DS 301: Homework 10
Due: April 27, 2022 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text file or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: MNIST handwritten digit database

Load the handwritten digits (MNIST) dataset into `R` using the `R` scripts we went over in class.

a. Randomly select 3000 observations from the training set and randomly select 100 observations from the test set. Implement KNN classification. Report the following:
   - Carry out 10-fold cross-validation on the training set to determine the optimal K. Try $K = 1, 5, 7, 9$. What is the optimal K?
   - Use this optimal K to implement KNN classification on the test set. Report your confusion matrix and overall misclassification error rate on the test set.
   - Report the specific misclassification rates for each digit on the test set.

b. Try to implement LDA on the MNIST dataset. What kind of error message do you obtain? Do some searching and explain what this error message means. Hint: check `var(train$x[,1])`.

c. Discuss how this dataset highlights some of the advantages of using KNN for classification.

Problem 2: Fashion MNIST

Many people consider the handwritten digits database to be far too easy for classifiers these days. More challenging datasets have appeared as new benchmarks. One example is the fashion MNIST dataset: https://github.com/zalandoresearch/fashion-mnist. Read up on the documentation for this dataset. Then, just like we did with the handwritten digits database, download the training and test set for this data. Load it into `R` using the same R scripts used in Problem 1.

a. Produce plots of the first 5 observations in the training set. What do you see?

b. Repeat Problem 1(a) for this dataset. How do your confusion matrices and misclassification error rates compare?

Problem 3: ROC Curve

Use the `Spam` data set, from HW 8, for this problem. Repeat your code from HW8, Problem 2 parts (b) and (c).

a. What type of mistake do we think is more critical here: reporting a meaningful email as spam (false positive) or a spam email as meaningful (false negative)?

b. Fit a logistic regression model here and apply it to the test set. Based on your answer to part (a), plot the ROC curve of true positive rate vs. false positive rate or true negative rate vs. false negative rate.

c. Output the confusion matrix. What is the false positive and false negative rate when we set the threshold to be 0.5?

d. Adjust the threshold such that your chosen error (false positive or false negative) is no more than 0.03. You should choose the threshold carefully so that the true positive (or true negative rate) are maximized. Report that threshold here.