Stat 5650

Statistical Learning and Data Mining I

Homework #2

Due: *Friday, February* 14.

1. This question concerns Fisher's iris data, one of the most well-known and, perhaps, overused datasets. I have placed the data ("Iris.csv") on the Canvas site for the class. For this homework, I would like you to apply linear and quadratic discriminant analysis to see if the measured values or petal length and width and sepal length and width can be used to discriminate among the three species of iris.

    a. Summarize the four measured variables for the three types of iris. Are the data approximately normal, and do they look like they have the same covariance matrix for all 3 species?
    b. Test to determine whether the covariance matrices for the three species may be pooled.
    c. Apply **both** LDA or QDA. Obtain the cross-validated confusion matrices and accuracy or error rates (by species and overall).
    d. Determine whether some of the measured variables are redundant and can be removed.

2. I have placed a dataset called "Nest.csv" on the Canvas site for the class. This dataset contains data on nest sites for three bird species—(*Northern*) *Flicker*, (*Mountain*) *Chickadee*, and (*Red-naped*) *Sapsucker*—plus a bunch of sites at which none of these birds are nesting. The variable Nest is the response or grouping variable. Species indicates the species of nesting bird, and StandType is a dummy variable coded as 0 for pure aspen forest and 1 for mixed aspen and conifer.

    a. Carry out numerical and graphical summaries of all the predictor variables except StandType. Are the variables approximately normal in distribution? If not, apply some transformation(s) to "improve" the distributions of these variables.
    b. Fit LDA and QDA to all the data (with the transformed predictor variables) treating the three birds as a single species. Compare the accuracies or error rates of your classifications using cross-validation.
    c. For each bird species separately, construct a dataset that comprises the data for all the **nest sites for that species**, and **all the non-nest sites**. Now, refit LDA and QDA, **for each bird species separately** and compare the results for the different methods.

3. Continuing with the Nest data, with the transformed predictor variables,

    a. Fit a logistic regression model with all the data. Compare the cross-validated accuracies or error rates with those for LDA and QDA that you obtained in the previous question.
    b. Now apply some variable selection procedure (in logistic regression) and identify variables important to the classification. By how much did the cross-validated accuracies/error rates change?
    c. Repeat part a. using the datasets for the individual bird species.
    d. Repeat part b. using the datasets for the individual bird species. Are there variables that are in the models for 2 or 3 of the species?

Classification and Discriminant Analysis (Supervised Learning)

Cavity Nesting birds in the Uintah Mountains, Utah

- Response variable is the presence (coded 1) or absence (coded 0) of a nest.
- Predictor variables (measured on 0.04 ha plots around the sites) are:
    - Numbers of trees in various size classes from less than 1 inch in diameter at breast height to greater than 15 inches in diameter.
    - Number of snags and number of downed snags.
    - Percent shrub cover.
    - Number of conifers.
    - Stand Type, coded as 0 for pure aspen and 1 for mixed aspen and conifer.