MDA project

John Ensley, Mohammad Mottahedi and Songshan Yang

April 20, 2016

1 Introduction of the data set

In this project, we applied the mixture discriminant analysis (MDA), linear discriminant analysis and quadratic discriminant analysis (QDA) for to the breast cancer data set. This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. There are 699 subjects in the data set and all the subjects are classified to 2 different classes–”Benign” and ”Malignant”. There are 9 features include clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell Size, bare nuclei, bland chromatin, normal nucleoli, mitoses and all of them range from 1 to 10.

2 Pseudocode

The pseudocode of the EM algorithm of MDA is shown below:

We use R programming to finish the algorithm by following the above pseudocode.

Procedure 1 EM algorithm for MDA

```
Require: $X$ = matrix of covariates; $Y$ = class; $R$ = number of subclasses in each class
Ensure: $\pi$ = mixture weights; $\mu$ = subclass means; $\Sigma$ = shared subclass variance
  $K :=$ number of classes
  $m :=$ number of covariates
  $n :=$ number of observations
  $\Sigma = I_{m \times m}$
  for $k = 1$ to $K$ do
    $X^*$ = rows of $X$ that belong to class $k$
    $\mu_{k1}, \ldots, \mu_{kr}$ = result from $k$-mean clustering on $X^*$ with $R$ clusters
    $\pi_{k1}, \ldots, \pi_{kr} = 1/R$
  end for
  while likelihood has not converged do
    for $i = 1$ to $n$ do
      $k :=$ class that $x_i$ belongs to
      for $r = 1$ to $R$ do
        $p_{i,r} = \frac{\pi_{kr}\phi(x_i|\mu_{kr},\Sigma)}{\sum_{r'=1}^{R} \pi_{kr'}\phi(x_i|\mu_{kr'},\Sigma)}$
      end for
    end for
    for $k = 1$ to $K$ do
      $X^*$ = rows of $X$ that belong to class $k$
      for $r = 1$ to $R$ do
        $\pi_{kr} = \frac{\sum_{i=1}^{n} I(X_i \in X^*) p_{ir}}{\sum_{i=1}^{n} I(X_i \in X^*)}$
        $\mu_{kr} = \frac{\sum_{i=1}^{n} X_i I(X_i \in X^*) p_{ir}}{\sum_{i=1}^{n} I(X_i \in X^*) p_{ir}}$
        $\Sigma = \frac{\sum_{i=1}^{n} \sum_{r=1}^{R} p_{ir}(X_i - \mu_{kr})(X_i - \mu_{kr})'}{n}$
      end for
    end for
  end while
  return $\pi$, $\mu$, $\Sigma$
```

Procedure 2 MDA

```
Require: $\pi$ = mixture weights; $\mu$ = subclass means; $\Sigma$ = shared subclass
    variance; $X$ = new covariate matrix
Ensure: $Y$ = classification labels for $X$
    $K$ := number of classes
    $n$ := number of observations
    $R$ := number of subclasses in each class
    for $i = 1$ to $n$ do
        for $k = 1$ to $K$ do
            $p_k = \sum_{r=1}^{R} \pi_{kr} \phi(X_i | \mu_{kr}, \Sigma)$
        end for
        $Y_i = \arg\max_k p_k$
    end for
    return $Y$
```

3 Mixture Discriminant Analysis

In the first step, we applied MDA to the whole data set to examine their classification error rate. We found that there are only 16 subjects having missing data and we delete these subjects, since the results will not be affected because of the large sample size.

There are 458 subjects belonging to Benign and 241 subjects belonging to Malignant, so the prior probabilities are $\pi_1 = 0.65$ and $\pi_2 = 0.35$ for benign and malignant, respectively.

We changed the number of components in each mixture distribution from $1 - 6$. the models were trained on 60 percent of the data and the rest of the data was used for testing the accuracy. Figure [?] show the mixture discriminant accuracy with respect to number of components.

The model with single component mixture distribution has the best accuracy on the test set. Increasing the number of component does not improve the accuracy.

Next, we used principle component analysis to reduce the dimentionality and used the first two component which allows us to visually inspect the the model. The first two component of the principal component analysis can describe 76.2% of the variation in the data. Figure [?] shows the result applying principal component analysis on the data set.

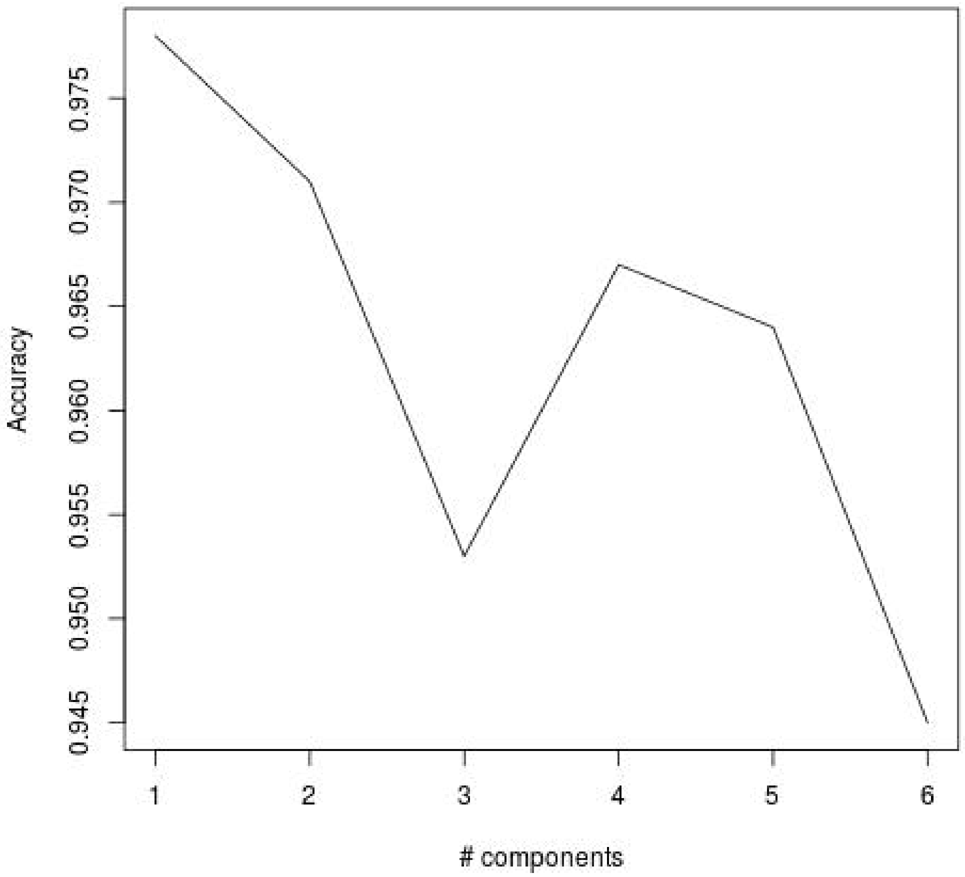

Figure 1: Mixture discriminant analysis accuracy

The accuracy of MDA applied to the first two component of PCA is shown in Figure [?]

Dimensional reduction did not improved the accuracy.

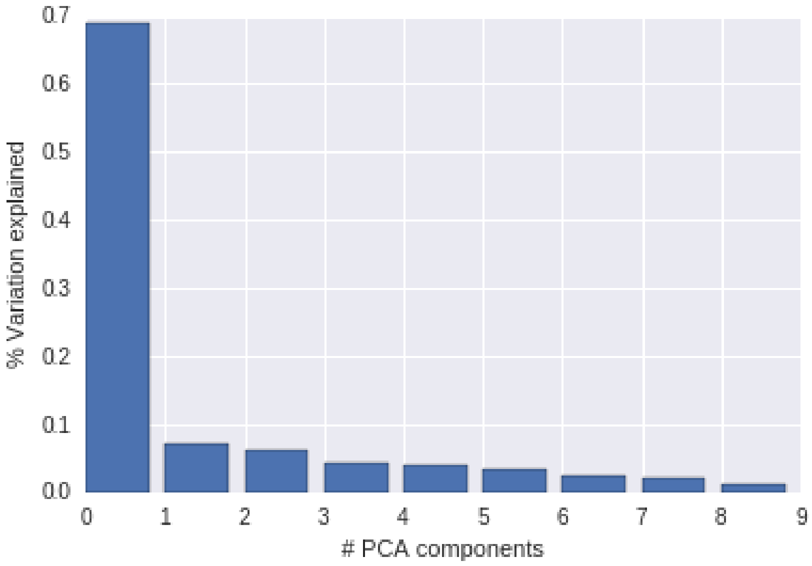

Figure 2: Principal Component analysis

4 Quadratic , Linear Discriminant Analysis, Logistic Regression, CAR

To compare the performance of the Mixture discriminant analysis we also trained Quadratic and Linear discriminant analysis on the breast cancer data set. For LDA, we compute the means of the class ”Benign” and ”Malignant” and the covariance matrix of the 9 features. Then we classify the subjects according to the maximum of their discriminant functions. For QDA, the process is similar to that of LDA, but we should compute the covariance matrix for each class and then classify the subjects.

The classification results are summarized in Table 1.

Table 1: Classification Error Rate of LDA, QDA and Logistic Regression

Method	MDA	CART	LDA	QDA	Logistic Regression
Error Rate	0.978	0.949	0.9605	0.974	0.9693

From Table 1, we can see that MDA and QDA have similar classification error rates and MDA and QDA outperforms the logistic regression and LDA and CART in the classification, CART performs the worst.

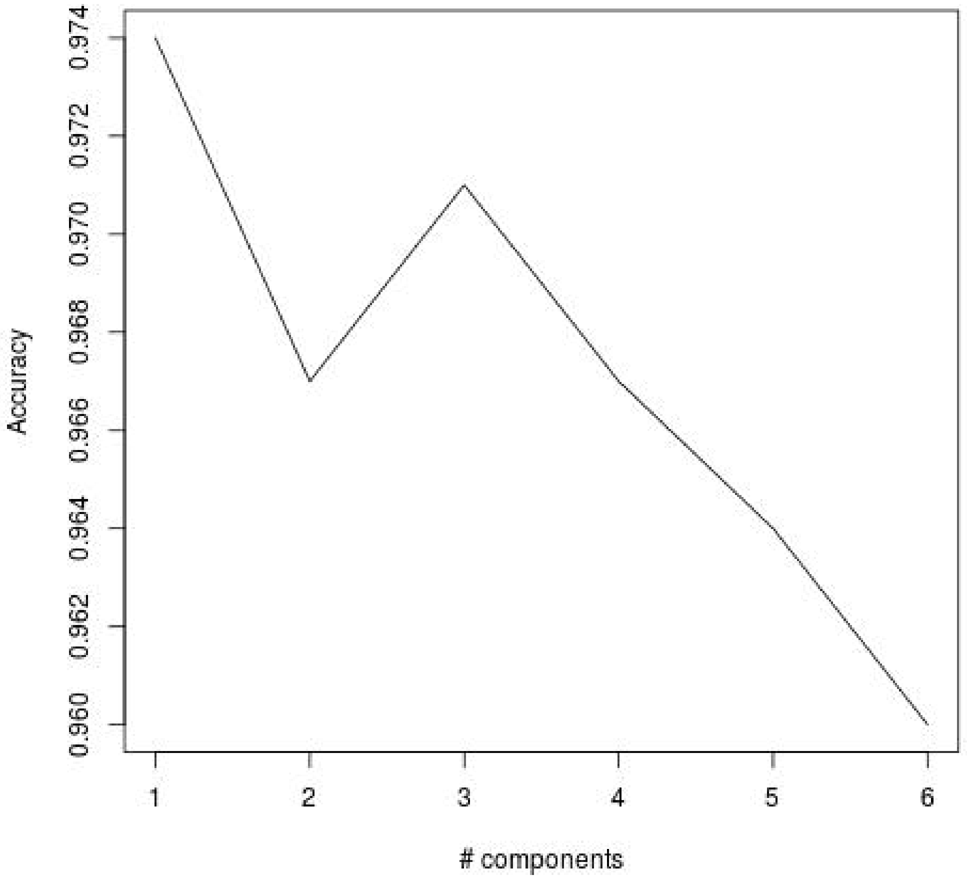

Figure 3: Mixture discriminant analysis accuracy after PCA

5 PCA on LDA, QDA and Logistic regression

In this section, we examined the effect of dimension reduction on classification. We first standardized the design matrix and did eigen decomposition of the covariance matrix of the standardized design matrix.

We decided to use the first two components which account for 76.2% of the total variance. Then the design matrix only has two columns. By repeating the process described in section 2, the classification error rates are summarized in Table 2. All the simulation results are based on 100 independent replicates. Compared Table 2 and Table 1, we can see that PCA decreased the accuracy of the QDA slightly, and it does not have effect on MDA, LDA and Logistic regression.

Table 2: Classification Error Rate of LDA, QDA and Logistic Regression

Method	MDA	CART	LDA	QDA	Logistic Regression
Error Rate	0.974	0.9498	.9605	0.9677	0.9693

6 Conclusion

From the results, we can see that MDA outperforms LDA and QDA in the classification and PCA method did not improve the performance of MDA.