SML 2025, Monsoon, EndSem, Total marks 20

Note:

- Symbols have their usual meanings. Duration: 2 hours. Number in [.] indicate marks. [COx] indicates the question is mapped to the respective course outcome.
- For MCQ, each question may have more than one correct answer. Select all correct options. Each MCQ carries 1.5 marks.

Q1. Bagging (Bootstrap Aggregating) mainly aims to: [CO1]

(a) Bagging, in general, should perform better than a single tree

(b) Reduce bias of the model

(c) Reduce variance of the model

(d) Increase both bias and variance of the model

answer. (a) and (c)

Q2. In Random Forests, which of the following techniques are used? [CO1]

(a) Update the weights of the incorrectly classified samples

(b) Not all features are selected at each split

(c) Boosting of weak learners

(d) Bootstrap sampling of data

answer. (b) and (d)

Q3. Boosting algorithms generally: [CO1]

(a) Focus more on correctly classified points at each step

(b) Assign higher weights to misclassified samples

(c) Combine weak learners sequentially

(d) Require weak learners to have high bias and low variance

answer. (b), (c) and (d)

Q4. Regarding Bias-Variance Tradeoff, which of the following are TRUE? [CO3]

(a) High-bias models are prone to overfitting.

(b) High-variance models are prone to underfitting.

(c) Increasing model complexity always reduces both bias and variance.

(d) Regularization methods (like l_2) can help control model complexity.

answer. d

Q5. In Fisher Discriminant Analysis (FDA), which of the following statements are correct? [CO2]

(a) FDA seeks a projection that maximizes between-class variance and minimizes within-class variance.

(b) FDA is equivalent to PCA when classes are well-separated.

(c) FDA uses the generalized Rayleigh quotient for optimization.

(d) FDA can only be used when class covariances are unequal.

answer. a, c

Q6. Which of the statement(s) is/are false for Maximum Likelihood Estimation (MLE)? [CO1]

(a) MLE finds parameters that minimize the likelihood of observed data.

(b) MLE always requires specifying prior distributions.

(c) MLE is always unbiased.

answer. all correct

Q7. Suppose for a binary classification task, there are two Rosenblatt' perceptrons to be used. To classify a point x_i, the decision rule is to compute "sign of the summation of distances of x_i from each perceptron' decision boundary". Suppose L denotes loss of Rosenblatt' perceptron. Now, as there are two Rosenblatt' perceptron, how does the loss change? Using the modified loss, find the update rule for one of the perceptrons. [2][CO1]

Ans. Since decision rule is sign of dsitance, Loss for perceptron $L = -y(\beta^T x + \beta_0)$.
With two perceptrons $L = -y(\beta^T x + \beta_0) - y(\alpha^T x + \alpha_0) = -y(\beta'^T x + \beta'_0)$
This is in standard form of Ronseblatt' perceptron
$\beta \leftarrow \beta + \eta y x$ as $dL/d\beta = -yx$
$\beta_0 \leftarrow \beta_0 + \eta x$

Q8. The idea of PCA is to find an orthogonal bases and project the data onto such bases such that the projected data preserves maximum variance. Let the data be $X = [X_1\ X_2]$, where $X_1 \in \mathcal{R}^{d\times n_1}$, $X_2 \in \mathcal{R}^{d\times n_2}$ and $X \in \mathcal{R}^{d\times(n_1+n_2)}$. X_1 denotes data from class 1 and X_2 denotes data from class 2. Suppose we apply PCA on X with an additional constraint - projected data must also maximize the absolute difference between means of the projected classes. That is, if M_1 and M_2 are the respective means of projected classes, then in addition to PCA objective, $|M_1 - M_2|$ must also be maximized. Solve for the first principal component vector $U \in \mathcal{R}^d$. While there may not be a closed form, you must still give an expression to compute U using a known form. [2] [CO2]

ans. PCA objective is $u'\Sigma u\ st. u'u = 1$. We need to incorporate constraint to maximize absolute distance of projected class means.
$u'\Sigma u + \beta|u'(\mu_1 - \mu_2)|\ st. u'u = 1$
where $\mu_i = \frac{1}{n_i} u' \sum_{j=1}^{n_i} x^j$, where x^j denotes j^{th} sample of X_1 or X_2.
We can use gradient descent to compute u where gradient wrt u is $2\Sigma u + \beta(\mu_1 - \mu_2) sign(u'(\mu_1 - \mu_2))$.
After each gradient step, we project u to unit l_2 ball, that is, normalize u to unit vector.

Q9. Consider a two-category classification problem. The likelihoods for both categories are multivariate Gaussian with the same covariance matrix but different means. Class 1 mean: $\mu_1 = [1\ 0\ 0]^T$, and, Class 2 mean: $\mu_2 = [0\ 1\ 0]^T$. Covariance matrix (common for both classes) is

$$\Sigma = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A test sample is given by $\boldsymbol{X} = [.5\ .25\ 1]^T$. Find the class of $\boldsymbol{X}$ using the discriminant function. The prior probability $P(\omega_1) = 1/3$. [2] [CO3]

The discriminant function for Gaussian class-conditional distributions with equal covariance is:

$$g_i(x) = x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(\omega_i)$$

$$\Sigma^{-1} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & 0 \\ -\frac{1}{3} & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Sigma^{-1}\mu_1 = \begin{bmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ 0 \end{bmatrix}$$

$$x^T\Sigma^{-1}\mu_1 = [0.5, 0.25, 1] \cdot \begin{bmatrix} \frac{2}{3} \\ -\frac{1}{3} \\ 0 \end{bmatrix} = 0.5 \cdot \frac{2}{3} + 0.25 \cdot (-\frac{1}{3}) = \frac{1}{3} - \frac{1}{12} = \frac{1}{4}$$

$$\mu_1^T\Sigma^{-1}\mu_1 = \frac{2}{3}$$

$$\log P(\omega_1) = \log\frac{1}{3} = -\log 3$$

$$g_1(x) = \frac{1}{4} - \frac{1}{2}\cdot\frac{2}{3} - \log 3 = -\frac{1}{12} - \log 3$$

$$\Sigma^{-1}\mu_2 = \begin{bmatrix} -\frac{1}{3} \\ \frac{2}{3} \\ 0 \end{bmatrix}$$

$$x^T\Sigma^{-1}\mu_2 = 0.5 \cdot (-\frac{1}{3}) + 0.25 \cdot \frac{2}{3} = -\frac{1}{6} + \frac{1}{6} = 0$$

$$\mu_2^T\Sigma^{-1}\mu_2 = \frac{2}{3}$$

$$\log P(\omega_2) = \log\frac{2}{3} = \log 2 - \log 3$$

$$g_2(x) = 0 - \frac{1}{2}\cdot\frac{2}{3} + \log 2 - \log 3 = -\frac{1}{3} + \log 2 - \log 3$$

$$g_1(x) \approx -\frac{1}{12} - \log 3 \approx -0.0833 - 1.0986 = -1.1819$$

$$g_2(x) \approx -\frac{1}{3} + \log 2 - \log 3 \approx -0.333 + 0.693 - 1.098 = -0.738$$

Since $g_2(x) > g_1(x)$, the classifier assigns x to class 2.

Q10. Being an ML enthusiast interested in applying theory to practice (say who will win IPL), you find a niche area of predicting whether a team will win or lose a match. Based on domain knowledge, you hypothesize that a team's probability of winning depends on three binary independent features:

- Past record p (1 = good, 0 = poor)
- Current record c (1 = good, 0 = poor)
- Health of players h (1 = good, 0 = poor)

Each of p, c, h is independently distributed and follows a Bernoulli distribution.
Define:

$$\theta_1 = \Pr(p = 1), \quad \theta_2 = \Pr(c = 1), \quad \theta_3 = \Pr(h = 1)$$

where $\theta_1, \theta_2, \theta_3$ are unknown parameters.

Suppose you collect survey responses from n independent individuals (denoted $S_1, S_2, \ldots, S_n$), each recording the corresponding values of (p, c, h). An excerpt of the responses is shown below:

Using only the three rows S-1, S-2, and S-i, estimate $\theta_1, \theta_2, \theta_3$ and compute the win probability Pr(Win) for a team with $p = 0, c = 1, h = 1$ using your estimates.[2] [CO1]

ans. Using MLE, each $\hat{\theta}_j$ is the average of observations.

$\hat{\theta}_1 = 2/3 = \hat{\theta}_3$, $\hat{\theta}_2 = 1/3$

Table 1: Survey responses

Response	p	c	h
S-1	1	0	0
S-2	1	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
S-i	0	1	1

$\Pr(\mathrm{X}) = \theta_1^p(1-\theta_1)^{1-p}\theta_2^c(1-\theta_2)^{1-c}\theta_3^h(1-\theta_3)^{1-h}$

$= (2/3)(1/3)(2/3).$

Q11. Derive the Adaboost algorithm using an exponential loss function. You must clearly derive the update rule for weights of the samples and coefficients of the classifiers. [3][CO3]