Minor-1

CSL2050 - Pattern Recognition and Machine Learning

NOTE:

1. This is Question-cum-Answer sheet. Maximum Points: 40, Total Questions: 6, Total Page: 4, Total Time: 1 Hour. **If there is anything not clear in the problems, go ahead with your own assumptions but state them clearly. No doubts will be entertained during the exam.** Be precise and write the answer in the box provided. Verbosity will be penalized. Use the other answer sheet for rough work and submit both.

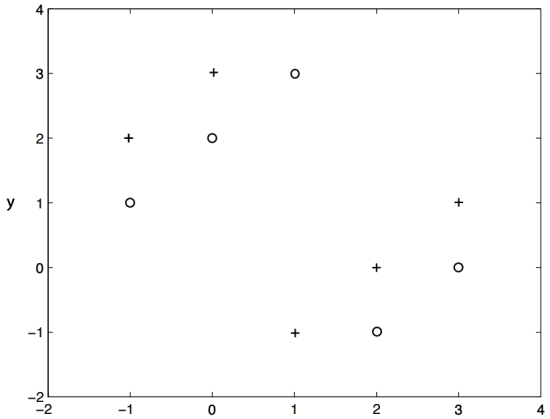
- Name: Roll Number: Signature:

---

1. Suppose you have a dataset with two features, X1 and X2, and you want to create a linear decision boundary to classify two classes, Class A (+ve) and Class B (-ve). The equation for the decision boundary is given as: $2X_1 - 3X_2 + 5 = 0$
   (2 points) (a) Draw the decision boundary in the 2D plane.

   (3 points) (b) Given a new data point with features (X1 = 4, X2 = 2), predict which class it belongs to based on the decision boundary.

2. Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and ∘).



(2 points) (a) What is the leave one out cross validation error when using 1-NN?

(3 points)(b) Which of the following values of k leads to the minimum number of validation errors: 3, 5 or 9? What is the error for that k?

3. The following data set will be used to learn a decision tree for predicting whether students are lazy (L) or diligent (D) based on their weight (Normal or Underweight), their eye color (Amber or Violet), and the number of study hours they have (2 or 3 or 4). The following numbers may be helpful as you answer this problem without using a calculator: $log_2 0.1 = -3.32, log_2 0.2 = -2.32, log_2 0.3 = -1.73, log_2 0.4 = -1.32, log_2 0.5 = -1$. **You don't need to show the derivation for your answers in this problem.**

Weight	Eye Color	Study Hours	Output
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

(a) **(3 points)** What is the conditional entropy $H(EyeColor|Weight = N)$?

(b) **(2 points)** What attribute would the ID3 algorithm choose to choose the root of the tree?

(c) **(3 points)** Draw full decision tree for this data (no pruning).

(d) **(2 points)** What is training error on this unpruned tree?

4. Define the following terms: (2 points each) (i) Machine Learning

(ii) Overfitting

(iii) Gini Impurity

(iv) Confusion Matrix

(v) Boosting

5. Consider fitting the linear regression model to the following data: (-1,1), (0,-1), (2,1)

(1.5 points) (a) Assume $\hat{y}_i = b + \eta_i$ where η_i is Gaussian noise.

(1.5 points) (b) Assume $\hat{y}_i = mx_i + \eta_i$ where η_i is Gaussian noise.

(2 points) (c) Assume $\hat{y}_i = mx_i + b + \eta_i$ where η_i is Gaussian noise.

6. **(TRUE/FALSE: 1 point each)**

(a) K-NN is an unsupervised Machine Learning technique.

(b) Gradient Descent ensures convergence to the global minima for convex functions, regardless of the selected learning rate.

(c) Having a well-defined performance metric is essential for measuring the progress of an ML task.

(d) For English Handwritten Digit classification, the confusion matrix size will be of size 2×2.

(e) In terms of storage, a KNN generally requires less memory compared to the Decision Tree.