UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Lecture 10 Linear Discriminant Analysis II

STAT 441/505: Applied Statistical Methods in Data Mining

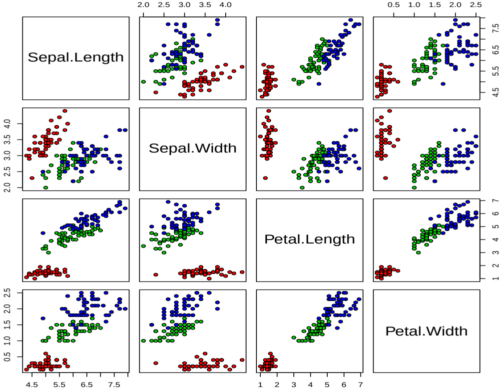
Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

Outline

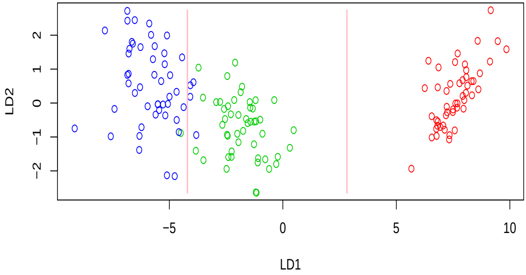
Fisher's Iris Data



- 4 variables, 3 species, and 50 samples per class
- Blue - Setosa, Orange - Versicolor, and Green Virginica
- LDA classifies all but 3 of the 150 training samples correctly.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Fisher's Iris Data



- When there are K classes, linear discriminant analysis can be viewed exactly in a $K - 1$ dimensional plot.
- Why? Because it essentially classifies to the closest centroid, and they span a $K - 1$ dimensional plane.
- Even when $K > 3$, we can find the best 2-dimensional plane for vizualizing the discriminant rule.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

From discriminant rule to probabilities

- Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

- So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.
- When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = k|X = x) > 0.5$ or else to class 1.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

LDA on credit data

```
> table(default.pred$class,defaultData$default)

        No   Yes
  No  9645   254
  Yes   22    79
> 22/9667
[1] 0.002275784
> 254/333
[1] 0.7627628
```

- $(22 + 254)/10000$ errors — 2.76% misclassification rate!
- However, this is training error, and we may be over fitting. Not a big concern here since $n = 10000$ and $p = 3$.
- If we classified to the prior — always to class `No` in this case — we would make $333/10000 = 3.33\%$ errors.
- Of the true `No` 's, we make $22/9667 = 0.2\%$ errors, of the true `Yes` 's, we make $254/333 = 76.3\%$ errors.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Types of errors

- False positive rate: The fraction of negative examples that are classified as positive 0.2% in example.
- False negative rate: The fraction of positive examples that are classified as negative 76.3% in example.
- We produced this table by classifying to class `Yes` if

$$\widehat{\Pr}(\texttt{Default=Yes}|\texttt{Balance, Incoming, Student}) \geq 0.5.$$

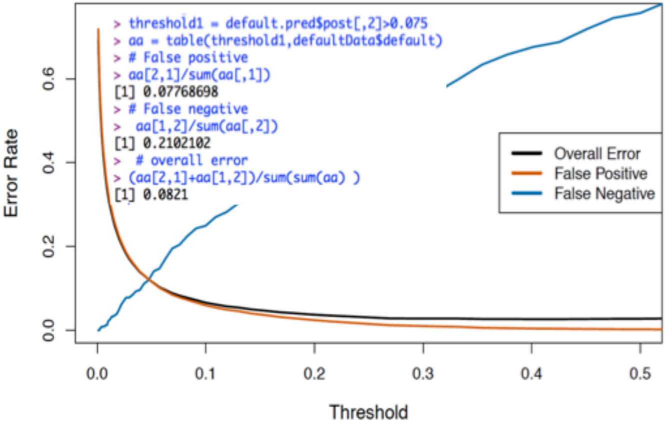
- We can change the two error rates by changing the threshold from 0.5 to some other value in $[0, 1]$:

$$\widehat{\Pr}(\texttt{Default=Yes}|\texttt{Balance, Incoming, Student}) \geq \textit{threshold}.$$

and vary *threshold*.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Varying the threshold



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

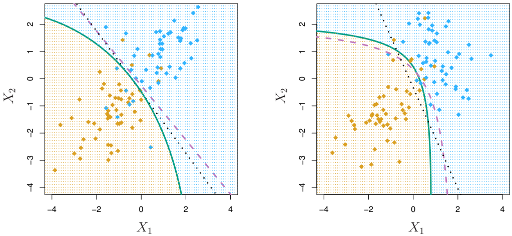
Other forms of Discriminant Analysis

- When $f_k(x)$ are Gaussian densities, with the same covariance matrix Σ in each class, the Bayes Theorem

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)},$$

  this leads to linear discriminant analysis.
- By altering the forms for $f_k(x)$, we get different classifiers.
  - With Gaussians but different Σ_k in each class, we get quadratic discriminant analysis.
  - With $f_k(x) = \Pi_{j=1}^p f_{jk}(x_j)$ (conditional independence model) in each class we get naive Bayes. For Gaussian this means the Σ_k are diagonal.
  - Many other forms, by proposing specific density models for $f_k(x)$, including nonparametric approaches.

Quadratic Discriminant Analysis



- As in the following Σ_k are different, so in QDA quadratic term matters

$$\delta_k(x) = \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k).$$

- Black dotted: LDA boundary; Purple dashed: Bayes' boundary; Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

QDA versus LDA

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic.
- QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances.
- LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Logistic Regression versus LDA

- For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p(x).$$

- So it has the same form as logistic regression. The difference is in how the parameters are estimated.
- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (aka discriminative learning).
- LDA uses the full likelihood based on $\Pr(Y|X)$ (aka generative learning).
- Despite these difference, in practice the results are often very similar.

Summary and Remark

- Linear Discriminant Analysis for $p > 1$
- From discriminant rule to probabilities
- Quadratic Discriminant Analysis
- Read textbook Chapter 4 and `R` code
- Do `R` lab