CS 2756, Spring 2022
Homework 3
Due: Mar 18, 2022, 11:59 PM, Gradescope

Instructions:

- The main submission of HW 3 must be uploaded on Gradescope. Your submission will be graded via Gradescope platform.
- Your PDF file should be named as "lastname_firstname_hw2". For example, if your name is Jane Doe, name your file as "doe_jane_hw2".
- Please write your name (First Name, Last Name) legibly, as it appears on Canvas. Please also include your PITT email.
- Type out your solutions on a separate blank file (using Word, Latex, etc.), and don't forget to convert the file to PDF extension before submitting. You should not type out on the original pdf file.
- Only a subset of the questions will be graded. However, such questions are not determined a priori, therefore please do your best to answer all the questions correctly.

Question 1.
The figure below shows a two-dimensional data set with two target classes (represented as circles and triangles) and the classification boundaries produced by each of the following classifiers: (1) Linear Support Vector Machine, (2) 1- Nearest Neighbor classifier and (3) Decision Tree. Match the classifiers (1), (2) and (3) with the corresponding classification boundaries in Figure 1 (A), (B) and (C) by filling the following table with the classifier selected and a short explanation of why you chose that classifier..

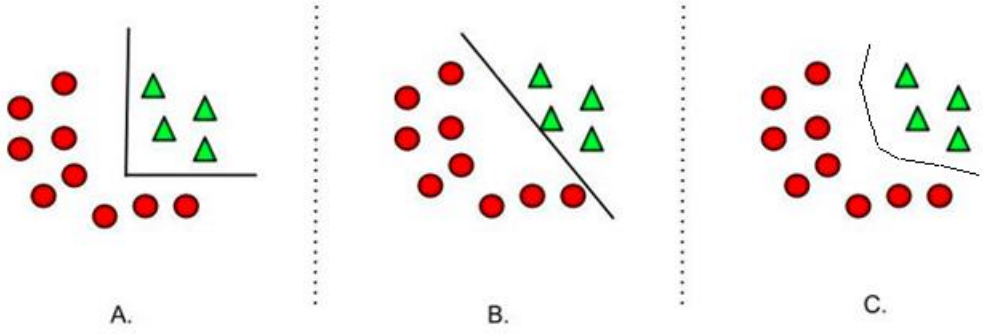


Figure 1: Two-dimensional dataset with classification boundaries from different classifiers

(a) Figure1(A)
Classifier (1,2, or 3):

(b) Figure1(B)
Classifier (1,2, or 3):

(c) Figure1(C)
Classifier (1,2, or 3):

Question 2.

Consider a data set with four binary attributes X1, X2, X3 and X4. The attribute X4 takes exactly the same value as X3 for each record, i.e., X4 is equal to X3. In each of the following three scenarios, find whether the decision boundary learnt by the two models would be similar, otherwise find which of the two models would perform better. Provide a brief justification for each.

(i) We build two KNN models:

a. M1 that is learnt using all the four attributes.
b. M2 that is learnt using the three attributes X1, X2, and X3.

(ii) We build two ANN models:

a. M1 that is learnt using all the four attributes.
b. M2 that is learnt using the three attributes X1, X2, and X3.

Question 3.

A realtor is studying housing values in the suburbs of Minneapolis and has given you a dataset with the following attributes: crime rate in the neighborhood, proximity to Mississippi river, number of rooms per dwelling, age of unit, distance to Minneapolis and Saint Paul Downtown, distance to shopping malls. The target variable is the cost of the house (with values high and low). Given this scenario, indicate the choice of classifier for each of the following questions and give a brief explanation.

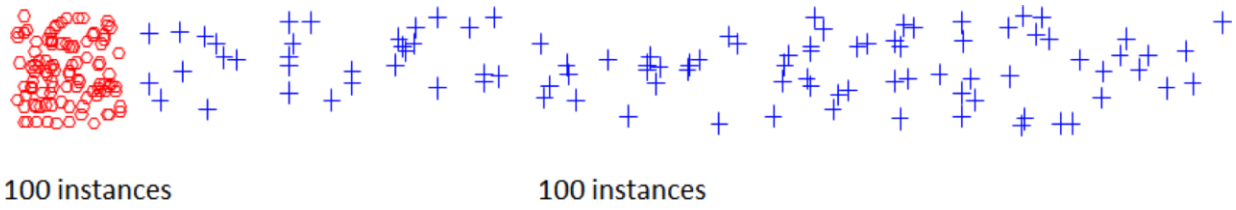
a) If the realtor wants a model that not only performs well but is also easy to interpret, which one would you choose between SVM, Decision Trees and kNN?

b) If you had to choose between RIPPER and Decision Trees, which one would you prefer for a classification problem where there are missing values in the training and test data?

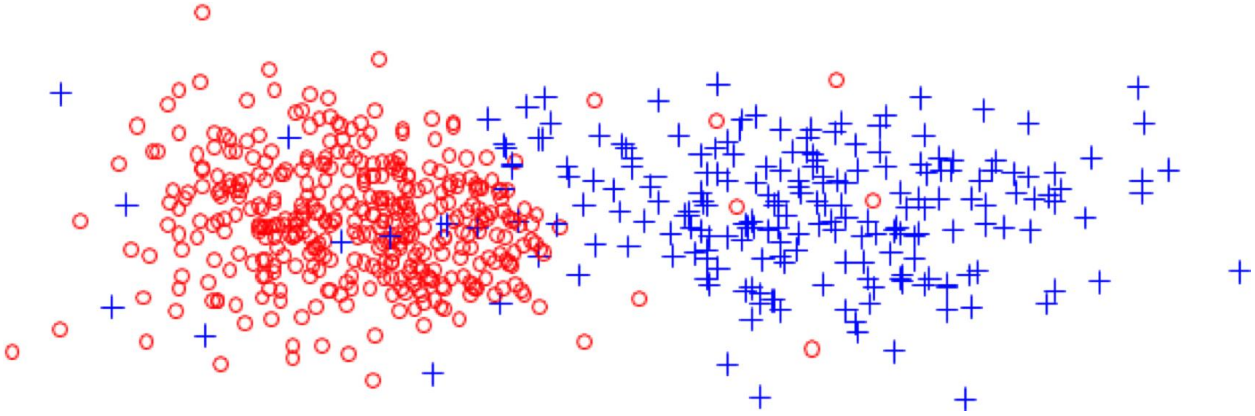
c) If you had to choose between RIPPER and KNN, which one would you prefer if it is known that there are very few houses that have high cost?

Question 4.

For each of the two given scenarios, make a right choice of K for the KNN classifier in order to obtain better performance with a brief explanation.



(a) K = 1 or K=5 or K = 50?



(b) K = 1 or K = 5 or K=50?

Question 5.

We have test data of 1000 samples with two classes: a + class (100 samples) and a – class (900 samples). Consider a random classifier C0 that classifies a test data instance to the + class randomly with a probability p.

a) What is the expected precision and recall for C0?

b) Write the expression for the F--measure of C0?

c) Consider another classifier C1 whose F--measure is known to be 0.15. Is C1 better than a random classifier?

Question 6.

A classifier is being tested on two datasets: Dataset 1, with 100 positives and 100 negatives, and Dataset 2, with 100 positives and 500 negatives. The confusion matrices of the classifier on the two datasets are provided below.

Dataset 1	Predicted '+'	Predicted '-'
Actual '+'	80	20
Actual '-'	20	80

Dataset 2	Predicted '+'	Predicted '-'
Actual '+'	80	20
Actual '-'	100	400

a) Calculate the Precision, Recall, TPR, and FPR for the classifier on Dataset 1 and Dataset 2.

Based on your observations from these results, if you had to choose between the following two evaluation metric pairs: {precision, recall} and {TPR, FPR}, which one would you choose in the following scenarios and why? Provide brief explanations in context with your observations from the results above.

b) The evaluation is required to be invariant to changes in the relative numbers of positives and negatives in the evaluation dataset.

c) Compute the accuracy of the classifier on Dataset 2 (you can leave your answer in fractions). Construct a trivial classifier that can achieve better accuracy on Dataset 2 without even looking at the attributes of the data. What is the accuracy of this trivial classifier?

Question 7.

Consider two classification algorithms that output real-valued scores. The class label prediction (positive or negative) is obtained by thresholding this score. The instances input to the algorithm and the corresponding scores are given below.

Instance ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
True Class Label	+	+	-	-	-	+	-	+	+	+	+	-	-	+	-	-
Score 1	7.9	8.2	1.2	2.8	2.5	9.2	5.8	6.3	8.9	5.2	2.1	3.2	4.5	6.2	1.9	0.7
Score 2	7.6	6.2	1.7	2.4	3.5	4.2	5.9	6.3	6.9	5.7	2.8	4.3	6.5	6.1	7.5	4.3

(a) Draw the ROC curve for the given data for both algorithms, thresholding the scores from 0 to 10, taking steps of 1. Note: For a threshold *th*, the instance is labeled as class positive (+) if score >= *th*, else it is labeled negative (-). You are free to either draw by hand, or code it using any programming language to generate the curve. If you decide on the latter, you do not need to submit the code, just put a screenshot of your resulting ROC curve in your submitted answer.

(b) Algorithm 2 is tested on the following two datasets with similar characteristics of positive and negative classes as in the training set -

a. 1000 positives, 100 negatives
b. 1000 positives, 1000 negatives

For each dataset, class predictions are generated using three thresholds on the score 2 - 3, 5 and 7. Mark these points on the ROC curve generated in part 1. For each of the two datasets, report the expected precision, recall, TPR, FPR and F-measure for the class predictions using the 3 different thresholds. Also, choose the best threshold in each case based on the computed F-measure.

(c) Based on the ROC curves plotted in part 1, choose the algorithm that is expected to perform better on a test data set with similar data characteristics. Justify.