भारतीय प्रौद्योगिकी संस्थान जम्मू
INDIAN INSTITUTE OF TECHNOLOGY JAMMU

Discipline of
Computer Science & Engineering

Assignment 1 - Executive M.Tech(AI)
Course: DL **Due Date: 20/05/2025**

Instructions

1. Assignment submissions will be accepted only via Google Classroom. Submissions through email or any other methods will NOT be accepted. Please join Google Classroom using the following link: https://shorturl.at/KEOdT
2. The submission deadline is 20/05/2025. Please submit a single pdf file .
3. Please read the assignment policy (uploaded on Google Classroom) carefully and follow all guidelines.

---

1. Let $\mathcal{H}$ be a hypothesis class of classifiers over $\mathbb{R}^2$ defined as follows:

   Each $h \in \mathcal{H}$ labels a point $(x, y) \in \mathbb{R}^2$ as positive if and only if it lies between two parallel lines. That is, for every $h \in \mathcal{H}$, there exists a direction vector $\vec{v} \in \mathbb{R}^2$ and two real numbers $a < b$ such that:

   $$h(x, y) = \begin{cases} 1 & \text{if } a < \vec{v} \cdot (x, y) < b \\ 0 & \text{otherwise} \end{cases}$$

   What is the VC dimension of this hypothesis class $\mathcal{H}$? Justify your answer.

2. Let $\mathcal{H}$ be the class of all linear classifiers (i.e., hyperplanes) in $\mathbb{R}^3$. Each hypothesis corresponds to a function:

   $$h(x) = \text{sign}(w \cdot x + b)$$

   where $w \in \mathbb{R}^3$, $b \in \mathbb{R}$.

   Determine the VC dimension of $\mathcal{H}$. Can it shatter 4, 5, or more points in $\mathbb{R}^3$? Justify your answer.

3. You are working on a wearable device that uses accelerometer data to detect whether a person is walking, running, or standing still.

   To simplify, you decide to classify each 1-second window of sensor readings as either "active" (walking or running) or "inactive" (standing still), based on features extracted from the sensor (e.g., average acceleration, variance, peak frequency, etc.).

   You plan to use a linear classifier on 5-dimensional feature vectors (i.e., each 1-second window is represented by a point in $\mathbb{R}^5$).

   (a) What is the VC dimension of your model?

   (b) Based on your answer to (a), what is the minimum number of training examples you would need to guarantee low generalization error (e.g., using the standard VC-based generalization bounds)?

4. You are building a machine learning model using a dataset of 60,000 examples. You split it into 70% training, 15% validation, and 15% test.

   After training several models and tuning hyperparameters using the validation set, you pick the best one and evaluate its accuracy on the test set.

   Later, you try a few more model variations and compare them using the test set again — finally reporting the best test set accuracy in your paper.

   (a) At first glance, your approach seems correct: training on training data, tuning on validation, and evaluating on test. But explain what mistake is being made here, and why it can hurt generalization.

   (b) What is the correct role of the test set in a machine learning workflow?

   (c) How can you modify your process to select the best model architecture while keeping the test set "untouched" for final evaluation?

5. A student trains different models on the same dataset and observes the following behaviors:

   - Model A has low training error but very high test error.
   - Model B has both high training error and high test error.
   - Model C has slightly higher training error than Model A but much lower test error.

   (a) For each model (A, B, and C), describe whether it suffers from high bias, high variance, or both. Justify your answers.

(b) Explain the bias-variance tradeoff in your own words. Why is it important to balance these two in machine learning?

(c) Suppose you increase the model complexity (e.g., move from linear regression to a 10th-degree polynomial regression). What is the expected effect on bias and variance? Explain with reasoning.

6. Consider the role of activation functions in neural networks.

(a) Why are activation functions necessary in neural networks? What would happen if we removed them from all layers?

(b) Compare the following activation functions in terms of their mathematical form and behavior:

- ReLU (Rectified Linear Unit)
- Sigmoid
- Tanh

(c) ReLU is widely used in deep learning models. What are two key advantages of ReLU over sigmoid and tanh?

(d) What is the "dying ReLU" problem, and how can it affect training? Mention one possible solution.

7. A neural network is trained for binary classification using a sigmoid activation in the output layer and mean squared error (MSE) as the loss function.

(a) Why might this choice of loss function lead to slow or unstable training, even though it seems reasonable at first?

(b) What would be a better choice of loss function for this setup, and why?

8. Let the loss function L be defined in terms of inputs x and y as follows:

$$a = x^2, \quad b = \sin(y), \quad c = a + b, \quad d = \ln(c), \quad L = d^2$$

(a) Draw the computational graph representing this sequence of operations. Clearly label each node with the operation and the variables.

(b) Identify all paths through which gradients must flow to compute $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial y}$.

(c) Based on your graph, explain briefly how the chain rule would be applied during backpropagation.

9. Consider the following computational graph:

   Let $z = x \cdot y$, $a = z + y$, and $L = a^2$, where $x = 2$, $y = 3$.

   (a) Compute the forward pass to find the value of L.

   (b) Compute the gradients $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial y}$ using the chain rule.

   (c) Suppose a student incorrectly treats the shared variable y as two independent copies in the graph. What gradient error will they likely make in $\frac{\partial L}{\partial y}$? Why is this incorrect?