Additional Multivariate Exercise 16

Brief solution only

Data: The Iris data were introduced in the lecture for Session 9 as an example of a dataset with well separated clusters. It contains four measurements of dimensions of flowers (sepal length and width, and petal length and width) of 50 samples from each of three different varieties/species of Iris. The data were actually collected by the American botanist Edgar Anderson, and two of the three species were collected in Gaspé Peninsula (in the province of Quebec).

Classification analyses: A linear discriminant analysis shows that the problem is almost one-dimensional because more than 99% of the between-group variance is on the first discriminant function. The focus here is however on comparing different classification algorithms on this, supposedly easy, classification problem. All assessment of performance should use leave-one-out classification. The Minitab implementation (for LDA and QDA) is applicable to the problem because there is no interest in anything else than uniform priors.

It is useful to start with descriptive statistics for the four variables across the groups, as shown in the table below. The table shows, among other things, that both petal variables completely separate the setosa species from the two others by their much lower values.

```
. tabstat sepallength-petalwidth, statistics( mean sd min max ) by(Species)

Summary statistics: mean, sd, min, max
  by categories of: Species (Species)

   Species |  sepall~h  sepalw~h  petall~h  petalw~h
-----------+--------------------------------------
    setosa |     5.006     3.428     1.462      .246
           |  .3524897  .3790644   .173664  .1053856
           |       4.3       2.3         1        .1
           |       5.8       4.4       1.9        .6
-----------+--------------------------------------
versicolor |     5.936      2.77      4.26     1.326
           |  .5161712  .3137983   .469911  .1977527
           |       4.9         2         3         1
           |         7       3.4       5.1       1.8
-----------+--------------------------------------
 virginica |     6.588     2.974     5.552     2.026
           |  .6358795  .3224966  .5518947  .2746501
           |       4.9       2.2       4.5       1.4
           |       7.9       3.8       6.9       2.5
-----------+--------------------------------------
     Total |  5.843333  3.057333     3.758  1.199333
           |  .8280661  .4358663  1.765298  .7622377
           |       4.3         2         1        .1
           |       7.9       4.4       6.9       2.5
---------------------------------------------------
```

The within-group distributions of the variables are helpful to understand the misclassifications. Among the methods we include the two discriminant analyses (LDA and QDA), logistic classification, and k-nearest neighbours for $k = 1, \ldots, 8$. The choice of at most eight 8 neighbours follows one of the recommendations for the choice of k as being not much larger than the square-root of the

typical group size. It turns out that logistic classification struggles quite a bit with these data because the classification becomes too good! If complete separation of the groups can be achieved, then the logistic model is trying to predict probabilities that are either 0 or 1, and we know from logistic regression that this causes trouble. Accordingly, analysis with the standard command for multinomial logistic regression in Stata produces a steady flow of warnings and error messages. Even with the more robust implementation via the discrimination commands, the leave-one-out cross-validation was unable to classify all observations. The (few) misclassifications by the four methods are tabulated below.

```
. list if ldagroup~=Species | qdagroup~=Species | logitgroup~=Species | knngroup~=Species,
      noobs compress
+-------------------------------------------------------------------------------------------+
| obs   s~gth   s~dth   p~gth   p~dth    Species    ldagroup    qdagroup  logitgroup    knngroup |
|-------------------------------------------------------------------------------------------|
|  17     5.4     3.9     1.3      .4     setosa      setosa      setosa           0      setosa |
|  38     4.9     3.6     1.4      .1     setosa      setosa      setosa           0      setosa |
|  69     6.2     2.2     4.5     1.5   versicol    versicol   virginica    versicol    versicol |
|  71     5.9     3.2     4.8     1.8   versicol   virginica   virginica   virginica    versicol |
|  73     6.3     2.5     4.9     1.5   versicol    versicol    versicol    versicol   virginica |
|-------------------------------------------------------------------------------------------|
|  84       6     2.7     5.1     1.6   versicol   virginica   virginica           0   virginica |
| 107     4.9     2.5     4.5     1.7  virginica   virginica   virginica   virginica    versicol |
| 134     6.3     2.8     5.1     1.5  virginica    versicol    versicol           0   virginica |
+-------------------------------------------------------------------------------------------+
```

Observations with `logitgroup` equal to zero could not be classified by the logistic classifier. All other methods perform a classification for these samples, and for three of them it is unanimously correct. Unsurprisingly, no misclassifications involves the setosa species; we already noted its complete separation in the data. Sample 84 was not classified correctly by any of the classifiers; apparently its high value of petal length was too difficult to reconcile with the versicolor species. A similar situation seemed to have occurred with sample 71, but here the KNN method managed to get the classification right. On the other hand, the KNN method failed for sample 107, which all other methods got right. For this sample, the sepal variables may have been behind the misclassification, and in view of those values it is perhaps a bit surprising that the other methods classified this sample correctly as virginica.

In summary, all methods performed well, but one should be aware of the limitations of logistic classification method when very high classification rates occur, due to its difficulties with estimating the corresponding model parameters on logistic scale.