Classification: Linear Discriminant Analysis, ROC Curve, and Quadratic Discriminant Analysis

(ISLR 4.4 - 4.5)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

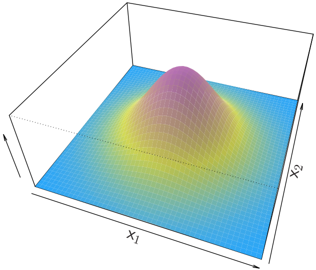
LDA with $p > 1$

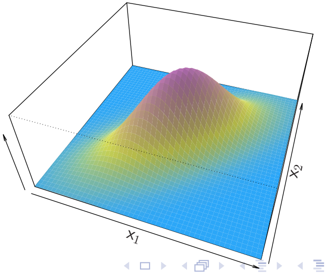
Suppose we have p (continuous) parameters $X = (X_1, X_2, \ldots, X_p)^T$, then under LDA, they are assumed to have multivariate normal distribution:

$$X \sim \mathrm{N}_p(\mu, \Sigma) \Longleftrightarrow f(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}$$

- Each class has its own μ_k.
- The covariance Σ is the same across classes.

Correlation between X_1, X_2: 0 (left) vs. 0.7 (right)





Linear discriminants

- The discriminant function is a linear combination of $X_1, \ldots, X_p$:

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

- When $K = 2$, `R` reports the coefficients of one linear discriminants.

$$\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

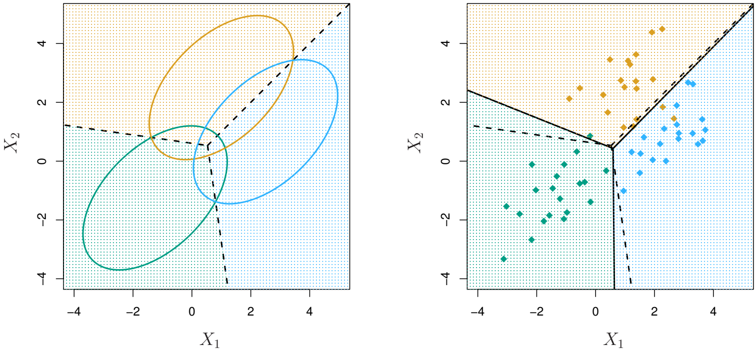
If this value is large, classify to one class; if small, the other class.

- Observation X will be classified to class 1 if

$$0 < \delta_1(X) - \delta_2(X) = C + \underbrace{(\mu_1 - \mu_2)^T \Sigma^{-1}}_{\text{proportional to } \beta_1, \ldots, \beta_p} X$$

An example: $p = 2, K = 3$

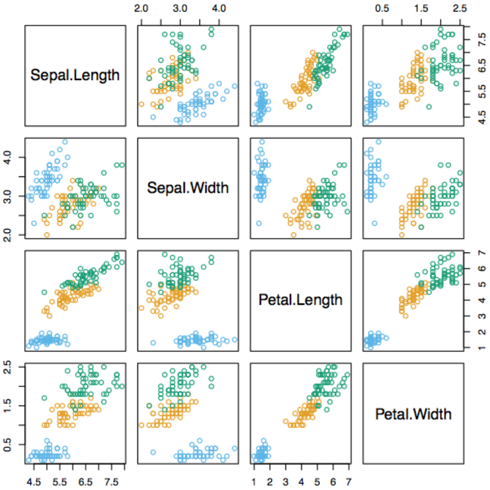
In the training data, $n_1 = n_2 = n_3 = 20$.



- Solid lines: LDA decision boundaries (error rate: 0.0770).
- Dashed lines: Bayes decision boundaries (error rate: 0.0746).

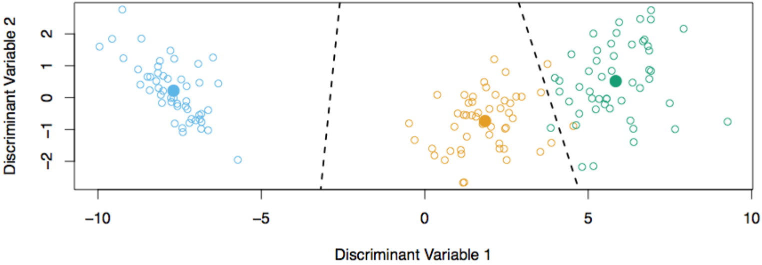
Fisher's iris data

$K = 3$: Setosa, Versicolor, Virginica, $p = 4$, $n_1 = n_2 = n_3 = 50$



Fisher's discriminant plot

LDA classifies all but 3 of the 150 training samples correctly.



When there are K classes, LDA can be viewed exactly in a $K-1$ dimensional plot.

Posterior probabilities

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

$$\hat{P}(Y = k \mid X) = \frac{e^{\hat{\delta}_k(X)}}{\sum_{j=1}^{K} e^{\hat{\delta}_j(X)}}$$

- In classification, the class with the largest $\hat{\delta}_k(X)$ also has the largest posterior probability.
- When $K = 2$, we classify to class k if

$$\hat{P}(Y = k \mid X) \geq 0.5$$

LDA on Credit data

```
lda(default ~ balance + student, data = Default);
```

We classify someone to $Y = 1$ if the predicted probability of default ≥ 0.5.

The *confusion matrix*

		True Default No	True Default Yes	Total
Predicted Default	No	9644	252	9896
	Yes	23	81	104
Total		9667	333	10000

- Overall error rate $\frac{23+252}{10000} = 2.75\%$ (training error rate).
- True proportion of default $\frac{333}{10000} = 3.33\%$. This is the error rate of a useless classifier that always predicts no default.
- Misclassification rate among who actually default $\frac{252}{333} = 75.68\%$!

Choosing a threshold

- Previously, we use the default threshold $\hat{P}(Y = 1 \mid X) \geq 0.5$, which misclassifies many Yes's.
- Now we lower the threshold $\hat{P}(Y = 1 \mid X) \geq 0.2$.

		True Default		Total
		No	Yes	
Predicted Default	No	9432	138	9570
	Yes	235	195	430
Total		9667	333	10000

- Overall error rate increases to $\frac{235+138}{10000} = 3.73\%$.
- Misclassification rate among who actually default $\frac{138}{333} = 41.4\%$.
- The trade-off in modifying the threshold: to credit card companies, false negatives may be more dangerous than false positives.

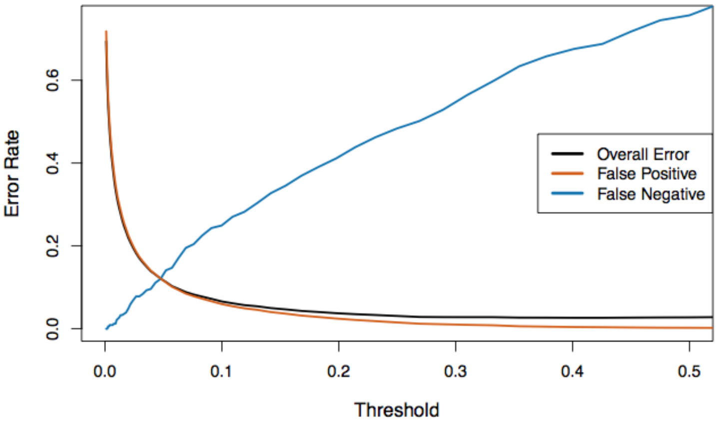
Types of errors

		True class $-$	True class $+$	Total
Predicted class	$-$	TN	FN (Type II error)	N^*
	$+$	FP (Type I error)	TP	P^*
Total		N	P	

- False positive rate: FP/N.
  - Type I error rate
  - $1-$ specificity
- True positive rate: TP/P
  - Power, i.e., $1-$ type II error rate
  - Sensitivity
- Depending on the specific problem, controlling type I errors (or type II errors) may be more important.

Visualization: choosing a threshold

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



Receiver operating characteristics (ROC) curve

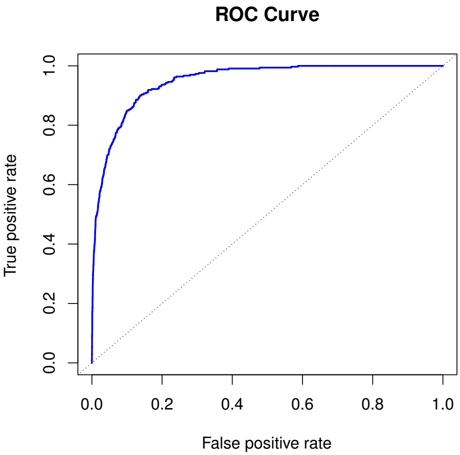
ROC curve

- Y-axis: true positive rate, sensitivity
- X-axis: false positive rate, $1-$ specificity

Area under the curve (AUC)

- The closer to 1 the better.
- Flipping a fair coin (dotted line): AUC $= 0.5$
- Let Z_k be a random variable from the distribution $f_k(X)$ for $k = 0, 1$, then

$$\text{AUC} = P(Z_0 < Z_1)$$



Quadratic discriminant analysis (QDA)

- In class k, X is assumed to have multivariate normal distribution:

$$X \sim \mathsf{N}_p(\mu_k, \Sigma_k)$$

  - Each class has its own mean μ_k.
  - Unlike LDA, QDA assumes that each class has its own covariance matrix Σ_k.

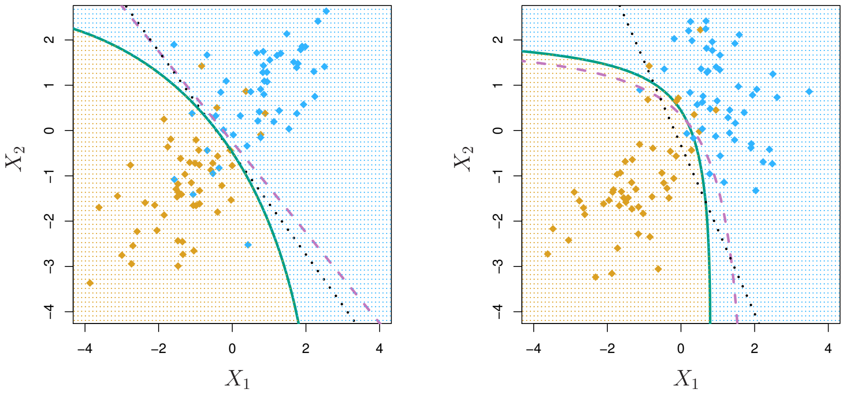
- The induced discriminant function is quadratic in $X_1, \ldots, X_p$.

$$\delta_k(X) = \underbrace{\log \pi_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \log |\Sigma_k|}_{\text{constant}} + \underbrace{\mu_k^T \Sigma_k^{-1} X}_{\text{linear}} - \underbrace{\frac{1}{2} X^T \Sigma_k^{-1} X}_{\text{quadratic}}$$

- More parameters to estimate.

An example: $K = 2, p = 2$

Decision boundaries: QDA, LDA, Bayes



Left: $\Sigma_1 = \Sigma_2$ Right: $\Sigma_1 \neq \Sigma_2$

QDA vs. LDA

- QDA is more flexible
- QDA will work better when
  - the variances are very different between classes, and
  - we have enough observations to accurately estimate the variances
- LDA will work better when
  - the variances are similar among classes, or
  - we don't have enough data to accurately estimate the variances

LDA vs. the logistic regression

- Similarity: both have linear decision boundaries.
  In logistic regression, when $K = 2$,

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- LDA will work better when its assumptions is satisfied
  - Observations X have normal distributions
  - There is a common variance (or covariance) across classes
- Logistic regression doesn't have the normality assumption. It outperforms LDA when normality doesn't hold.

KNN vs. the parametric methods

- KNN is a non-parametric method: No assumptions are made about the shape of the decision boundary!
- Advantage of KNN
  - It dominates both LDA and logistic regression when the decision boundary is highly non-linear
  - It also dominates QDA when the decision boundary is non-quadratic
- Disadvantage of KNN
  - It does not tell us which predictors are important (no table of coefficients)
  - Its performance varies with the choice of K (number of neighbors)