Handwritten Assignment-1

CSL2050 - Pattern Recognition and Machine Learning

NOTE:

1. This Handwritten Assignment contains seven problems. Total points for this assignment is 50. Please use A4 sheets and neatly solve all the problems. Be precise, verbosity will be penalized.

2. **(IMPORTANT)** Please write the following pledge in your handwriting with your signature on the first page of your assignment sheet (without this, the assignment will not be considered for evaluation):

   Honesty Pledge:

   "I affirm that this assignment is solely my work. I have not used unauthorized assistance, engaged in plagiarism, or violated ethical standards. Further, all references used and any discussion with anyone have been appropriately cited. Any breach may lead to disciplinary actions as per the course academic honesty policy discussed in Lecture-1."

3. **Deadline:** Feb 7, 2024, 9 AM before the class.

4. **Late Submission Policy:** Late submissions beyond the due date will incur a 10% penalty for each day. Plan the submission ahead and do not wait until the last minute.

---

1. **(6 points)** Given one-dimensional training points and their binary labels: (1,+1), (2,+1), (4,-1), (5,+1), (6,-1), (7,-1).

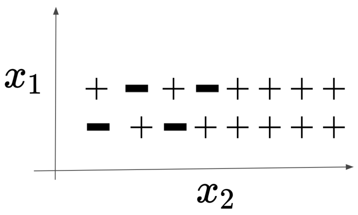
   (a) Show the points using appropriate markers such as $\oplus$ for positive samples and $\ominus$ for negative samples. Further, show the regions where test samples will be classified as positive if 1-NN is used as classification. (Assume test samples can be between 0 and 7 inclusive). (b) Compute training accuracy in percentage if k=3 is used in k-NN.

2. **(4 points)** Consider an image retrieval engine. For a query – *Query-1* it retrieves 10 images as follows in rank 1 to 10: +1, +1, -1, -1, +1, +1, +1, -1, -1, -1 where +1 and -1 denote the retrieved image is relevant and irrelevant to the query respectively. Assuming there were 10 relevant images with respect to *Query-1* in the database, find out: (A) Precision (B) Recall

3. **(4 points)** Suppose in population of 100 students, RTPCR tested students s_1 and s_2 as positive and remaining as negative. But, students $\{s_1, s_2, s_5, s_9, s_{10}, s_{11}, s_{100}\}$ were actually positive. Compute True Positive, False Positive, True Negative, False Negative, Precision, Recall, and F1-score.

4. **(10 points)** Compute drop in impurity for all the nodes in the decision tree worked out example discussed in class. Use Gini and Misclassification impurity to show the drop.

5. **(10 points)** Consider the data points shown in the Figure:



Now, consider the following two extreme decision tree algorithms: (i) The ABC algorithm constructs a decision tree using the conventional approach but refrains from pruning at any stage. (ii) Conversely, the XYZ algorithm avoids the risk of splitting altogether, resulting in the entire decision tree being a single leaf node.

(a) What is the precise count of leaf nodes in the ABC decision tree generated on this dataset?

(b) Could you provide the leave-one-out classification error of applying ABC to this dataset? Please report the total number of misclassifications.

(c) What is the leave-one-out classification error when utilizing XYZ on our dataset? Kindly report the total number of misclassifications.

(d) Which among ABC and XYZ will overfit to the training data?

6. **(10 points)** Define the following Machine Learning terms: (a). Machine Learning (b). Overfitting (c). Supervised Learning (d). CART Decision Tree (e). ID3 Decision Tree.

7. **(6 points)** Consider fitting the linear regression model to the following data: (-1,0), (0,-1), (2,1)

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

8. **Additional practice:** Problem 2 and Problem 4 from here: `https://courses.cs.washington.edu/courses/cse546/14au/exams/14au_midterm_sol.pdf` Feel free to refrain from submitting a solution for this problem; however, it is strongly encouraged that you take the time to comprehend its nuances.

End of Paper