Indian Institute of Technology (IIT) Roorkee
Department of Electrical Engineering
EEC-351: Fundamentals of AI/ML
Mid-Term Examination (Autumn 2025–26)

Set-A

Timing: 90 Mins | Date: Sep 15, 2025 Time: 2:00 PM - 3:45 PM | Max. Marks: 50

Instructions

- Answer all questions. Each question must begin on a **new page**.
- Write **Set Name** on **Top Right Corner of Main Answer Sheet**, to avoid any penalty.
- Provide complete justifications for your answers. Partial/unjustified answers may not receive full credit.
- Mention assumptions and conditions clearly wherever necessary.

---

1. Let x is some integer in the set $X = \{1, 2, \ldots, 125, 126, 127\}$, and where each hypothesis $h \in \mathcal{H}$ is an interval of the form $a \leq x \leq b$, with a and b as any integers between 1 and 127 (inclusive), so long as $a \leq b$. A hypothesis $a \leq x \leq b$ labels instance x positive if x falls into the interval defined by a and b, and labels the instance negative otherwise.
   (a) How many distinct hypotheses are there in such $\mathcal{H}$?. (No explanation required) **[1]**
   (b) Suppose we draw N independent examples uniformly from X with $\mathcal{H}$ hypothesis space. Using Hoeffding's inequality: If $\epsilon = 0.05$, calculate the minimum number of samples N needed to ensure that the confidence is at least 95%. **[1]**

2. (a) The VC dimension depends on the input space as well as $\mathcal{H}$. For a fixed $\mathcal{H}$, consider two input spaces $\mathcal{X}_1 \subseteq \mathcal{X}_2$. Show that the VC dimension of $\mathcal{H}$ with respect to input space $\mathcal{X}_1$ is at most the VC dimension of $\mathcal{H}$ with respect to input space $\mathcal{X}_2$. **[1]**
   (b) The monotonically increasing hypothesis set is $\mathcal{H} = \{h \mid x_1 \geq x_2 \;\; \Rightarrow \;\; h(x_1) \geq h(x_2)\}$, where $x_1 \geq x_2$ if and only if the inequality is satisfied for every component. Give an example of a monotonic classifier in two dimensions, clearly showing the $+1$ and -1 regions. (Just show labeled diagram. Any sentence will result in [**-1**]). **[1]**
   (c) Consider a model trained on a hypothesis set $\mathcal{H}$ of 5-dimensional perceptron using a training set and later tested on an independent test set. Model achieved training error $E_{train} = 0.10$ on $N = 200$ training samples and test error $E_{test} = 0.15$ on $N_{test} = 100$ test samples. Using Hoeffding bound, compute the tightest bound on the E_{out} with at least 95% confidence. **[2]**

3. (a) Show that if $\mathcal{H}$ is closed under linear combination (any linear combination of hypotheses in $\mathcal{H}$ is also a hypothesis in $\mathcal{H}$) then $\bar{g} \in \mathcal{H}$. **[2]**
   (b) Give an example of $\mathcal{H}$ (any type) for which the expected final hypothesis function $\bar{g} \notin \mathcal{H}$. **[2]**

4. Suppose a random variable X has the Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, and its probability density function (PDF) is given by:

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

   where, $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the Beta function, $a, b \in \mathbb{Z}^+$ and the Gamma function is $\Gamma(N) = (N-1)!$
   (a) Compute the mean $\mathbb{E}[X]$ and variance $\mathrm{Var}(X)$. **[1+2]**
   (b) Let N samples $\{X_i\}_{i=1}^N$ are independently drawn from $f_X(x)$, and $\mu = \frac{1}{N}\sum_{i=1}^N X_i$
      A. Using **Chebyshev's inequality**, provide an upper bound for bad event: $P\Big(|\mu - \mathbb{E}[X]| > \epsilon\Big)$. Express in terms of α, β, N, and ϵ. **[2]**
      B. If **Hoeffding's inequality** can be applied on the random variable X, bound the bad event probability. **[2]**
      C. Consider that if Hoeffding's inequality is not applicable, it results in confidence as infinity. Now, for $\alpha = 3$, $\beta = 5$, $N = 50$, and $\epsilon = 0.1$, **compute the Chebyshev and Hoeffding bounds**. Also, **show the condition** on N under which Hoeffding's bound becomes tighter than Chebyshev's bound (if at all) for $\alpha = 3$, $\beta = 5$, and $\epsilon = 0.1$. In such a case, **approximate the value** of N where Hoeffding starts outperforming Chebyshev or vice-versa. **[1+1+1]**

5. The expected value of $E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_x\left[\left(g^{(\mathcal{D})}(x) - y(x)\right)^2\right]$ over training data can be decomposed into bias and variance as

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \text{bias} + \text{var}.$$

Now assume, there is noise in the data, $y(x) = f(x) + \varepsilon$, then

(a) If $\varepsilon = \mathcal{N}(0, \sigma^2)$ is data-independent, derive bias-variance decomposition of $\mathbb{E}_{\mathcal{D}}\big[E_{\text{out}}(g^{(\mathcal{D})})\big]$. **[2]**

(b) If $y(x) = f(x) + \varepsilon(\mathcal{D}, x)$, where the data-dependent error has the form

$$\varepsilon(\mathcal{D}, x) = \mu(x) + \lambda(x)\big(g^{\mathcal{D}}(x) - \bar{g}(x)\big) + \eta(x),$$

where $\bar{g}(x)$ is average predictor, $\mu(x)$ is a deterministic shift (bias in the labels), $\lambda(x)$ a coupling factor that relates dataset noise to model fluctuations, and $\eta(x)$ an independent zero-mean noise with variance $\text{Var}[\eta(x)] = \tau^2(x)$. Compute the simplified expression of bias-variance decomposition for a fixed x. **[4]**

6. Consider a simplified learning scenario. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval $[-1, 1]$. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$ (You Just Got Lucky to Know $f(\cdot)$ This Time). Thus, the full data set is $\mathcal{D} = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm returns the line fitting these two points as $g(\cdot)$ (the hypothesis set $\mathcal{H}$ consists of functions of the form $h(x) = ax + b$).

(a) In the above given scenario, give analytical expression for the function $\bar{g}(x)$. **[3]**

(b) Compute analytically what E_{out}, bias and var should be. **[3]**

7. Let $H \in \mathbb{R}^{n\times n}$ be an *idempotent* matrix, i.e. $H^2 = H$ (no symmetry assumed).

(a) Prove that every eigenvalue λ of H satisfies $\lambda \in \{0, 1\}$. **[1]**

(b) Consider H is diagonalizable over $\mathbb{R}$ then show $\text{Tr}(H) = \text{rank}(H)$ is True or False. **[2]**

Hint: Recall the rank–nullity theorem: for $A \in \mathbb{R}^{n\times n}$, $\text{rank}(A) + \text{nullity}(A) = n$. If A has $k \leq n$ zero eigenvalues, then they are associated with eigenvectors that span the null space of A.

8. $L_{sq-sq}(\cdot)$ pins correct solution energy to zero and incorrect to above margin m. What properties should the energy function $E(\cdot)$ have have in order for L_{sq-sq} to be effectively applied, and why? (Mathematical expression of property and 2-3 lines of reasoning at max. Extra text will result in [**-1**]) **[2]**

9. You are training a binary classifier to predict whether a customer will leave a service (churn). However, the dataset labels are imperfect: a label of 1 indicates the customer called support, while 0 indicates they did not. This creates misalignment, since some customers who churned never called, and some who called did not churn.

(a) Let the true churn label be $y^\star \in \{0, 1\}$, but suppose you only observe:

$$y = y^\star \oplus \varepsilon \text{ with } \varepsilon \sim \text{Bernoulli}(p)$$

, where, $\oplus$ denotes XOR. Derive the ***expected cross-entropy loss*** that the model is effectively minimizing under this noise process. How does the noise level p influence the model's learned decision boundary and plot p v/s loss curve? **[4+2]**

(b) Propose and justify a method to mitigate the effect of label noise so that the model's predictions align more closely with true churn. (Only theory will result in [**-1**] marks.) **[2]**

10. An energy-based model assigns an energy $E_\theta(x, y) \in \mathbb{R}$ to each input x and label $y \in \{1, \ldots, K\}$. Consider the mixed loss, as convex combination of two terms with $\alpha \in [0, 1]$, $m > 0$ is a margin, and $[z]_+ := \max(0, z)$:

$$\mathcal{L}_{\mathbf{351}}(\theta; \mathbf{x}, \mathbf{y}) = \alpha \log\left(\sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{K}} \mathbf{e}^{-\mathbf{E}_\theta(\mathbf{x},\mathbf{j})}\right) + (\mathbf{1} - \alpha)\left[\mathbf{m} + \mathbf{E}_\theta(\mathbf{x}, \mathbf{y}) - \min_{\mathbf{i}\neq\mathbf{y}} \mathbf{E}_\theta(\mathbf{x}, \mathbf{i})\right]_+$$

Determine whether $\mathcal{L}_{351}(\theta; x, y)$ is consistent with the general philosophy of loss functions for energy-based models, i.e., whether minimizing this loss lowers the energy of the correct label and raises the energies of incorrect labels. Justify your answer with α range for appropriate operation of $\mathcal{L}_{351}$. **[5]**

End of Paper