**Linear Discrimination on Breast Cancer Data**

**Written by: Juan Wei**

**Dec. 22, 2013**

**Summary**

Linear discrimination analysis is algorithmically simple yet provides powerful discriminators and reliable results. In this report, three different linear discriminators, namely linear regression, linear discriminant analysis and logistic regression, are applied on a binomial response data. Dimension of the data is reduced after processed by PCA and FDA and only two scores are retained. Errors are below 17.7% among various discriminators. To get more insights, training set and test set are switched and some other linear regressions discriminators are played upon this data.

**I. Data Source**

The data used in this report is Breast Cancer Wisconsin (Diagnostic) Data Set, consisting of 30 real attributes and one binomial response. The real-valued features are computed for each cell nucleus from radius, texture, perimeter, area, smoothness, etc., which are measured from digitized images. The data provides 569 observations and the response is either malignant or benign.

**II. Linear discrimination analysis on PCA and FDA scores**

The data is split into training set and data set, containing 300 and 269 data points respectively. The benign group is labelled as group one and colored in black in all the plots below, while the malignant group is labelled as group two and colored in red. For two-classes data, linear regression and LDA end up with same results.

**A. Principal component analysis**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Singular values | Training set scores | Test set scores |

Figure 1

In figure 1, singular values of the training set are plotted and two principal components are distinguishable from the rest. The first two scores are plotted for both training set and test set. The benign group is concentrated in a corner while the malignant cells spread widely out. This partly reflects the properties of the breast fat cells which is that the benign cells are more regular than the malignant cells.

|  |  |  |  |
| --- | --- | --- | --- |
| Linear regression/LDA | | Logistic regression | |
|  |  |  |  |
| Training error: 0.133 | Test error: 0.150 | Training error: 0.067 | Test error: 0.177 |

Figure 2

In figure 2, the errors of the two sets from all three methods are below 17.7%.

**B. Fisher discriminant analysis**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Singular values | Training set scores | Test set scores |

Figure 3

In figure 3, singular values of the training set are plotted and one component stands out. The first two scores are plotted for both training set and test set. The benign groups in both training set and test set are again densely located in a corner and linearly stretching out, but in a different direction from PCA scores.

|  |  |  |  |
| --- | --- | --- | --- |
| Linear regression/LDA | | Logistic regression | |
|  |  |  |  |
| Training error: 0.137 | Test error: 0.150 | Training error: 0.067 | Test error: 0.177 |

Figure 4

In figure 4, the errors of the two sets from all three methods are very close to those in PCA. PCA and FDA show no significant difference to this extent.

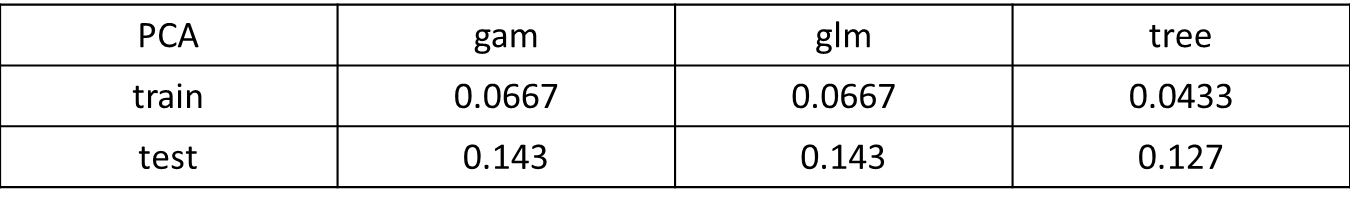
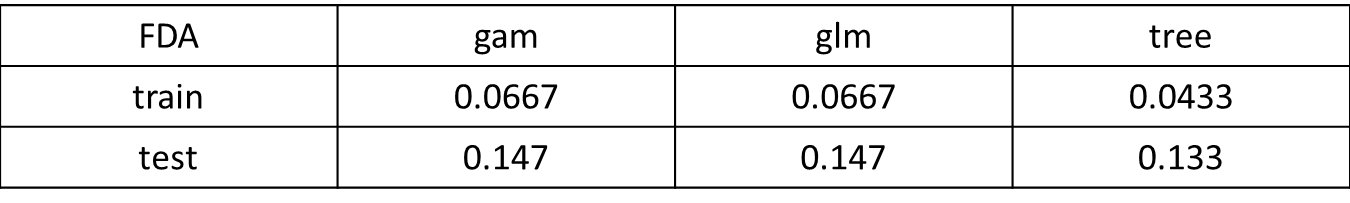
**C. Data structure analysis**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Singular values | Training set scores | Test set scores | Test error: 0.693 |

Figure 5

In figure 5, the two sets of data are switched and the PCA singular values and fires two scores are plotted. The scores of the benign group are again linearly expanding but yet to a different direction than before. This is because we are now projecting the data onto the loadings of the latter 269 data points which is different from the loadings of the first 300 data points. It is important to notice that the data should be projected onto the same direction before discrimination since different data set might have distinct data structure. Huge error, as large as 69.3% will be proposed if the model was built upon the data projected onto the loadings of the training set, but yet applied to the test data which are projected onto the loadings of the test set.

**D. Generalized additive model, Generalized linear model and Tree**

**** Table 1

More linear discriminators such as gam and glm are applied on this data yet with no significant improvement. Tree method is also tried and provides nearly same error rates.

**Conclusions**

1. Linear discriminators are reliable with errors as low as 17.7%. PCA and FDA show no difference in this data set.
2. It is important to notice that the data should be projected onto the same direction when using PCA, otherwise it will result in huge errors.
3. Other linear discriminators and tree method provides nearly same error rates for this data.