



B.M.S. COLLEGE OF ENGINEERING, BANGALORE-19		
(Autonomous Institute, Affiliated to VTU)		
Department Name: Computer Science and Engineering		
Course Code: 20CS6PCMAL	Course Title: MACHINE LEARNING	
Semester : 6	Maximum Marks : 40	Date : 07-07-2022
SCHEME AND SOLUTION		
Instructions: Choice is provided in Part C		

PART-A

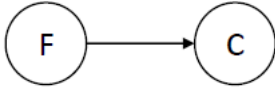
Total 5 Marks (No choice)

Sl No	Question	Marks
1	<p>Bayes optimal Classifier</p> <p>The most probable classification of the new instance is obtained by combining the prediction of all hypothesis, weighted by their posterior probabilities</p> $\operatorname{argmax}_{v_j \in V} \sum_{h_i \in H} P(v_j h_i) P(h_i D)$ <p>Gibbs Algorithm</p> <p>Choose a hypothesis h from H at random, according to the posterior probability distribution over H.</p> <p>Use h to predict the classification of the next instance x.</p> <p>Given a new instance to classify, the Gibbs algorithm simply applies a hypothesis drawn at random according to the current posterior probability distribution. Surprisingly, it can be shown that under certain conditions the expected misclassification error for the Gibbs algorithm is at most twice the expected error of the Bayes optimal classifier</p>	5M

PART-B

Total 5 Marks (No choice)

Sl No	Question	Marks
2a	<p>$P(\text{positive} \text{Covid19}) = 0.99$</p> <p>$P(\text{covid19}) = 0.6$</p> <p>$P(\text{positive}) = 0.598$</p>	5M

	$P(\text{covid19} \text{positive}) = 0.99 \times 0.6 / 0.598$ $= 0.993$																
2b	$17 / 100 = 0.17$ $100 (0.17)(1 - 0.17) = 14.11$ $\text{square root}(14.11) = 3.76$ $3.76 / 100 = 0.0376$ $= \text{standard deviation estimate for } Error_D(h)$ 95% confidence interval for $Error_D(h) =$ $Error_D(h) \pm 1.96 * (\text{square root}[Error_D(h) * (1 - Error_D(h)) / n])$ $= 0.17 \pm 1.96 (\text{square root}[0.17 * (1 - 0.17) / 100])$ $= 0.17 \pm 0.0736$	5M															
2c	<p>Consider the following Bayesian network, where F = having the flu and C = coughing:</p> <p> $P(F) = 0.1$  $P(C F) = 0.8$ $P(C \neg F) = 0.3$ </p> <p>Write down the joint probability table specified by the Bayesian network.</p> <table border="1"> <thead> <tr> <th>F</th><th>C</th><th></th></tr> </thead> <tbody> <tr> <td>t</td><td>t</td><td>$0.1 \times 0.8 = 0.08$</td></tr> <tr> <td>t</td><td>f</td><td>$0.1 \times 0.2 = 0.02$</td></tr> <tr> <td>f</td><td>t</td><td>$0.9 \times 0.3 = 0.27$</td></tr> <tr> <td>f</td><td>f</td><td>$0.9 \times 0.7 = 0.63$</td></tr> </tbody> </table>	F	C		t	t	$0.1 \times 0.8 = 0.08$	t	f	$0.1 \times 0.2 = 0.02$	f	t	$0.9 \times 0.3 = 0.27$	f	f	$0.9 \times 0.7 = 0.63$	5M
F	C																
t	t	$0.1 \times 0.8 = 0.08$															
t	f	$0.1 \times 0.2 = 0.02$															
f	t	$0.9 \times 0.3 = 0.27$															
f	f	$0.9 \times 0.7 = 0.63$															

PART-C

Total 20 Marks (Answer 3 or 4)

Sl No	Question	Marks
3a		10M

	$P(Accident_{yes} A_{new}) = P(Weather\ condition_{rain} Accident_{yes})$ $* P(Road\ condition_{good} Accident_{yes})$ $* P(Traffic\ condition_{normal} Accident_{yes})$ $* P(Engine\ problem_{no} Accident_{yes})$ $* P(Accident_{yes})$ $P(Accident_{no} A_{new}) = P(Weather\ condition_{rain} Accident_{no})$ $* P(Road\ condition_{good} Accident_{no})$ $* P(Traffic\ condition_{normal} Accident_{no})$ $* P(Engine\ problem_{no} Accident_{no})$ $* P(Accident_{no})$ $P(Accident_{yes} A_{new}) = \frac{1}{5} * \frac{1}{5} * \frac{1}{5} * \frac{2}{5} * \frac{5}{10} = \frac{10}{6250}$ $P(Accident_{no} A_{new}) = \frac{2}{5} * \frac{3}{5} * \frac{2}{5} * \frac{4}{5} * \frac{5}{10} = \frac{240}{6250}$ <p>Since $P(Accident_{no} A_{new}) > P(Accident_{yes} A_{new})$ the prediction is Accident='no'</p>	
	(OR)	
3b	<ul style="list-style-type: none"> The learner considers some set of candidate hypotheses H and it is interested in finding the most probable hypothesis $h \in H$ given the observed data D Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis h_{MAP}. We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis. $h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h D)$ $= \underset{h \in H}{\operatorname{argmax}} \frac{P(D h) P(h)}{P(D)}$ $= \underset{h \in H}{\operatorname{argmax}} P(D h) P(h)$ <p>Brute – Force MAP Learning Algorithm</p>	10M

	<p>For each hypothesis h in H, calculate the posterior probability</p> $P(h D) = \frac{P(D h)P(h)}{P(D)}$ <p>Output the hypothesis h_{MAP} with the highest posterior probability</p> $h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h D)$ <ul style="list-style-type: none"> • BF MAP learning algorithm must specify values for $P(h)$ and $P(D h)$. • $P(h)$ and $P(D h)$ must be chosen to be consistent with the assumptions: <ol style="list-style-type: none"> 1. The training data D is noise free (i.e., $d_i = c(x_i)$). 2. The target concept c is contained in the hypothesis space H 3. We have no a priori reason to believe that any hypothesis is more probable than any other. <p>With these assumptions:</p> $P(h) = \frac{1}{ H } \quad \text{for all } h \text{ in } H$ $P(D h) = \begin{cases} 1 & \text{if } d_i = h(x_i) \text{ for all } d_i \text{ in } D \\ 0 & \text{otherwise} \end{cases}$ <p>So, the values of $P(h D)$ will be:</p> $P(h D) = \frac{0 \cdot P(h)}{P(D)} = 0 \quad \text{if } h \text{ is inconsistent with } D$ $P(h D) = \frac{1 \cdot \frac{1}{ H }}{\frac{ VS_{H,D} }{ H }} = \frac{1}{ VS_{H,D} } \quad \text{if } h \text{ is consistent with } D$ $P(D) = \sum_{h_i \in H} P(D h_i)P(h_i) = \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{ H } + \sum_{h_i \notin VS_{H,D}} 0 \cdot \frac{1}{ H }$ $= \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{ H } = \frac{ VS_{H,D} }{ H }$	
4a	<ul style="list-style-type: none"> • We have to find the difference d between true errors of h_1 and h_2 hypotheses $d \equiv \operatorname{error}_{\mathcal{D}}(h_1) - \operatorname{error}_{\mathcal{D}}(h_2)$ <ul style="list-style-type: none"> • The estimator of the true error is the sample errors (\hat{a}) 	10M

	$\hat{d} \equiv error_{s_1}(h_1) - error_{s_2}(h_2)$ <ul style="list-style-type: none"> Since the sample errors of h_1 and h_2 follows normal distribution <p>$error_{s_1}(h_1)$, $error_{s_2}(h_2)$, the difference of errors between h_1 and h_2 also follow normal distribution</p> <ul style="list-style-type: none"> Variance of this distribution is the sum of the variances of $error_{s_1}(h_1)$, $error_{s_2}(h_2)$ to obtain the approximate variance of each of these distribution $\sigma_d^2 \approx \frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2}$ <p>Normal distribution with mean d and variance σ^2, the $N\%$ confidence interval estimate for d is $\hat{d} \pm z_N \sigma$. Using the approximate variance σ_d^2 given above, this approximate $N\%$ confidence interval estimate for d is</p> $\hat{d} \pm z_N \sqrt{\frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2}}$	
	(OR)	
4b	<hr/> <ul style="list-style-type: none"> Each tuple can be represented as set of attribute pairs as $E = \{ \{x_1=\text{sunny}, x_2=\text{hot}, x_3=\text{high}, x_4=\text{weak}\}, \{x_1=\text{sunny}, x_2=\text{hot}, x_3=\text{high}, x_4=\text{strong}\}, \dots \}$ Total number of different attribute values? are 10 Each tuple have 4 attribute that could be chosen in at most ${}^{10}C_4$ ways =210 Class attributes are two so total number of ways are 210×2 	10M

- Code length in bits $\log(210 \times 2) = 8.715$
- Code length of whole database $L(E) = 8.715 \times 14 = 122.01$

Consider two Hypothesis

H1: [play=yes]

If { *outlook*=overcast}
If { *humidity*=normal, *wind*=weak}

H2: [play=yes]

If { *outlook*=overcast}
If { *humidity*=normal, *wind*=weak}
If { *temperature*=mild, *humidity*=normal}

- Length of Hypothesis $L(H1) = \log({}^{10}C_1) + \log({}^{10}C_2) = 8.81$
- Length of Hypothesis
 $L(H2) = \log({}^{10}C_1) + \log({}^{10}C_2) + \log({}^{10}C_2) = 14.30$

H1

Actual/Predicted	Yes	No
Yes	7	2
No	0	5

H2

Actual/Predicted	Yes	No
Yes	8	1
No	0	5

- $L(E/H1) = \log({}^7C_0) + \log({}^7C_2) = 4.39$
- $L(E/H2) = \log({}^8C_0) + \log({}^6C_1) = 2.59$
- Compression of H1: $L(E) - L(H1) - L(E/H1) = 108.81$
- Compression of H2: $L(E) - L(H2) - L(E/H2) = 105.12$

Result shows that H1 has better compression than H2.