Linear Classification Methods and QDA

John Ensley and Songshan Yang

February 19, 2023

1 Introduction of the data set

In this project, we applied the linear classification methods and QDA to a breast cancer data set. This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. There are 699 subjects with tumors in the data set. All of the tumors are classified as one of two different classes — either "Benign" or "Malignant". There are 9 features measured for each tumor, including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell Size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. All of them range in value from 1 to 10.

2 LDA, QDA and Logistic Regression

In the first step, we applied LDA, QDA and logistic regression to the whole data set to examine their classification error rate. We found that there are only 16 subjects that have missing data and we deleted 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. 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 instead compute the covariance matrix for each class and then classify the subjects. For logistic regression, we use the Newton-Raphson algorithm to finish the optimization problem.

We randomly selected approximately half of the data set to use as training data, and used the rest for testing. The classification results are summarized in Table **??**.

From Table **??**, we can see that LDA and QDA have similar classification error rates and logistic regression outperforms the LDA and QDA in the classification.

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

Method	LDA	QDA	Logistic Regression
Error Rate	0.0395	0.0401	0.0307

3 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. The ratio of how much each principle component explains the total variance is summarized in Table **??**. We decided to use the first two components which account for about 74% 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 **??**. Comparing Table **??** and Table **??**, we can see that PCA only improves the performance of QDA, and it does not have an effect on LDA or logistic regression. See Figure **??** for the decision boundary resulting from LDA with dimension reduction.

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

Ratio	0.655	0.086	0.060	0.051	0.042	0.034	0.033	0.029	0.010

Table 3: Classification Error Rate of LDA, QDA and Logistic Regression

Method	LDA	QDA	Logistic Regression
Error Rate	0.0395	0.0337	0.0307

4 Classification Using Cross-validation

In this section we explored the classification accuracy using cross-validation. We use 2, 4, 6, 8 and 10-fold cross-validation to examine the effect of cross-validation on the classification method and also the number of folds on the error rate. The results are summarized in Table **??**. All the results are based on 100 simulations.

From Table **??**, we can see that the performances of LDA, QDA and logistic regression are not changed too much when we use cross-validation and the number of folds does not affect the performances significantly.

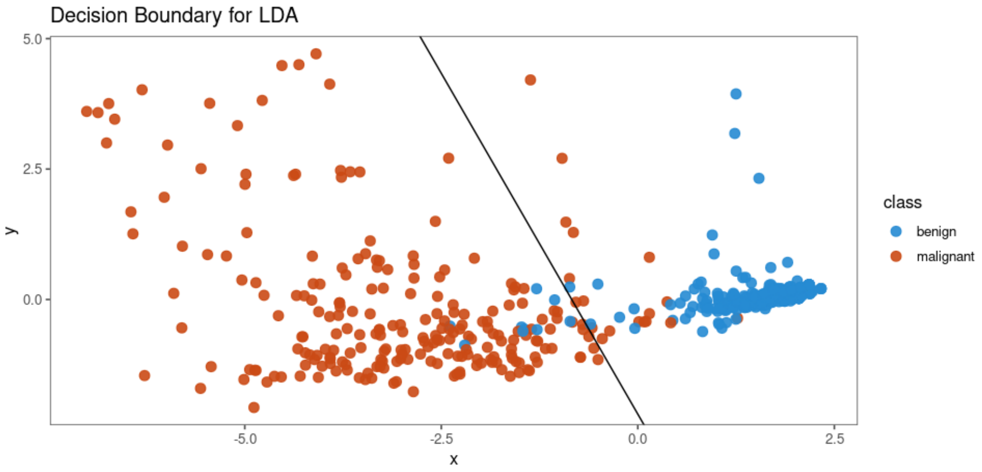

Figure 1: Decision boundary for LDA after reducing to two dimensions using PCA.

Table 4: Classification Error Rate of LDA, QDA and Logistic Regression using cross-validation

		LDA	QDA	Logistic Regression
Error Rate	2	0.0468	0.0454	0.0410
	4	0.0395	0.0512	0.0322
	6	0.0395	0.0468	0.0307
	8	0.0395	0.0483	0.0307
	10	0.0395	0.0483	0.0351

We then evaluate the classification method with reduced dimension using cross-validation. We again use 2, 4, 6, 8 and 10-fold cross-validation to examine the effect of cross-validation on the classification method and also the number of folds on the error rate. The results are summarized in Table **??**. All the results are based on 100 simulations.

Compared Table **??** to Table **??**, the performance of QDA improves but the performance of LDA and logistic regression are not improved. Besides, the number of folds does not have significant effect on the classification methods.

Table 5: Classification Error Rate of LDA, QDA and Logistic Regression using PCA and cross-validation

		LDA	QDA	Logistic Regression
Error Rate	2	0.0380	0.0351	0.0366
	4	0.0380	0.0337	0.0292
	6	0.0395	0.0351	0.0322
	8	0.0395	0.0351	0.0307
	10	0.0395	0.0337	0.0307

5 Roles of Each Team Member

We met up and wrote the code for this project together. Songshan wrote the report and John put together the presentation slides and made the various plots and visualizations.