Assignment 2

October 6, 2020

All files referred to in this homework can be found on CourseWorks.

Your hand-in should be made on Gradescope. Please note that we are *changing* the submission instructions. Your submission is going to be a jupyter notebook that contains answers to all the questions. Please use headings to separate your notebook into answers to each question.

To see examples of how to write latex code in jupyter, see this link.

You must submit *two* files:

1. An `.ipynb` file (i.e. jupyter notebook file) with your data analysis for answering questions
2. A pdf file with your jupyter notebook output. Please use the workflow described here for generating this pdf.

1 ISLR Classification Lab

Logistic Regression, LDA, QDA, KNN

Complete the lab from Section 4.6 of ISLR. Feel free to utilize the provided worked jupyter notebook as inspiration.

You do not need to submit your code for this question, it is purely for you to practice with.

2 Classification Models for Stock Market Data

Solve exercise 10 from Section 4.7 of ISLR.

Logistic Regression, LDA, QDA, KNN

The `Weekly` dataset can be found in the Data folder. A description of it can be found here on page 14.

3 Reduced-Rank LDA

Let B and W be positive definite matrices and consider the following problem of maximizing the *Rayleigh quotient*:

Symmetric

$$\max_a \frac{a^T B a}{a^T W a} \tag{1}$$

1. Use the method of Lagrange multipliers to solve this problem. In particular, show that the optimal solution a^* is an eigenvector of a certain matrix related to B and W. What is this matrix, and which eigenvector does a^* correspond to?

   Hint: Use the *scale invariance* of the Rayleigh quotient to rewrite the unconstrained maximization as a constrained maximization problem where B appears in the objective, and W appears in the constraint.

2. By identifying B and W with the between-class and within-class covariance matrices, we can interpret the problem in (1) as the problem of finding the linear combination $a^{*T}x$ so as to maximize the *between-class variance* relative to the *within-class variance*. Show that $a^{*T}x$ is the first discriminant variable.

   Hint: First note that $W = \Sigma$ from the lecture slides, and that $B^* = D^{-1/2}U^T BUD^{-1/2}$

4 Logistic Regression

1. Show that binary classification using logistic regression yields a linear classifier.

Consider a naive Bayes classifier for a binary classification problem where all the class-conditional distributions are assumed to be Gaussian with the variance of each feature X_j being equal across the two classes. That is we assume $(X_j|Y = k) \sim N(\mu_{jk}, \sigma_j^2)$

2. Show that the decision boundary is a linear function of $X = (X_1, ..., X_d)$ and hence that it has the same parametric form as the decision boundary given by logistic regression.

3. Does the result of part (2) imply that in this case, Gaussian naive Bayes and logistic regression will find the same decision boundary? Justify your answer.

4. If indeed the class conditional distributions are Gaussian with $(X_j|Y = k) \sim N(\mu_{jk}, \sigma_j^2)$ and the assumptions of naive Bayes are true, which classifier do you think will be "better": the naive Bayes classifier of part (2) or logistic regression? Justify your answer.

5 Bootstrap Probabilities

Solve exercise 2 from Section 5.4 of ISLR.