Linear Classification Methods and QDA

John Ensley & Songshan Yang

February 25, 2016

Penn State University
STAT 557

Data Description

- 699 tumor samples
- 458 classified as benign, 241 as malignant
- 9 observations for each: uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses
- Each predictor is a value from 1 to 10

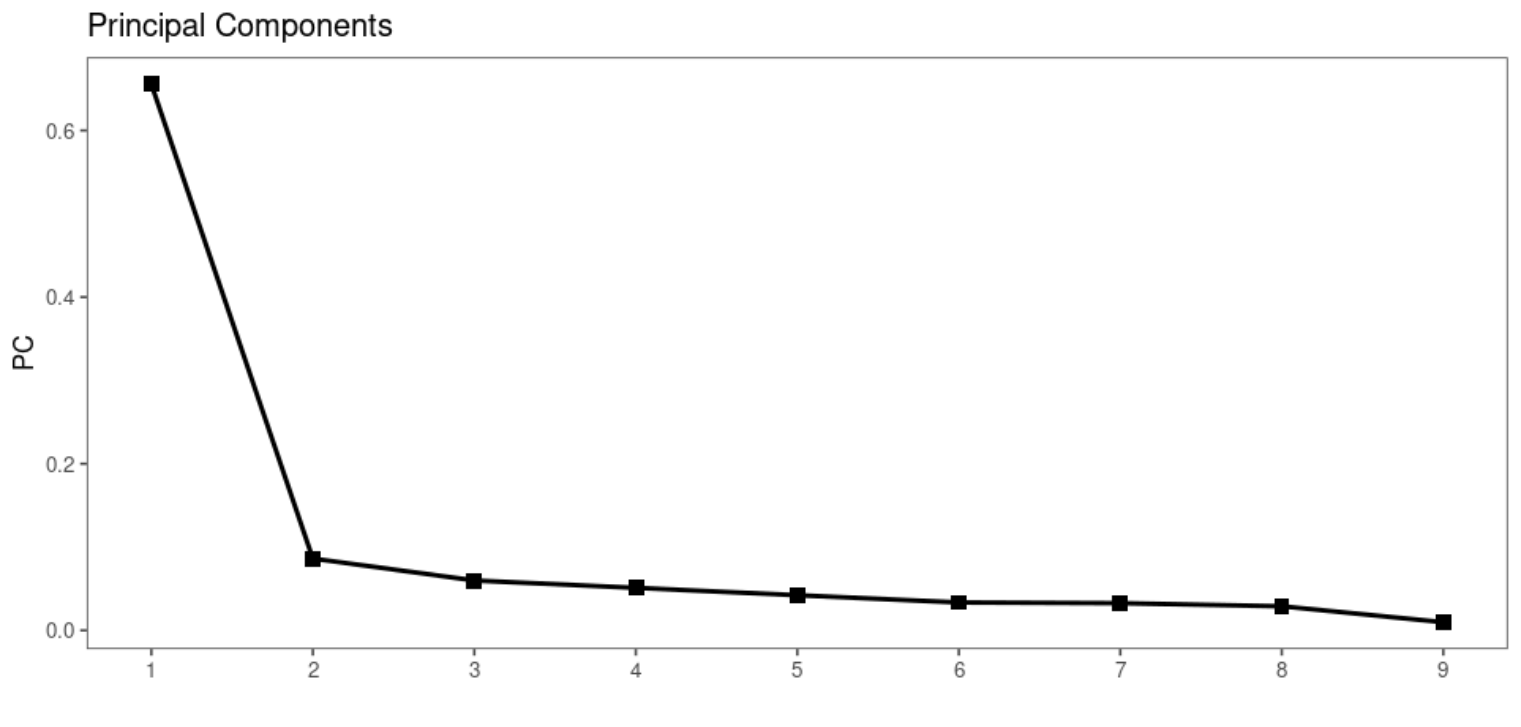
Results, No Dimension Reduction

50% of observations randomly chosen for training data, rest used for testing

Method	LDA	QDA	Logistic Regression
Error Rate	0.0395	0.0401	0.0307

Principal Component Analysis

PCs	0.655	0.086	0.060	0.051	0.042	0.034	0.033	0.029	0.010

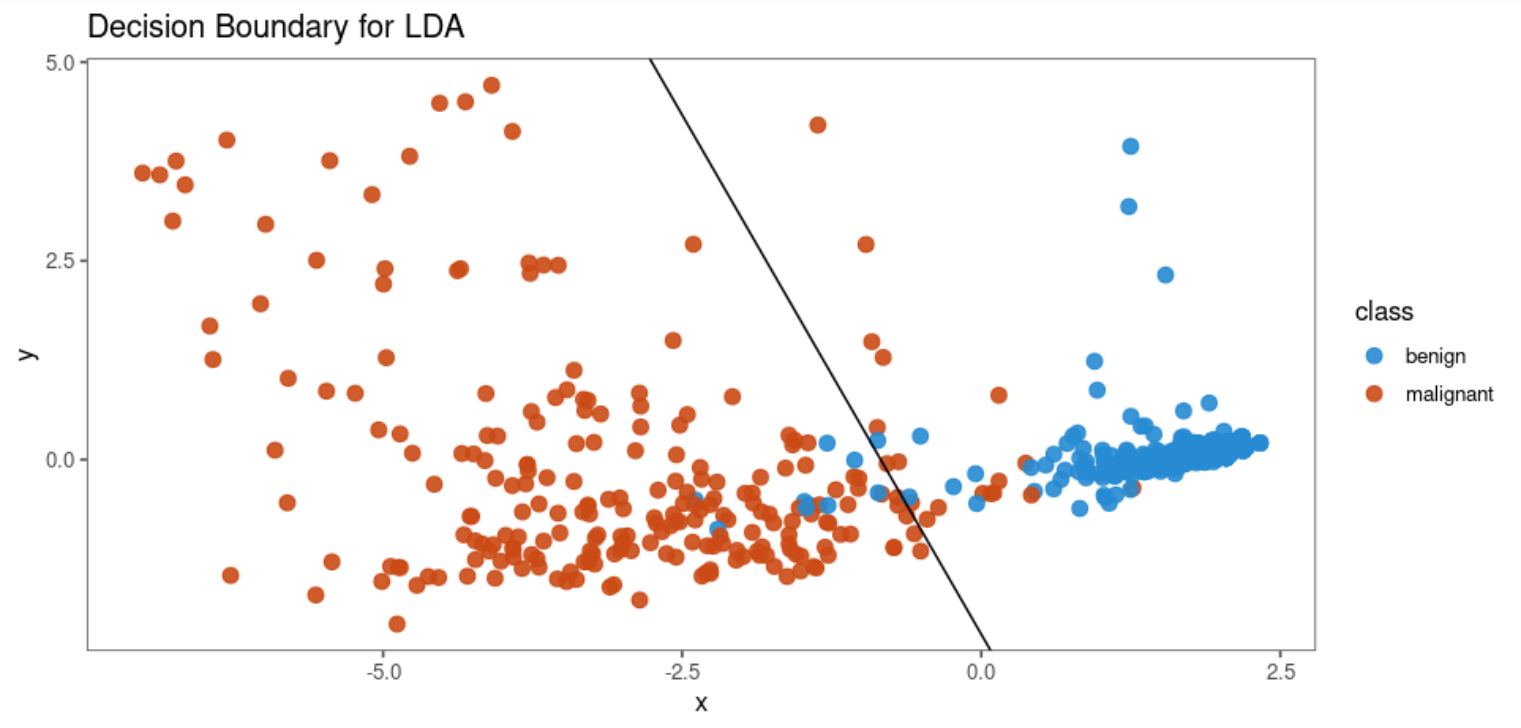


The first 2 principal components make up 74% of the variance

Results, Dimension Reduction

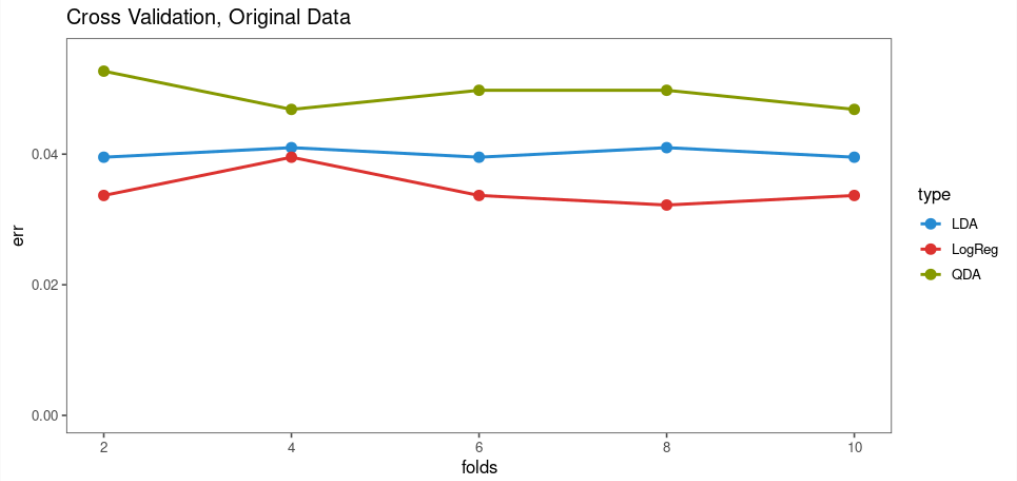
50% of observations randomly chosen for training data, rest used for testing

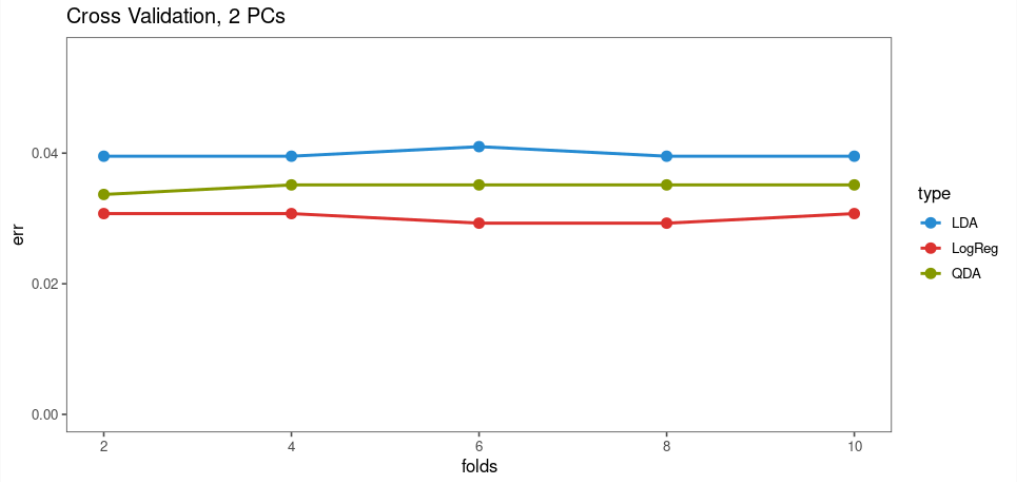
Error Rates	LDA	QDA	Logistic Regression
Original Data	0.0395	0.0401	0.0307
2 PCs	0.0395	0.0337	0.0307



Results, Cross Validation

10-fold CV	LDA	QDA	Logistic Regression
Original Data	0.0395	0.0483	0.0351
2 PCs	0.0395	0.0337	0.0307





Results, Cross Validation

No PCA:

Folds	LDA	QDA	Logistic Regression
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

PCA:

Folds	LDA	QDA	Logistic Regression
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

Conclusions

- All classification methods perform well on this dataset
- Logistic regression performs the best overall
- Using 2 principal components improves the error rates, particularly with QDA