Name:____________ PID:__________

Pledge: I have neither given nor received aid in this examination.

Signature:____________

CMSE381 QUIZ 2

Oct 15th, 2021

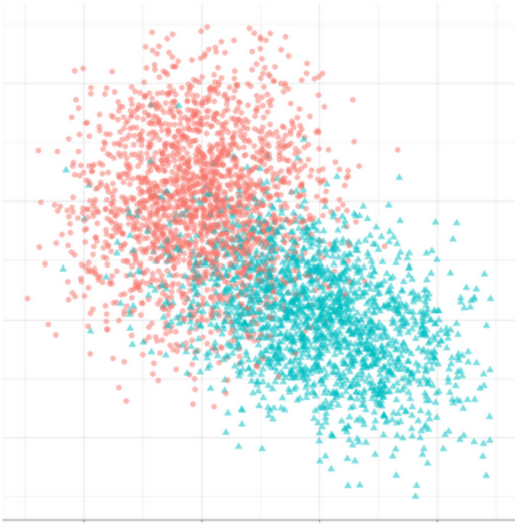
Instructions:
This is a closed book and closed notes examination. The best way to earn partial credits is to show all of your work. The instructor reserves the right to remove points if not all steps are shown. The total points are 20.You have 30 minus for this QUIZ. Good luck!

1.
  a (5 pts) For a classification problem with $K = 2$ ($Y \in \{0, 1\}$) and loss function of $\sum_{i=1}^{n} I(y_i \neq C(x_i))$, what is the oracle (the best classifier)?

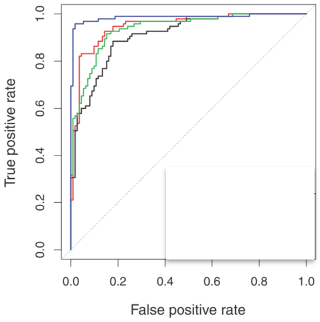
  b (10 pts) For a classification problem with $K = 2$, if we know Type I erorr will cost \$10 while Type II error will cost \$20. Derive the new oracle classifier which minimizes this cost. (Please write your answer on a white paper)

  c (5 pts) If $K = 3$ ($Y \in \{0, 1, 2\}$), will you prefer Logistic regression or LDA? Justify your answer

2. [10 pts] We want to build a binary classification model for the following data. Should we use a LDA or QDA model? Justify your choice.(Please write your answer on a white paper)



3. [5 pts] We are trying to predict whether a patient will have a heart attach within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.



4. [10 pts] We want to study the relationship between cups of coffee (Y) a people may drink per day and the stress score (X). Here X and Y are random quantities. From Dr. Sparty's research, we think the relationship is

$$Y = \alpha \exp(\beta \sin^2(X) + 100).$$

We then collected 10000 data $\{(x_1, y_1), \ldots (x_{10000}, y_{10000})\}$ and can estimate $\hat{\alpha}$ and $\hat{\beta}$ from this data. If we want to quantify the accuracy of our estimate of α and β in terms of confidence interval, what procedure will you follow to obtain it (Validation set, Cross validation, bootstrap, forward selection, backward selection, best subset, etc)? Outline your procedure. (Please write your answer on a white paper)

5. [5 pts] If we want to estimate the variance of testing error for our model, should we use CV or Bootstrap? Which method will produce a larger estimate of the testing error? Justify your answer

6. (Extra 2 pts) Prove that when sample size n goes to infinity, a bootstrap dataset will contain $1 - e^{-1}$ of original data.