x}, & x > 0 \\ 0, & o/w \end{cases}$$

$\Phi = (\pi_1, \pi_2, \theta_1, \theta_2)'$

$\underset{\sim}{x}$: incomplete data w/o class labels

$\underset{\sim}{y}' = (\underset{\sim}{x}', \underset{\sim}{z}')$

$\underset{\sim}{Z}$: latent variables

$$g(\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n \mid \Phi) = \prod_{i=1}^{n} \left(\prod_{j=1}^{2} \left(f(x_i \mid \theta_j) \pi_j \right)^{z_{ji}} \right)$$

log likelihood f^n

$$\log g = \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ji} \log f(x_i \mid \theta_j) + \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ji} \log \pi_j$$ (2)

E-step: $\omega_{ji} = E(Z_{ji} \mid x_i, \Phi^{(m)})$

$$= \frac{\pi_j^{(m)} f(x_i \mid \theta_j^{(m)})}{\sum_{j=1}^{2} \pi_j^{(m)} f(x_i \mid \theta_j^{(m)})}$$ — (2)

form the function

$$Q(\Phi, \Phi^{(m)}) = \sum_{i=1}^{n} \sum_{j=1}^{2} \omega_{ji} \log f(x_i \mid \theta_j) + \sum_{i=1}^{n} \sum_{j=1}^{2} \omega_{ji} \log \pi_j$$ (2)

M-step: Maximise Q w.r.t. π_j & θ_j

Maximization of Q w.r.t. π_j subject to $\sum_{j=1}^{2} \pi_j = 1$

gives $\hat{\pi}_j = \sum_{i=1}^{n} \omega_{ji} / n$ (2)

Further,

$$\frac{\partial Q}{\partial \theta_j} = \sum_{i=1}^{n} w_{ji}\left(\frac{1}{\theta_j} - x_i\right)$$

$$\Rightarrow \frac{\partial Q}{\partial \theta_j} = 0 \Rightarrow \sum_{i=1}^{n} \frac{w_{ji}}{\theta_j} = \sum_{i=1}^{n} w_{ji} x_i$$

$$\Rightarrow \quad \hat{\theta}_j = \sum_{i=1}^{n} w_{ji} \Big/ \sum_{i=1}^{n} w_{ji} x_i \quad - (4)$$

(5) $\pi_i \sim N_p(\underline{\mu}_i, \Sigma) \quad i = 1, 2 \; ; \; \Sigma > 0$

(a) $\int \left(f(\underline{x} | \pi_1) f(\underline{x} | \pi_2) \right)^{1/2} d\underline{x}$

$$= \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \int \left[\exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu}_1)' \Sigma^{-1} (\underline{x} - \underline{\mu}_1) - \frac{1}{2}(\underline{x} - \underline{\mu}_2)' \Sigma^{-1} (\underline{x} - \underline{\mu}_2) \right) \right]^{1/2} d\underline{x}$$

$$= \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \int \left[\exp\left(-\left\{ \underline{x}' \Sigma^{-1} \underline{x} - (\underline{\mu}_1 + \underline{\mu}_2)' \Sigma^{-1} \underline{x} + \frac{1}{2} \underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1 + \frac{1}{2} \underline{\mu}_2' \Sigma^{-1} \underline{\mu}_2 \right\} \right) \right]^{1/2} d\underline{x}$$

$$= \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \int \exp\left(-\frac{1}{2} \left\{ \left(\underline{x}' \Sigma^{-1} \underline{x} - 2\left(\frac{\underline{\mu}_1 + \underline{\mu}_2}{2} \right)' \Sigma^{-1} \underline{x} + \left(\frac{\underline{\mu}_1 + \underline{\mu}_2}{2} \right)' \Sigma^{-1} \left(\frac{\underline{\mu}_1 + \underline{\mu}_2}{2} \right) \right) - \left(\frac{\underline{\mu}_1 + \underline{\mu}_2}{2} \right)' \Sigma^{-1} \left(\frac{\underline{\mu}_1 + \underline{\mu}_2}{2} \right) + \frac{1}{2} \underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1 + \frac{1}{2} \underline{\mu}_2' \Sigma^{-1} \underline{\mu}_2 \right\} \right) d\underline{x}$$

$$= \exp\left(-\frac{1}{2} \left(-\frac{1}{4} (\underline{\mu}_1 + \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) + \frac{1}{2} (\underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1 + \underline{\mu}_2' \Sigma^{-1} \underline{\mu}_2) \right) \right)$$

4

$$= \exp\left(-\frac{1}{8} (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \right)$$

$$= \exp\left(-\frac{1}{8} \Delta^2 \right)$$

$$\Rightarrow J_B = -\log\left(e^{-\frac{1}{8}\Delta^2}\right) = \frac{1}{8}\Delta^2$$ 3

i.e. $8 J_B = \Delta^2$

(b) Bayes classifier rule is:

Assign $\underset{\sim}{x}$ to π_K if

$$p_K f_K(\underline{x}) \geq p_i f_i(\underline{x}) \quad \forall i$$

i.e. assign $\underset{\sim}{x}$ to π_1 if

$$p_1 f_1(\underline{x}) > p_2 f_2(\underline{x})$$

i.e. $$\frac{1}{4}\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1)\right)\right)$$

$$> \frac{3}{4}\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2)\right)\right)$$ 2

i.e. $(\underline{x}-\underline{\mu}_1)'\Sigma^{-1}(\underline{x}-\underline{\mu}_1) < (\underset{\sim}{x}-\underline{\mu}_2)'\Sigma^{-1}(\underline{x}-\underline{\mu}_2) + \log 3$

i.e. $\left(\underline{x}'\Sigma^{-1}\underset{\sim}{x} + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - 2\underline{\mu}_1'\Sigma^{-1}\underset{\sim}{x}\right) < \left(\underline{x}'\Sigma^{-1}\underline{x} + \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2 - 2\underline{\mu}_2'\Sigma^{-1}\underline{x}\right) + \log 3$

i.e. $(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} > \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2) + \log 3$

i.e. $R_1^* = \{\underset{\sim}{x} : (\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} > \frac{1}{2}(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1+\underline{\mu}_2) + \log 3\}$

$R_2^* = \{\underline{x} : \quad \leq \quad . \; - \; . \}$

5

(c) FLDF is

assign $\underset{\sim}{x}$ to π_1 if

$$(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{x} > \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2)$$

$$P(1|2) = P_{\pi_2}\left((\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} > \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \right)$$

$$(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} \overset{\pi_2}{\sim} N\left((\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{\mu}_2 , \Delta^2 \right)$$

$$\Rightarrow P(1|2) = P_{\pi_2}\left(Z > \frac{1}{2}\Delta \right) \quad ; \quad Z \sim N(0,1)$$

$$= \cancel{\Phi} = 1 - P\left(Z \le \frac{\Delta}{2} \right) = 1 - \Phi(\Delta/2)$$

$$= \Phi(-\Delta/2)$$ $2\frac{1}{2}$

$$P(2|1) = P_{\pi_1}\left((\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} \le \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 + \underline{\mu}_2) \right)$$

$$(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underset{\sim}{X} \overset{\pi_1}{\sim} N_1\left((\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \underline{\mu}_1 , \Delta^2 \right)$$

$$\Rightarrow P(2|1) = P_{\pi_1}\left(Z \le -\Delta/2 \right) \quad ; \quad Z \sim N(0,1)$$

$$= \Phi(-\Delta/2)$$ $2\frac{1}{2}$

$$\Rightarrow TPM = \frac{1}{4} \Phi(-\Delta/2) + \frac{3}{4} \Phi(-\Delta/2)$$

$$= \Phi(-\Delta/2)$$ 2

(6) $\Pi_1 : f_1(\underline{x}) = e^{-x} \quad x > 0$

$\Pi_2 : f_2(x) = 2e^{-2x} \quad x > 0$

$\Pi_3 : f_3(\underline{x}) = 3e^{-3x} \quad x > 0$

(a)

$R_1^* : \underline{x} \ni f_1(\underline{x}) > f_2(x)$ & $f_1(x) > f_3(x)$

i.e. $\underline{x} \ni x > \log 2$ & $x > \frac{1}{2}\log 3$

$R_2^* : \underline{x} \ni f_2(x) > f_1(x)$ & $f_2(x) > f_3(x)$

i.e. $x \ni x < \log 2$ & $x > \log 3/2$

$R_3^* : \mathcal{X} - R_1^* - R_2^*$

$\Rightarrow R_1^* : x > \log 2$

$R_2^* : \log 3/2 < x < \log 2$

$R_3^* : x < \log 3/2$

5

(b) $P(2|1) = \int_{R_2^*} f_1(x)\,dx = \int_{\log 3/2}^{\log 2} e^{-x}\,dx = \frac{1}{6}$

3

(c)

$TPM = p_1 P(\text{error}|1) + p_2 P(\text{error}|2) + p_3 P(\text{error}|3)$

$P(\text{error}|1) = \int_{R_1^{*c}} f_1(x) = \int_0^{\log 2} e^{-x}\,dx$

$= \frac{1}{2}$

2

$$P(\text{error} \mid 2) = \int_{R_2^{*c}} f_2(x)\,dx$$

$$= 2\int_0^{\log 3/2} e^{-2x} + 2\int_{\log 2}^{\infty} e^{-2x}\,dx$$

$$= \frac{29}{36}$$ 2

$$P(\text{error} \mid 3) = \int_{R_3^{*c}} f_3(x)\,dx = \int_{\log 3/2}^{\infty} 3e^{-3x}\,dx = \frac{8}{27}$$ 2

$$\Rightarrow TPM = \frac{1}{3}\left(\frac{1}{2} + \frac{29}{30} + \frac{8}{27}\right) = 0.53395$$ 2

(7)

(a) $(9, 1, 0, 4, 12, 2)$ reaches

$\Rightarrow$ assignment is π_2 2

(b) Impurity of tree $= \sum_{t \in \tilde{T}} p(t)\, Imp(t)$

$$= \underset{\uparrow \atop 4}{0} + \underset{\uparrow \atop 6}{0} + \underset{\uparrow \atop 8}{\frac{20}{100}} \times \frac{5}{20} + \underset{\uparrow \atop 9}{\frac{30}{100}} \times \frac{5}{30} + \underset{\uparrow \atop 10}{\frac{10}{100}} \times \frac{1}{10} + \underset{\uparrow \atop 11}{\frac{14}{100}} \times \frac{3}{14}$$

4

$$= \frac{14}{100}$$

(c) <u>strength of node 3</u>

$$g(3) = \frac{R(3) - R(T_3)}{|\tilde{T}_3| - 1}$$

$$R(3) = p(3)\, r(3) = \frac{20}{100}$$

$$R(T_3) = \sum_{6,10,11} p(t)\, r(t) = \frac{10}{100} \times \frac{1}{10} + \frac{14}{100} \times \frac{3}{14} = \frac{4}{100}$$

$$\Rightarrow g(3) = \frac{\frac{20}{100} - \frac{4}{100}}{3 - 1} = \frac{8}{100}$$

4

(d) Gini Index of node 5

$$= \sum_{i=1}^{2} \sum_{\substack{j=1 \\ j \neq i}}^{2} p(\pi_i | 5)\, p(\pi_j | 5)$$

$$= 2\, p(\pi_1 | 5)\, p(\pi_2 | 5)$$

$$= 2 \quad \frac{10}{50} \times \frac{40}{50} = \frac{8}{25}$$ 3

(e) $$\mathrm{Imp}(2) = \frac{20}{60}$$

After split at 2, impurity is:

$$p_4\, \mathrm{Imp}(4) + p_5\, \mathrm{Imp}(5) \quad - (*)$$

$$p_4 = \frac{N(4)}{N(2)} = \frac{10}{60} \; ; \; p_5 = \frac{N(5)}{N(2)} = \frac{50}{60}$$

$$(*) = \frac{10}{60} \times 0 + \frac{50}{60} \times \frac{10}{50} = \frac{1}{6}$$

$$\Delta\, \mathrm{Imp} = \frac{2}{6} - \frac{1}{6} = \frac{1}{6}$$ 4

(f) Strengths of internal nodes

$$g(1) = \frac{R(1) - R(T)}{|\tilde{T}| - 1} = \frac{\frac{40}{100} - \frac{14}{100}}{5} = \frac{26}{100} \times \frac{1}{5}$$

$1\frac{1}{2}$

$$g(2) = \frac{R(2) - R(T_2)}{|\tilde{T}_2| - 1}$$

$$= \frac{\frac{20}{100} - \frac{10}{100}}{2} = \frac{5}{100} \quad 1\frac{1}{2}$$

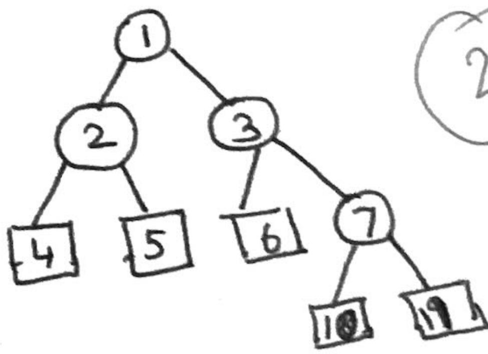
$$g(3) = \frac{16}{100}$$

$$g(5) = \frac{R(5) - R(T_5)}{|\tilde{T}_5| - 1} = \frac{\frac{10}{100} - \frac{10}{100}}{1} = 0 \quad 1\frac{1}{2}$$

$$g(7) = \frac{R(7) - R(T_7)}{|\tilde{T}_7| - 1} = \frac{\frac{4}{100} - \frac{4}{100}}{1} = 0 \quad 1\frac{1}{2}$$

weakest link prunning will be obtained by prunning either at 5 or 7

i.e. T_1 is either



2

or