Linear Discriminant Analysis, Part II

Patrick Breheny

September 20

Anderson's Iris Data

- To illustrate the application of LDA to a real data set, we will use a famous data set collected by Anderson and published in "The irises of the Gaspé Peninsula", and which originally inspired Fisher to develop LDA
- Anderson collected and measured hundreds of irises in an effort to study variation between and among the different species
- There are 260 species of iris; this data set focuses of three of them (*Iris setosa*, *Iris virginica* and *Iris versicolor*)
- Four features were measured on 50 samples for each species: sepal width, sepal length, petal width, and petal length

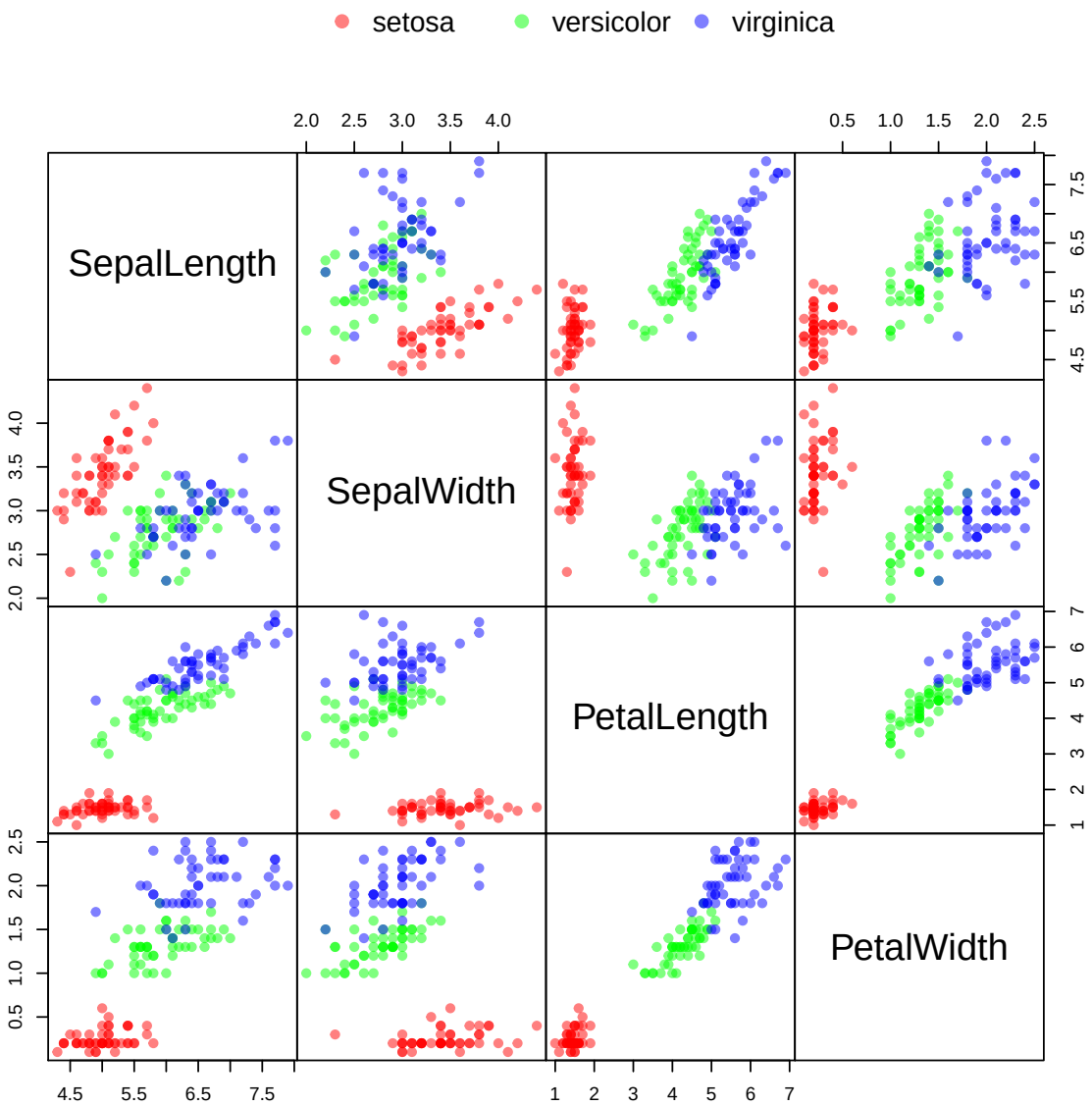
Iris species

(a) setosa

(b) virginica

(c) versicolor

Scatterplot matrix



LDA in SAS/R

- Fitting LDA models in SAS/R is straightforward
- SAS code:

```
PROC DISCRIM DATA=iris;
  CLASS Species;
RUN;
```

- R code (requires the MASS package):

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

Confusion matrix

The cross-classification table of predicted and actual species assignments (sometimes called the *confusion matrix*):

		Actual		
		setosa	versicolor	virginica
Predicted	setosa	50	0	0
	versicolor	0	48	1
	virginica	0	2	49

Mahalanobis distance

- The "distance" between classes k and l can be quantified using the *Mahalanobis distance*:

$$\Delta = \sqrt{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)},$$

- Essentially, this is a scale-invariant version of how far apart the means, and which also adjusts for the correlation between variables
- The result is a multivariate extension of the notion of "how many standard deviations apart are X and Y"?

Mahalanobis distance

	setosa	versicolor	virginica
setosa	0.00	9.48	13.39
versicolor	9.48	0.00	4.15
virginica	13.39	4.15	0.00

These distances are rather large; hence the ease with which LDA was able to classify the species

Prediction

- An important feature of LDA is the ability to estimate the conditional probability of the class given the identifying features
- This is valuable in two distinct situations:
  - To predict future classes
  - To illustrate the model and the relationship of the explanatory variables to the outcome
- For example, suppose we only had five observations per species; would that be enough to build an accurate classifier?

Making predictions in SAS/R

- To explore this, let's split our sample randomly into a *training set* used to fit the model, and a *test set* we can use to see how well our model predicts new observations
- Once this is done, it is straightforward in both SAS and R to make predictions on a new set of data:

```
PROC DISCRIM DATA=Train TESTDATA=Test TESTOUT=Pred;
  CLASS Species;
RUN;
```

- Or in R:

```
fit <- lda(Species~.,Train)
pred <- predict(fit,Test)
```

Prediction results

Results from one such test/train split:

		Actual		
		setosa	versicolor	virginica
Predicted	setosa	45	0	0
	versicolor	0	42	4
	virginica	0	3	41

The misclassification error goes up slightly, but the differences between the species are big enough that we have a rather good classifier even with only 5 observations per class

Multinomial logistic regression

- If you are familiar with multinomial logistic regression, you may be thinking to yourself: what's the big deal? I already have a perfectly good tool for dealing with this problem
- To refresh your memory, the multinomial logistic regression model consists of defining one class to be the reference and fitting separate logistic regression models for $k = 2, \ldots, K$, comparing each outcome to the baseline:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \beta_{k0} + \mathbf{x}_i^T \boldsymbol{\beta}_k$$

where π_{ik} denotes the probability that the ith individual's outcome belongs to the kth class

LDA = logistic regression?

- Recall, however, that LDA satisfies:

$$\begin{aligned}\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) &= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) \\ &\quad + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l) \\ &= \alpha_{k0} + \mathbf{x}_i^T\boldsymbol{\alpha}_k\end{aligned}$$

- At first glance, then, it seems the models are the same

Difference between LDA and logistic regression

- However, although the two approaches have the same form, they do not estimate their coefficients in the same manner
- LDA operates by maximizing the log-likelihood based on an assumption of normality and homogeneity
- Logistic regression, on the other hand, makes no assumption about $\Pr(\mathbf{X})$, and estimates the parameters of $\Pr(G|\mathbf{x})$ by maximizing the conditional likelihood

Difference between LDA and logistic regression (cont'd)

- Intuitively, it would seem that if the distribution of $\mathbf{x}$ is indeed multivariate normal, then we will be able to estimate our coefficients more efficiently by making use of that information
- On the other hand, logistic regression would presumably be more robust if LDA's distributional assumptions are violated
- Indeed, this intuition is borne out, both by theoretical work and simulation studies, although in practice, the two approaches do usually give similar results

Iris data comparison

- For the iris data, multinomial logistic regression classifies the data even better (slightly) than LDA:

		Actual		
		setosa	versicolor	virginica
Predicted	setosa	50	0	0
	versicolor	0	49	1
	virginica	0	1	49

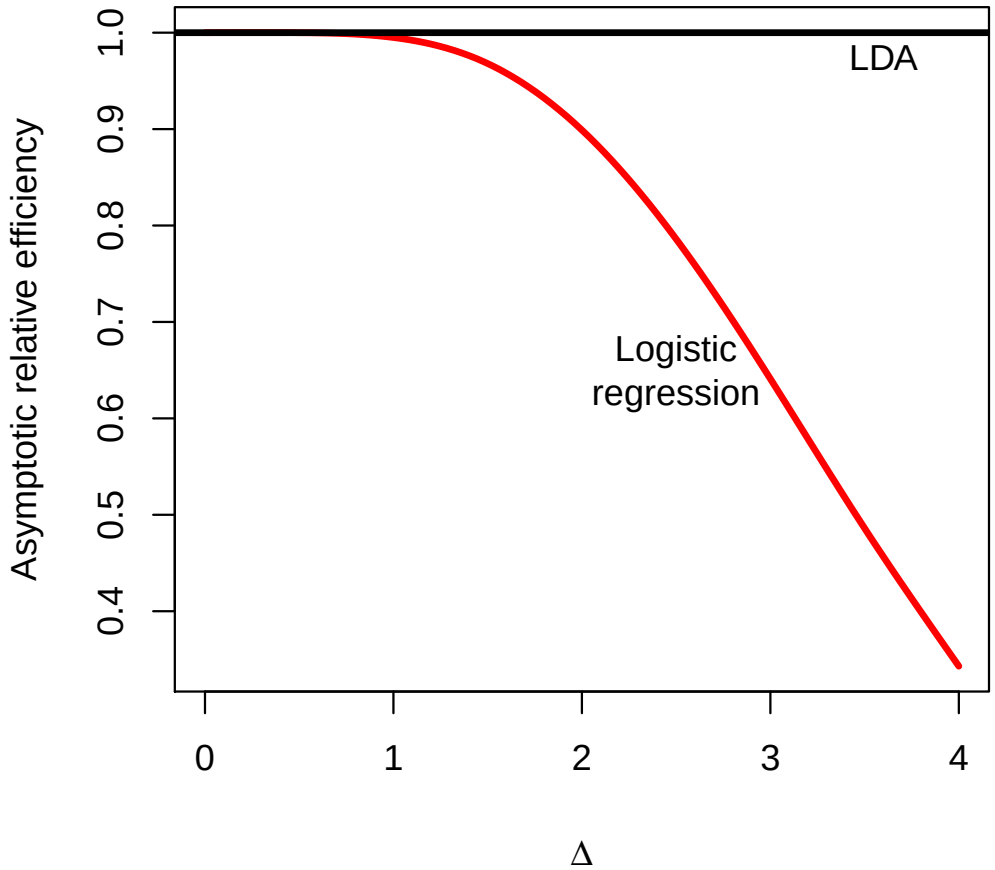
- However, this is not convincing; what matters is the ability to predict observations that the model doesn't already know the answers for

Iris cross-validation

- Consider a cross-validation study with the iris data, randomly splitting it up into a training set containing 5 observations per species, with the remainder used as a test set
- The results: LDA has a misclassification rate of 5.2%, while logistic regression has a misclassification rate of 7.7%

Asymptotic results

Efron (1975) derived the asymptotic relative efficiency of logistic regression compared to LDA in the two-class case when the true distribution of $\mathbf{x}$ is normal and homogeneous, and found the logistic regression estimates to be considerably more variable:



Final remarks

- Recall the problem of complete separation in logistic regression: when there is no overlap between the classes, the logistic regression MLEs go to $\pm\infty$
- This does not happen with LDA, however: estimates are always well-defined and finite
- In principle, LDA should perform poorly when outliers are present, as these usually cause problems when assuming normality
- In practice, however, the two approaches usually give similar results, even in cases where $\mathbf{x}$ is obviously not normal (such as for categorical explanatory variables)