DS 303 Homework 9
Due: Nov. 06, 2023 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` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word 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 misclassification error rate on the test set.

b. Try to implement LDA on the MNIST dataset. What kind of error message do you obtain? Do some digging and explain what this error message means.

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: Concept Review

(a) Suppose you just took on a new consulting client. He tells you he has a large dataset (say $100,000$ observations) and he wants to use this to classify whether or not to invest in a stock based on a set of $p = 10,000$ predictors. He claims KNN will work really well in this case because it is non-parametric and therefore makes no assumptions on the data. Present an argument to your client on why KNN might fail when p is large relative to the sample size.

(b) For each of the following classification problems, state whether you would advise a client to use LDA, logistic regression, or KNN and explain why:

i. We want to predict gender based on height and weight. The training set consists of heights and weights for 82 men and 63 women.

ii. We want to predict gender based on annual income and weekly working hours. The training set consists of 770 mean and 820 women.

iii. We want to predict gender based on a set of predictors where the decision boundary is complicated and highly non-linear. The training set consists of 960 men and 1040 women.

Problem 4: k-NN

Assume our outcome Y can take on $Y = 0$, $Y = 1$ or $Y = 2$ (3 categories). Suppose we have a training data set with 5 observations. We want to classify a test observation using KNN. Below are all the distances between each of the 5 observations in training set and the test observation.

Training Observation (i):	1	2	3	4	5
Y_i label:	0	1	2	0	2
distance:	0.45	0.23	0.61	0.31	0.57

(a) Based on the above, how would we classify our test observation using $K = 1$?

(b) How would we classify our test observation using $K = 3$?

(c) KNN is highly dependent on the choice of K. Discuss the bias/variance tradeoff we make in choosing K.

Problem 5: Email Spam Part 2

Use the `Spam` data set, from HW 8, for this problem. Repeat your code from Problem 2 parts (a), (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 and true negative rate are also maximized.Report that threshold here.

(e) Implement LDA and repeat parts (b) -(d).

(f) Carry out QDA, Naive Bayes and KNN on the training set. You should experiment with values for K in the KNN classifier using cross-validation. Remember to standardize your predictors for KNN. For each classifier, report the confusion matrix and overall test error rates for each of the classifiers.

(g) Which classifier would you recommend for this data? Justify your answer.

End of assignment