3AS/3AS4: Applied Statistics (Spring Term)

ASSIGNMENT 4 Date: March 7, 2020

To be submitted by 5pm, March 20, 2020
Submit in a single .pdf file

1. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{Class is Red}|X)$:

   $$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, \text{ and } 0.75.$$

   There are two common ways to combine these results together into a single class prediction. One is the majority vote approach. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

2. Consider the `Carseats` data in the `ISLR` package. Now we will seek to predict `Sales` using regression trees and related approaches, treating the response as a quantitative variable.

   (a) Split the data set into a training set of 300 observations and a test set of remaining observations.

   (b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test error rate do you obtain?

   (c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?

   (d) Use the bagging approach in order to analyze this data. What test error rate do you obtain? Use the `importance()` function to determine which variables are most important.

   (e) Use random forests to analyze this data. What test error rate do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

3. (a) Generate a data set with $n = 500$ and $p = 2$, such that the observations belong to two classes with a quadratic decision boundary between them. For instance, you can do this as follows:

   ```
   > x1=runif (500) -0.5
   > x2=runif (500) -0.5
   > y=1*( x1^2-x2^2 > 0)
   ```

   (b) Plot the observations, coloured according to their class labels. Your plot should display X_1 on the x-axis, and X_2 on the y-axis.

(c) Fit a support vector classifier to the data with X_1 and X_2 as predictors. Obtain a class prediction for each training observation. Plot the observations, coloured according to the predicted class labels.

(d) Fit a SVM using a non-linear kernel to the data. Obtain a class prediction for each training observation. Plot the observations, coloured according to the predicted class labels.

(e) Comment on your results.

4. A researcher collects expression measurements for 1,000 genes in 100 tissue samples. The data can be written as a $1,000 \times 100$ matrix, which we call X, in which each row represents a gene and each column a tissue sample. Each tissue sample was processed on a different day, and the columns of X are ordered so that the samples that were processed earliest are on the left, and the samples that were processed later are on the right. The tissue samples belong to two groups: control (C) and treatment (T). The C and T samples were processed in a random order across the days. The researcher wishes to determine whether each gene's expression measurements differ between the treatment and control groups.

   As a pre-analysis (before comparing T versus C), the researcher performs a principal component analysis of the data, and finds that the first principal component (a vector of length 100) has a strong linear trend from left to right, and explains 10% of the variation. The researcher now remembers that each patient sample was run on one of two machines, A and B, and machine A was used more often in the earlier times while B was used more often later. The researcher has a record of which sample was run on which machine.

   (a) Explain what it means that the first principal component "explains 10% of the variation".

   (b) The researcher decides to replace the (i, j)th element of X with

   $$x_{ij} - u_{i1}v_{j1}$$

   where u_{i1} is the i-th score, and v_{j1} is the jth loading, for the first principal component. He will then perform a two-sample t-test on each gene in this new data set in order to determine whether its expression differs between the two conditions. Critique this idea, and suggest a better approach.

   (c) Design and run a small simulation experiment to demonstrate the superiority of your idea.