STAT 5650

Statistical Learning and Data Mining I

Homework #3

Due: *Friday, March* 13.

1. This is a continuation of the analyses on the data for three bird species—(*Northern*) *Flicker*, (*Mountain*) *Chickadee*, and (*Red-naped*) *Sapsucker*—plus a bunch of sites at which none of these species of birds are nesting. In the previous homework you analyzed these data using logistic regression and LDA/QDA; in this question I would like you to apply k-Nearest Neighbor classification to the data and to compare your results with the results for LDA, QDA, and logistic regression.

    a) Apply *k*-NN classification to the combined dataset for all 3 species using 'Nest' as the response variable. What value of *k* did you select. Compare the results of this classification with the results you obtained for LDA, QDA, and logistic regression in the previous homework.

    b) Now apply *k*-NN classification to the three datasets for the individual. What values of *k* did you select. Compare the results of this classification with the results you obtained for LDA, QDA, and logistic regression in the previous homework.

2. This question is also a continuation of the analyses on the data for three bird species—(*Northern*) *Flicker*, (*Mountain*) *Chickadee*, and (*Red-naped*) *Sapsucker*—plus a bunch of sites at which none of these species of birds are nesting. In the previous homework you analyzed these data using logistic regression and LDA/QDA; in this homework I would like you to apply classification trees to the data. The first priority is to come up with accurate classifications of the nest sites, the second priority is to determine important variables to the birds in selecting nest sites, and the third priority is to determine whether the three species can be treated as one species (with regard to selection of bird nest sites) or need to be treated separately.

    a) First, fit a classification tree to all the data treating the three birds as a single species. Compute the accuracy of your classification using 10-fold cross-validation, and compare it with cross–validated accuracy rates for LDA, QDA, and logistic regression that you computed before.

    b) Fit classification trees for each bird species separately and, again, compute estimates of the accuracies by 10-fold cross validation. Qualitatively compare the classification trees for the three species.

c) Another way to get at the issue of the similarity of the bird species might be to do the following:

- Fit a classification tree to the combined data using **Species** as the response variable.

- Look at the cross-validated confusion matrix for the classification tree to see where the misclassifications are occurring.

3. This problem concerns the forensic glass data set labeled "Glass.csv" in Canvas. There are six different types of glass, coded 1—6, and nine measured variables. The first of the measured variables is the refractive index of the glass, and the remaining eight are weight percentages of eight chemical elements. The purpose of the analysis is to classify the six types of glass using the refractive index and the chemical percentages.

   a) Fit a classification tree to the data using the 1-SE rule or choosing a tree just a little smaller or larger than the one selected by the 1-SE rule. Briefly summarize the tree.

   b) Compute the 10-fold cross-validated confusion matrix. If you have trouble doing this, you may have to consider eliminating some types of glass or collapsing categories of glass that may be similar and have small numbers of observations.

Glass Code	**Type of Glass**
1	*Building windows, float processed*
2	*Building windows, non-float processed*
3	*Vehicle windows, float processed*
4	*Containers*
5	*Tableware*
6	*Auto headlamps*