Homework 5: Classification

This homework covers material covered in Chapters 2,3, & 4 of *Introduction to Statistical Learning with Applications in R (ISLR)*. Some questions come directly from the book exercises, others are new questions.

Complete this homework as a jupyter notebook titled “LASTNAME_Homework5.ipynb” posted on your GitHub account. Turn in your assignment by emailing a link to the notebook to timothyv@andrew.cmu.edu by no later than 3pm on Tuesday April 3rd, 2018.

Conceptual: Short answer questions and proofs. Be concise. Show stepwise solutions to proofs where appropriate

1. Consider the problem of classifying a binary response variable (i.e., $y \in \{1, 0\}$). If there is no overlap in the values of X when y = 1 and when y=0, such that there is a large “gap” between the two distributions of X values, then this is problematic for one of the classifiers discussed in class and the text. What classifier does this situation pose a problem for? Explain conceptually why this is a problem and compare it with another classifier approach that does not suffer this limitation.

2. Compare logistic regression, LDA, and kNN classification approaches. Which are parametric which are non-parametric? For parametric models what functions do they assume? For non-parametric methods, how do the classifiers separate groups? How is the flexibility/bias tradeoff adjusted for each method?

3. What is the *curse of dimensionality*? Why is it especially problematic for kNN classification (i.e., why does kNN fail in high dimensional contexts)?

4. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Applied: Submit all R code, along with plots and written responses.

Do applied exercises 4.10 & 4.11