Extensions to LDA and multinomial regression

Patrick Breheny

September 22

Introduction

- Linear discriminant analysis begins by assuming that $\mathbf{x}|G$ follows a multivariate normal distribution with a variance that does not depend on the class G
- Consider now doing away with this assumption and allowing the class-conditional distributions to have unequal variances; denote the variance given class k as $\mathbf{\Sigma}_k$

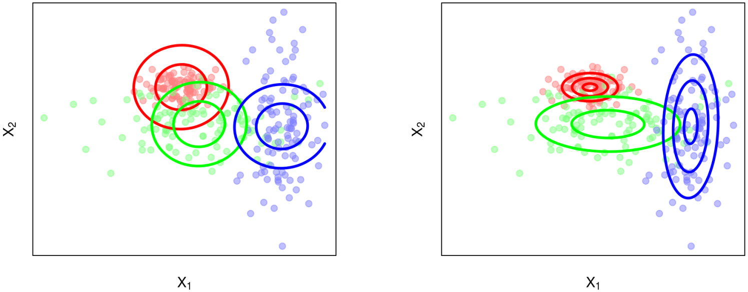
Quadratic discriminant analysis

- Without the equal variance assumption, the convenient cancellations of LDA do not occur:

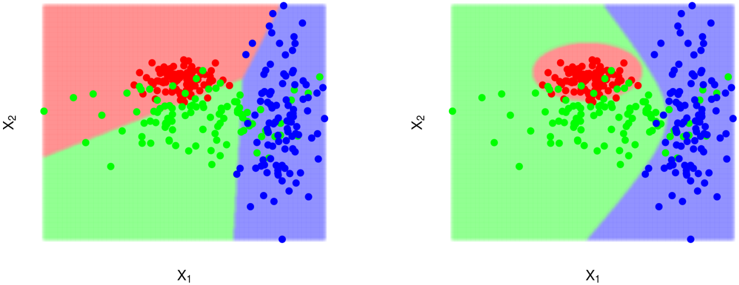
  $$\delta_k(\mathbf{x}) = \log \pi_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k);$$

  in particular, note that the quadratic terms remain
- The discriminant functions and decision boundaries therefore become quadratic in $\mathbf{x}$
- This procedure is therefore called *quadratic discriminant analysis* (QDA)

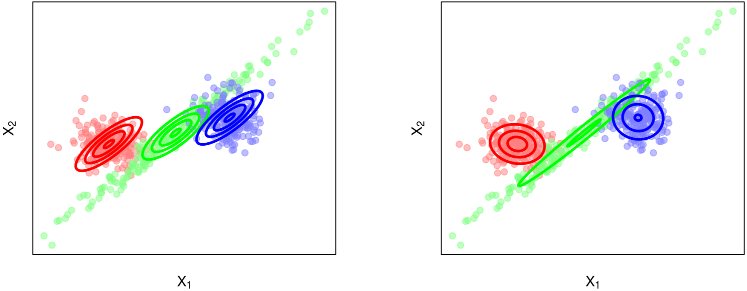
Example 1



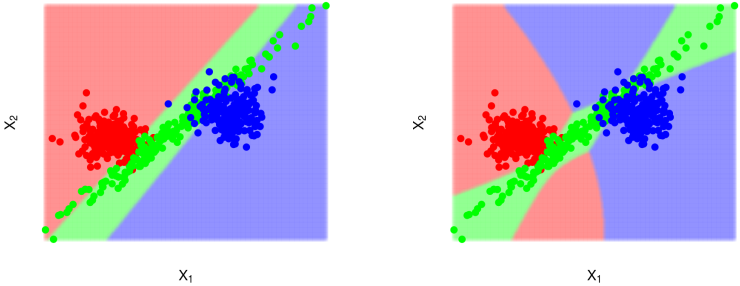
Example 1



Example 2



Example 2



R/SAS

- QDA is implemented and easily run in both SAS and R
- In SAS:

```
PROC DISCRIM DATA=Iris POOL=NO;
  CLASS Species;
RUN;
```

- In R (again requiring the MASS package):

```
fit <- qda(Species~.,Data)
```

Car silhouette data

- For the iris data set, whether we use LDA or QDA makes no difference: we end up with the same predictions either way (the probabilities differ, of course, but the class with the highest probability does not)
- Of course, this is not always the case
- An example of a data set where QDA performs extremely well is the Vehicle Silhouettes data set, in which, based on computer images of the silhouettes of vehicles, the model is made to predict between four vehicles: a bus, a van, and two cars (a Saab and an Opel)

Vehicle silhouette results

Leave-one-out cross validation results for LDA and QDA on the vehicle silhouette data:

	LDA			
	bus	opel	saab	van
bus	209	8	11	2
opel	4	130	65	3
saab	2	67	128	2
van	3	7	13	192

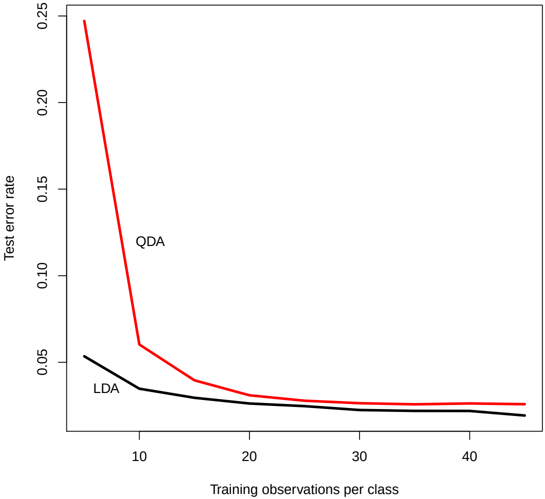
	QDA			
	bus	opel	saab	van
bus	214	0	2	1
opel	0	160	54	2
saab	0	43	154	0
van	4	9	7	196

Overall CV error rates: 22.1% for LDA, 14.4% for QDA

QDA on the small-sample iris data

- Of course, QDA does not always outperform LDA; in fact it often does quite a bit worse
- For example, let's compare LDA with QDA using the iris data in the same manner as our earlier comparison involving multinomial regression (randomly choose 5 observations per class used as training data, the rest to test the fit of the model)
- QDA has a test error rate of 24.7%; far worse than LDA's 5.2% and multinomial regression's 7.7%

Iris cross-validation comparison



Remarks

- Why does QDA do so poorly with $n = 5$ per class?
- Consider the number of additional parameters that have to be estimated by QDA: LDA must estimate one 4×4 covariance matrix (with 10 distinct parameters), whereas QDA must estimate three such matrices
- Thus, QDA must estimate 20 additional parameters, which is a lot to ask with a sample size of 15
- As usual, the tradeoff between LDA or QDA is one of bias and variance – LDA makes stronger assumptions and obtains estimates with lower variance, but has the potential for biased decision boundaries if heterogeneity is truly present

Ridge/lasso penalized multinomial regression

- As a final topic for this section of the course, we will take a brief tour of some of the ways in which penalization and regularization can be applied to the methods of multinomial regression and discriminant analysis
- Extending ridge and lasso to logistic and multinomial regression is fairly straightforward; the lasso estimate $\widehat{\boldsymbol{\beta}}$ would be minimize

$$Q(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{k=1}^{K}\sum_{i|g_i=k}\log\pi_{ik} + \sum_{k=1}^{K}\sum_{j=1}^{p}|\beta_{kj}|,$$

where

$$\pi_{ik} = \frac{\exp(\eta_{ki})}{\sum_l \exp(\eta_{li})}$$

and $\eta_{ki} = \mathbf{x}_i^T\boldsymbol{\beta}_k$

Penalization and identifiability

- The ridge estimates would of course be defined similarly, albeit with β_{ki}^2 replacing $|\beta_{kj}|$
- For the most part, the extensions are straightforward; the only noticeable change is with regard to the notion of a reference category
- In traditional multinomial regression, one category must be chosen as a reference group (*i.e.*, have $\boldsymbol{\beta}_k$ set to $\mathbf{0}$) or else the problem is not identifiable

Penalization and identifiability (cont'd)

- With penalized regression, this restriction is not necessary
- For example, suppose $K = 2$; then $\beta_{1j} = 1, \beta_{2j} = -1$ produces the exact same $\{\pi_{ik}\}$ as $\beta_{1j} = 2, \beta_{2j} = 0$
- As it is impossible to tell the two models apart (and an infinite range of other models), we cannot estimate $\{\boldsymbol{\beta}_k\}$

Penalization and identifiability (cont'd)

- With, say, a ridge penalty, this is no longer the case, as $\sum_k \beta_{jk}^2 = 2$ in the first situation and 4 in the second; the proper estimate is clear
- The same holds for the lasso penalty, although of course there is now the possibility of sparsity, perhaps with respect to multiple classes
- For example, with $\lambda = 0.1$, the coefficient vector for petal width in the iris data is $(0, 0, 1.57)$
- In words, an increase of one cm in petal width increases the log odds of *virginica* by 1.57 with respect to both *setosa* and *versicolor* (those two species serving, in effect, as a common reference group)

Regularized discriminant analysis

- The notion of regularization can also be extended to discriminant analysis
- For example, why commit ourselves to LDA or QDA when we could consider the range of intermediary estimates given by

$$\hat{\mathbf{\Sigma}}_k(\alpha) = \alpha\hat{\mathbf{\Sigma}}_k + (1-\alpha)\hat{\mathbf{\Sigma}}$$

- With $\alpha = 0$ we have LDA and with $\alpha = 1$ we have QDA; in between we have estimates that allow for class-specific variances, but with shrinkage toward the common covariance matrix

Regularized discriminant analysis (cont'd)

- Furthermore, one could consider regularizing $\hat{\mathbf{\Sigma}}$ itself in a ridge-like manner:

$$\hat{\mathbf{\Sigma}}(\gamma) = \gamma\hat{\mathbf{\Sigma}} + (1-\gamma)\mathbf{I}$$

- Recall that the LDA/QDA estimates rely on $\hat{\mathbf{\Sigma}}^{-1}$; thus, the addition of a ridge down the diagonal is perhaps a very good idea, as it stabilizes this inverse
- Furthermore, it ensures that the inverse always exists

Shrunken centroids

- As final example of regularization, we could apply a lasso-type penalty to the means $\{\boldsymbol{\mu}_k\}$
- This would have the advantage of parsimony (a smaller number of variables would actually be involved in the model) as well as reduction of variance, especially in settings where p is large
- The means $\{\boldsymbol{\mu}_k\}$ are sometimes called the *centroids* of the classes; the method described on this slide is therefore called *nearest shrunken centroids*