For: Dr. Trenary
By: Christopher Carlson

CS5950 Machine Learning Assignment 2: Iris Data Classification

Results:

The overall results of my classification attempts were very good for all methods used. I had near perfect results using Quadratic Discriminant Analysis – depending on the partition sample the results ranged in the upper ninety percents, frequently it reached one-hundred percent. Linear Discriminant Analysis was also quite good – depending on the data-partition it reported results in the high eighty percent range up to full one-hundred percent accuracy. Logistic Regression showed acceptable results as well, though its results were in the sixty to eighty-five percent accuracy range.

I partitioned the data into a training set and a testing set. I did this by randomly sampling one quarter of all the records and putting them into the testing set, and then I used all the records that were not selected for testing to create the training set. This partitioning scheme is what I used for most of the testing, except for my initial logistic regression test. For that, because I was unaware of the vglm library, I used ruby scripts to partition the data – this procedure is described below.

In order to do Logistic Regression I initially used Ruby Scripts to restructure the data into training data and testing data, though I did a poor job structuring this data. I arbitrarily chose 5 observations from each species and put them into the testing data file and I used the rest for training data. I then created three new sets of training data, one for each species of iris, such that the species column only contained the values, “'species_for_which_training_data_was_made'”, and “other”. This way I could still use the logistic regression model with family “binomial”. This way a predictor can be made for each species which predicts either, “Yes, this is species x” or “No, this belongs to some other group.”

The interesting thing about the logistic regression models was that for Iris-Virginica and for Iris-Setosa, the logistic regression models were highly effective. They rarely mis-categorized anything. For the Versicolor species though the results were terrible. The only way I could get decent results was to raise the prediction threshold up higher. When I moved the threshold to .6, the model correctly identified 2 out of five as iris-versicolor. Moving the threshold to .7, the model predicted 4 out of five as iris-versicolor, but now it misidentified 1 iris-virginica as iris-versicolor.

Logistic Regression Initial Output Sample

```
                  Iris-virginica | other
Iris-setosa               0        5
Iris-versicolor           0        5
Iris-virginica            5        0


                   Iris-setosa | other
Iris-setosa              5       0
Iris-versicolor          0       5
Iris-virginica           0       5

                  Iris-versicolor | other
Iris-setosa               0          5
Iris-versicolor           4          1
Iris-virginica            1          4
```

After class on Monday I decided to try out the 'vglm' library to do the logistic regression, and it worked well, though the results were still imperfect. It was easier to implement and faster to use. It is the preferred method in the future.

Logistic Regression Output Sample from VGLM

```
glm.pred     setosa | versicolor | virginica
  setosa        3           0            0
  versicolor    0          10            3
  virginica     8           0           13
[1] 0.7027027
```

Sample Output from Linear Discriminant Analysis

```
             setosa | versicolor | virginica
  setosa       11           0           0
  versicolor    0           9           3
  virginica     0           1          13
[1] 0.8918919
```

Sample Output from Quadratic Discriminant Analysis

```
             setosa | versicolor | virginica
  setosa        11         0            0
  versicolor     0         9            1
  virginica      0         1           15
[1] 0.9459459
```