STAT 5650

Statistical Learning and Data Mining I

Homework #3

Due: *Wednesday, March* 1.

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 LDA and QDA; in this question I would like you to apply k nearest neighbor (k-NN) classification to the birds nest data.

   Important: *For these data, in all your k-NN classification you should use the transformed predictor variables that you derived in the previous homework*.

   a) Apply k-NN classification to the combined dataset for all 3 species using 'nest' as the response variable (and excluding 'Species' from the analyses). Summarize your results and compare the accuracy of prediction to your previous analyses using LDA and QDA.

   b) Apply k-NN classification to the three individual datasets for the three species that you obtained in the previous homework. Summarize your results and compare the accuracy of the predictions to your previous analyses using LDA and QDA, and for the three species individually.

2. This problem is also 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. Previously you analyzed these data using LDA, QDA and, in Q1 of this homework, k-NN classification; in this question I would like you to apply logistic regression with and without variable selection to the birds nest data.

   Important: *For these data, in all your logistic regressions you should use the transformed predictor variables that you derived in the previous homework*.

   a) Apply logistic regression *with no variable selection* to the combined dataset for all 3 species using 'nest' as the response variable (and excluding 'Species' from the analyses). Summarize your results and compare the accuracy of prediction to your previous analyses using LDA, QDA, and k-NN.

b) Now carry out logistic regression *with variable selection* on the combined dataset with 'Nest' as the response. You may decide which variable selection method to use. It could be backward elimination with *P*-values (SAS) or stepwise variable elimination in R using the 'step' function and a criterion like the AIC, or the LASSO in either package. Which variables were removed and which were retained in the model? Compare the predictive accuracy for this model to LDA, QDA, k-NN, and logistic regression with no variable selection.

c) Apply logistic regression *with no variable selection* to the three individual datasets for the three species that you obtained in the previous homework. Summarize your results and compare the accuracy of the predictions to your previous analyses using LDA, QDA, and k-NN for the three species individually.

d) Now carry out logistic regression *with variable selection* on each of the three datasets for different species. Again, you may decide which variable selection method to use. Which variables were removed and which were retained in each of the models? Compare the predictive accuracy for these models to the accuracies for LDA, QDA, k-NN that you obtained in the previous homework, and to logistic regression with no variable selection.

3. For the Pima diabetes data, 'Pima Diabetes 3.csv,' we would like to fit logistic regression models to the data and determine which of the measured variables is useful in predicting whether a person has diabetes. Carry out an appropriate analysis and report on your findings. Some aspects that you may wish to take into account are the distributions of the variables and whether or not to apply variable selection. You should include cross-validated error rates for your final model.