HKU BUSINESS SCHOOL
港大經管學院

Business Statistics

Discriminant Analysis

Weichen Wang

Assistant Professor
Innovation and Information Management

ISLR Chapter 4.4-4.5

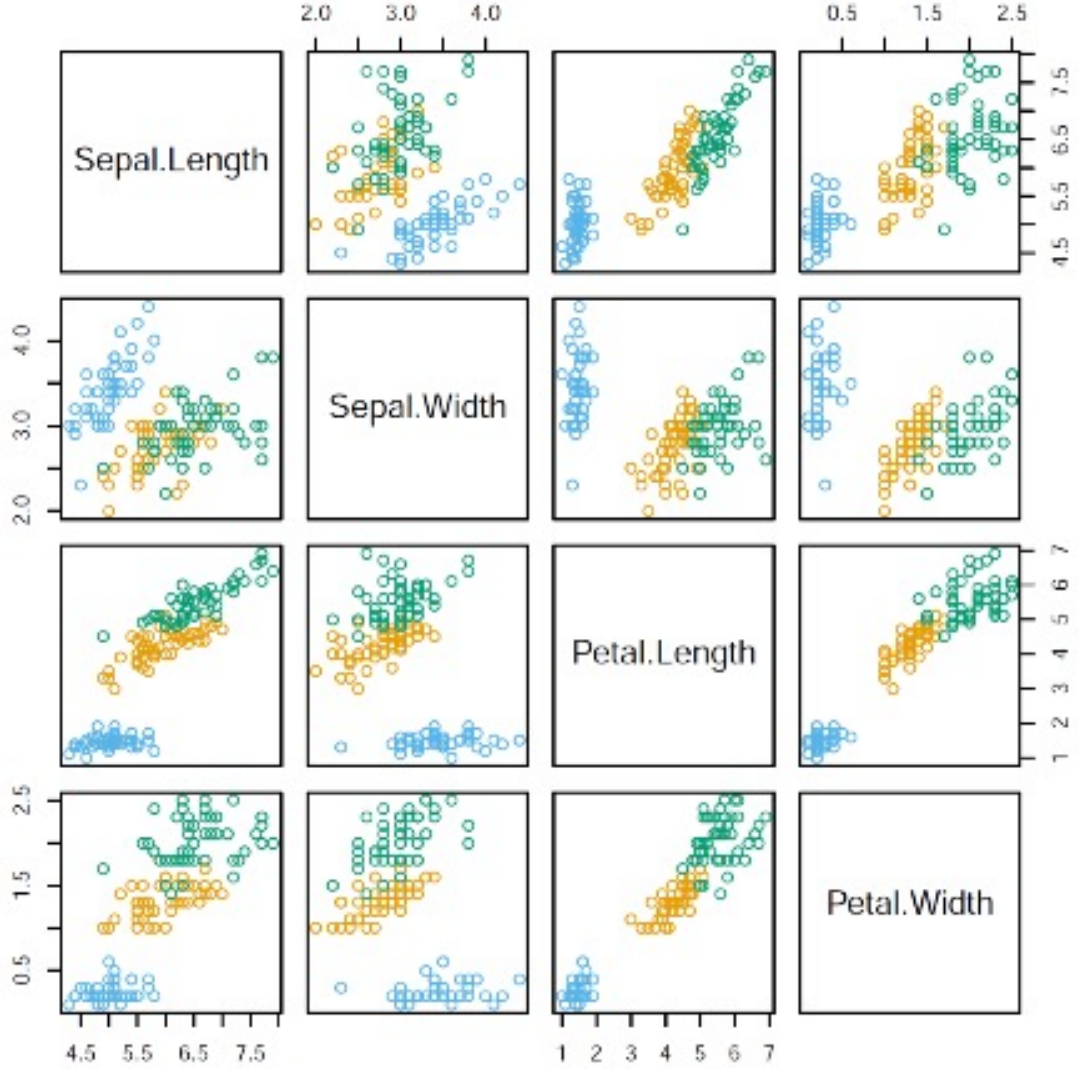
Fisher's Iris Data

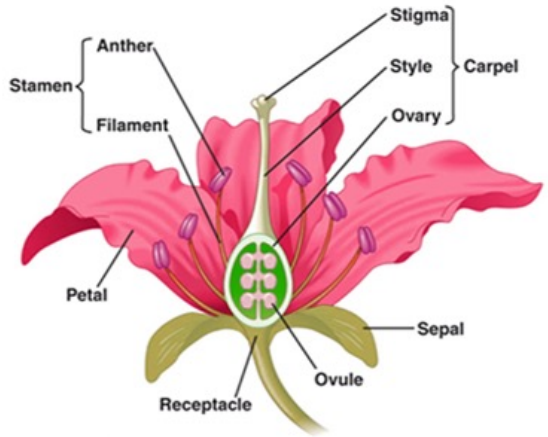
- 4 variables, 50 samples, 3 species (or classes)

Iris setosa

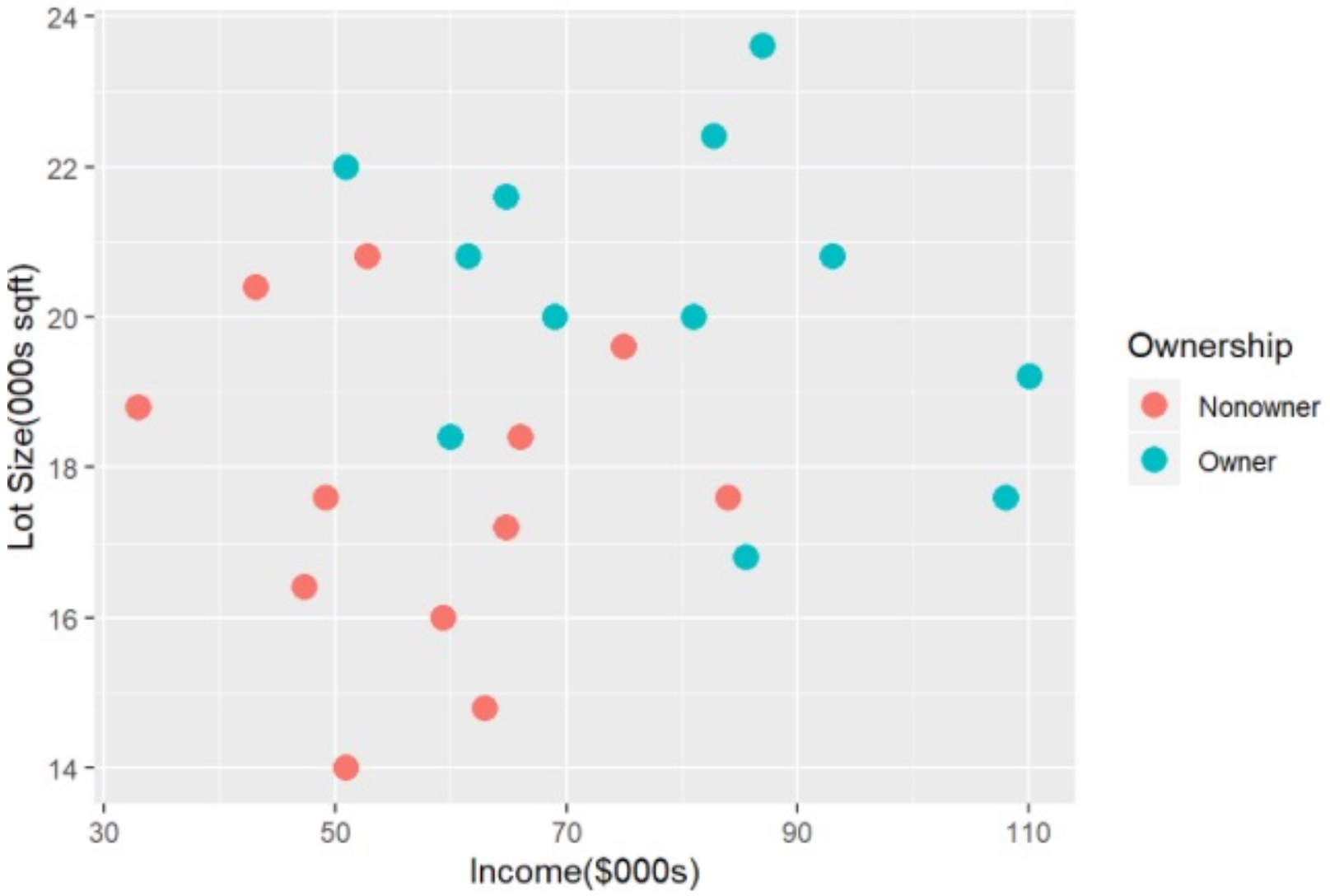
Iris versicolor

Iris virginica

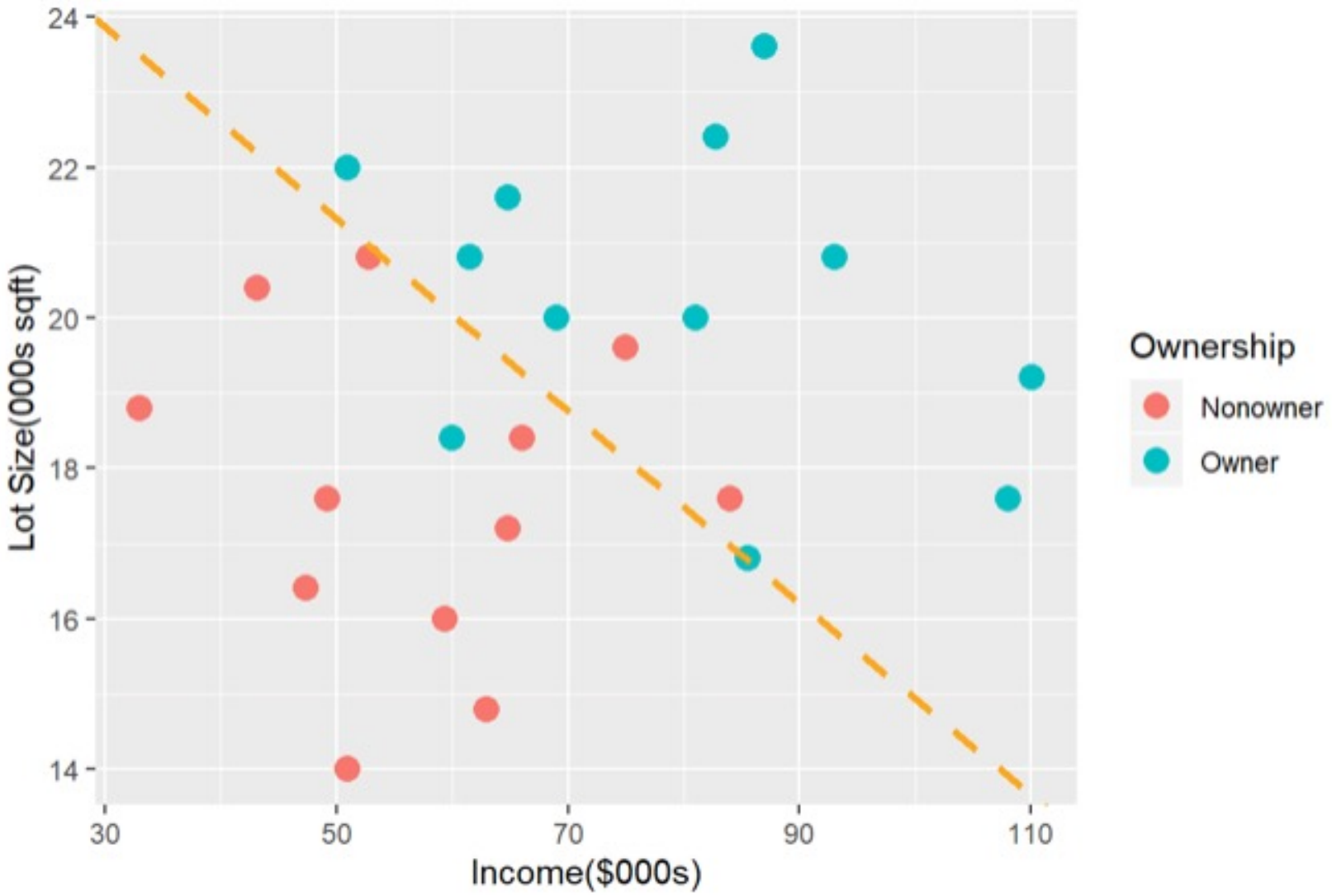




Riding Mowers

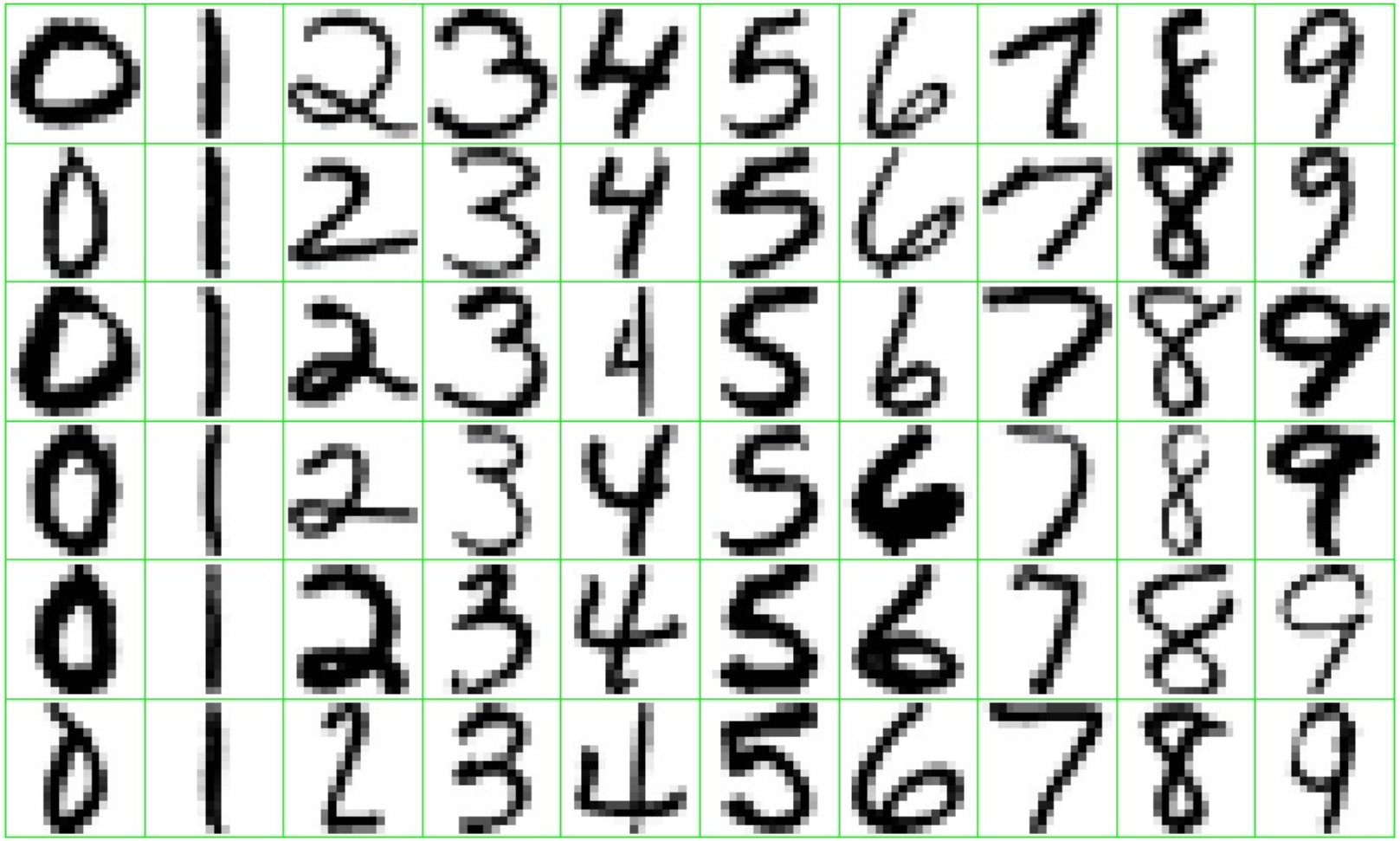


Riding Mowers



Universal Bank Personal Loan Acceptance

Handwritten Digits



Discriminant Analysis

- Identify a set of variables that best discriminate between groups (usually >2)
- Choose a separating line that
  - maximizes the similarity between members of the same group
  - minimizes the similarity between members belonging to different groups

Discriminant Analysis

- The approach is to model the distribution of X in each of the classes separately, and then use Bayes Theorem to flip things around and obtain $\Pr(Y|X)$.
- When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.
- This approach is quite general, so other distributions can be used as well.
- We will focus on normal distributions.

Bayes Theorem for Classification

Thomas Bayes was a famous mathematician whose name represents a big subfield of statistical and probablistic modeling. Here we focus on a simple result, known as Bayes theorem:

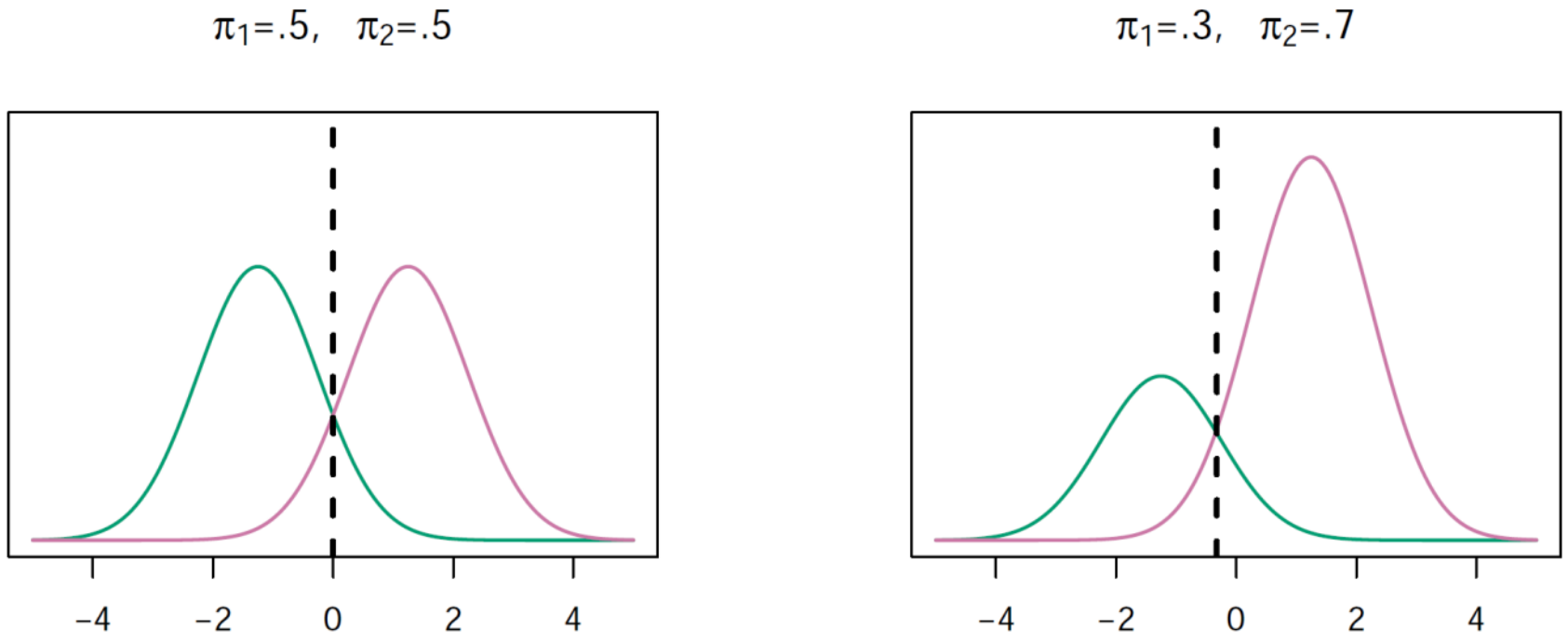
$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

One writes this slightly differently for discriminant analysis:

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

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.

Classify to the Highest Density



- We classify a new point according to which density is highest.
- When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$.
- On the right, we favor the pink class -- the decision boundary has shifted to the left.

Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here μ_k is the mean, and σ_k^2 the variance (in class k). We will assume that all the $\sigma_k = \sigma$ are the same.

Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \Pr(Y = k|X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Happily, there are simplifications and cancellations.

Discriminant Functions

To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest *discriminant score*:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

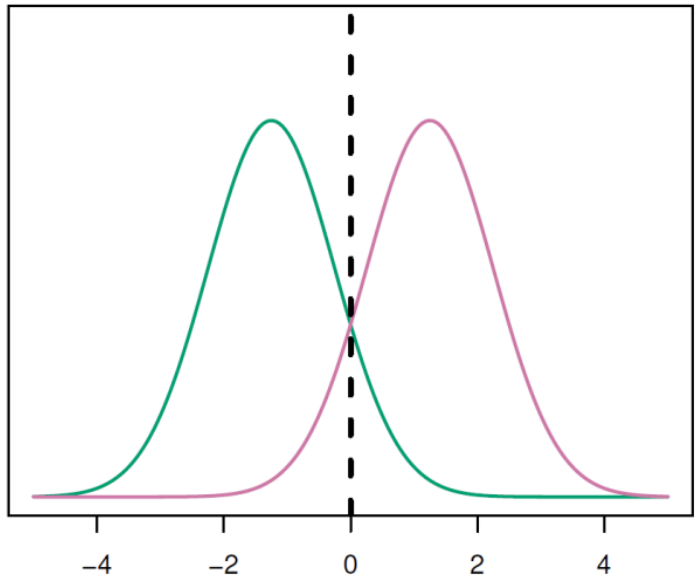
Note that $\delta_k(x)$ is a *linear* function of x.

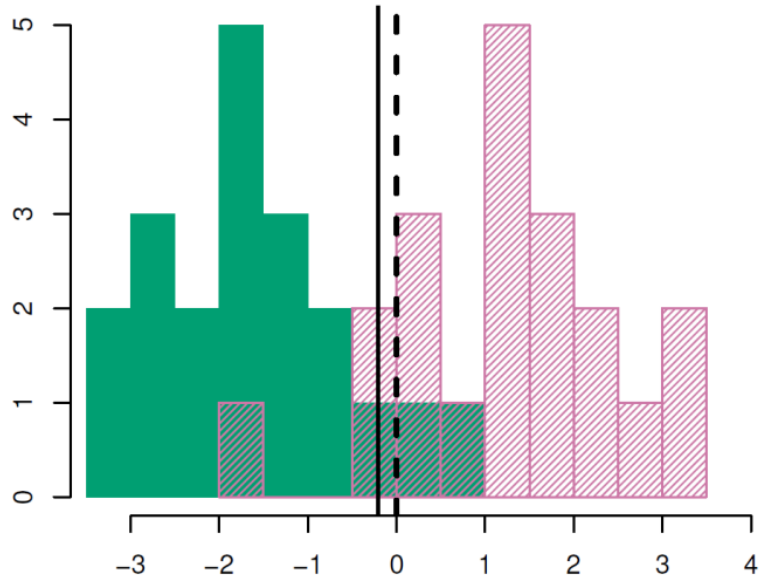
If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can see that the *decision boundary* is at

$$x = \frac{\mu_1 + \mu_2}{2}.$$

(See if you can show this)

Decision Boundary Example





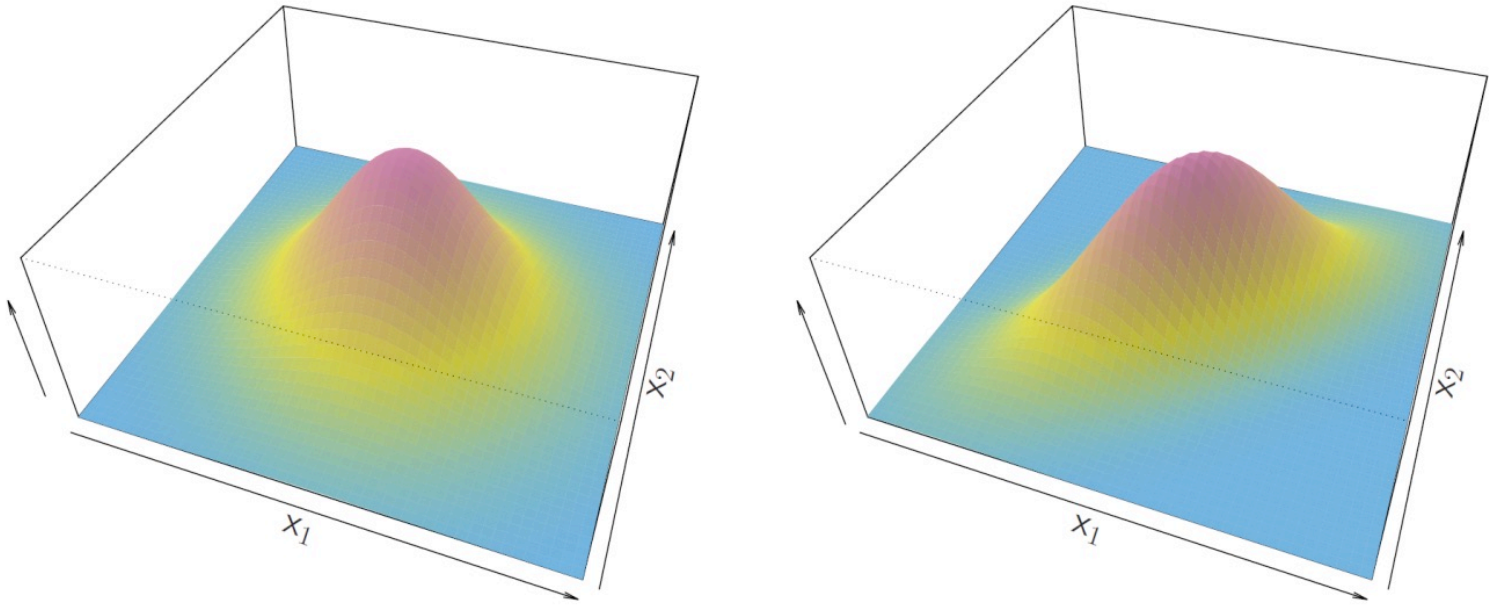
Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$. Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into the rule.

Estimating the Parameters

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2
\end{aligned}
$$

where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the kth class.

Linear Discriminant Analysis when $p > 1$



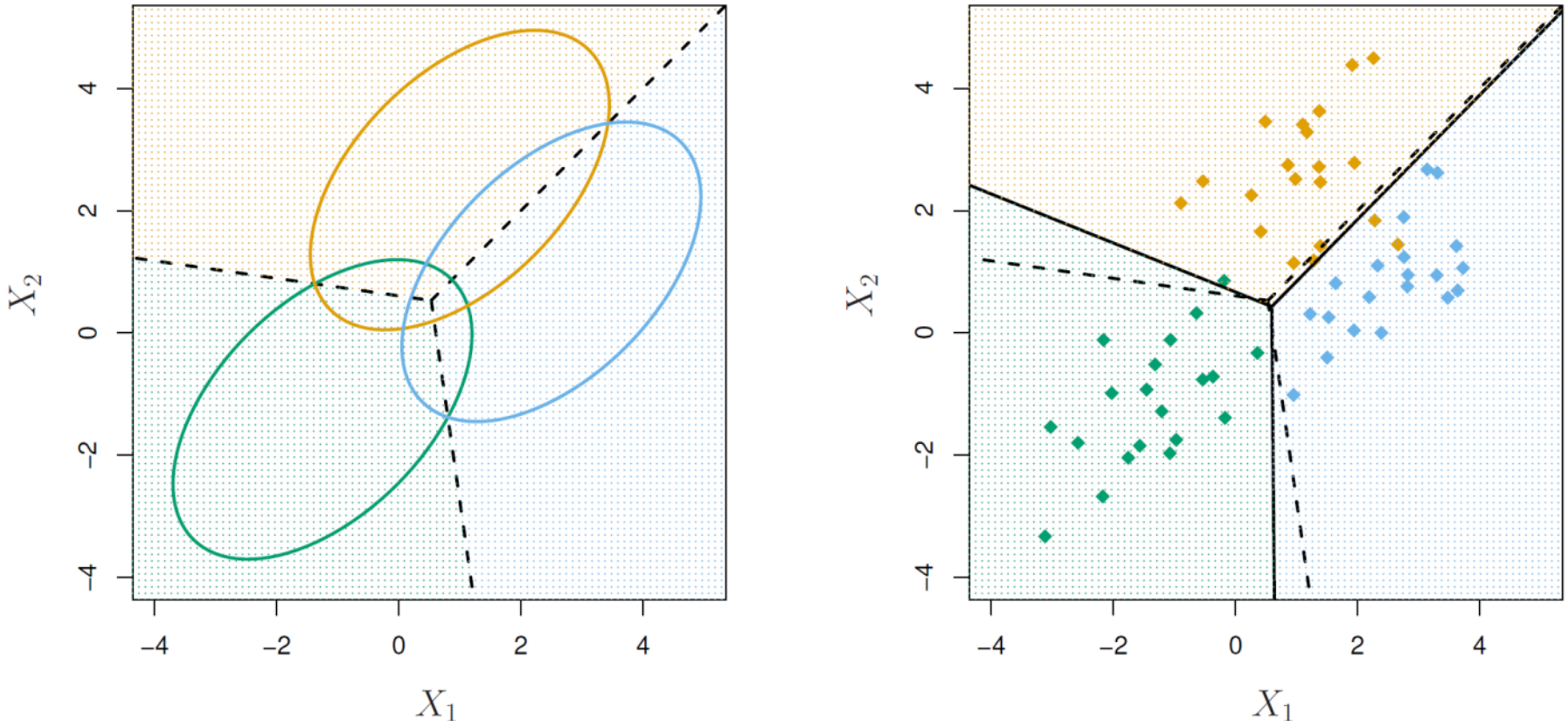
$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

$$\text{Discriminant function: } \delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

Despite its complex form,

$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \ldots + c_{kp}x_p$ — a linear function.

$p = 2$ and $K = 3$ Classes



Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.

The dashed lines are known as the *Bayes decision boundaries*.

Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

LDA of Fisher's Iris Data

- Use "*linDA*" to find out $\delta_k(x)$

```
iris.linda <- linDA(iris[,-5], iris$Species, prior = c(1/3,1/3,1/3))
iris.linda$functions
```

```
##                  setosa versicolor  virginica
## constant      -86.30847 -72.852607 -104.36832
## Sepal.Length   23.54417  15.698209   12.44585
## Sepal.Width    23.58787   7.072510    3.68528
## Petal.Length -16.43064   5.211451   12.76654
## Petal.Width  -17.39841   6.434229   21.07911
```

```
iris.linda$scores[c(1:10),]
```

```
##       setosa versicolor  virginica
## 1   89.84175   40.54492  -5.907026
## 2   73.33898   33.86902 -10.238836
## 3   74.99079   31.62274 -13.267604
## 4   66.99145   30.38796 -12.327408
## 5   89.84612   39.68235  -6.783083
## 6   97.93127   50.93367   7.346627
## 7   73.97104   32.63199 -10.390567
## 8   83.48548   38.78899  -6.243484
## 9   59.20811   25.31267 -16.830288
## 10 75.79455   34.45400 -10.701564
```

We could caculate the classification score for each class.

$$\hat{\delta}(setosa|Sepal, Petal) = -86.31 + 23.54 * SL + 23.59 * SW - 16.43 * PL - 17.40 * PW$$

$$\hat{\delta}(versicolor|Sepal, Petal) = -72.85 + 15.70 * SL + 7.073 * SW + 5.211 * PL + 6.434 * PW$$

$$\hat{\delta}(virginica|Sepal, Petal) = -104.4 + 12.45 * SL + 3.685 * SW + 12.77 * PL + 21.08 * PW$$

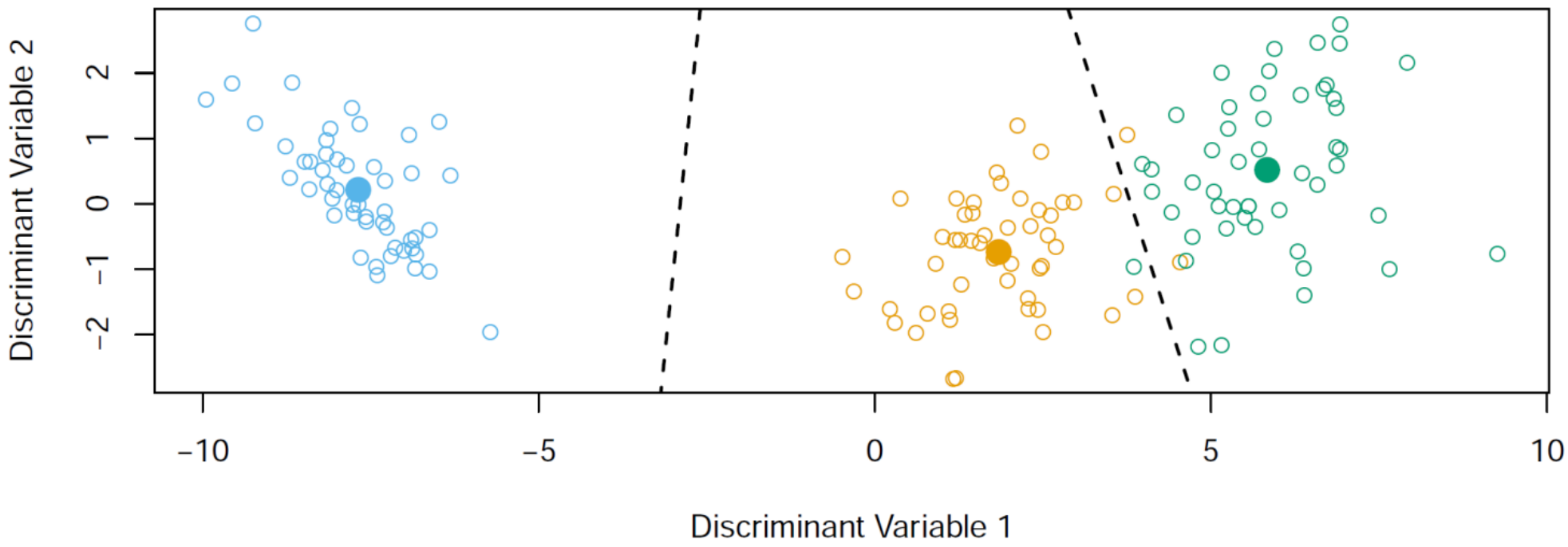
LDA of Fisher's Iris Data

- Use "*lda*" for visualization

```
iris.lda <- lda(iris$Species~., data=iris)
iris.lda
predict.iris_LDA <- predict(iris.lda)
table(iris$Species, predict.iris_LDA$class)
```

```
Call:
lda(iris$Species ~ ., data = iris)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088

             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          1        49
```

Fisher's Discriminant Plot



From $\delta_k(x)$ to Probabilities

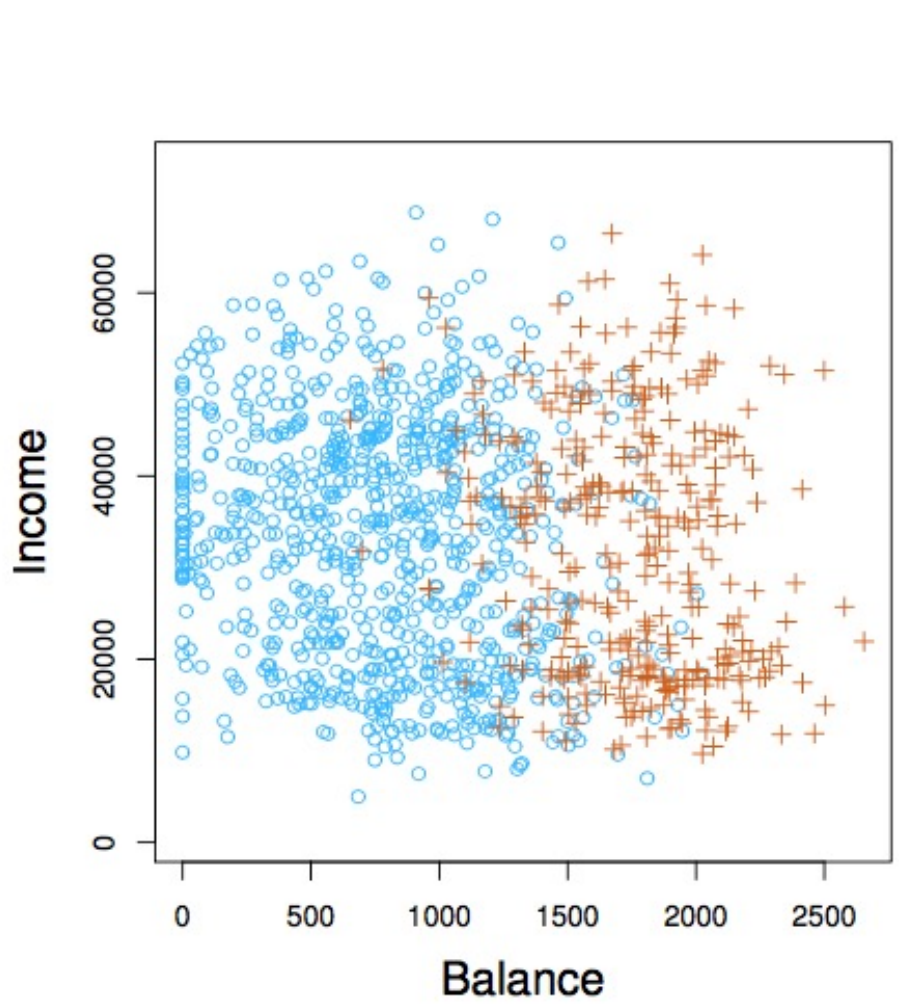
Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

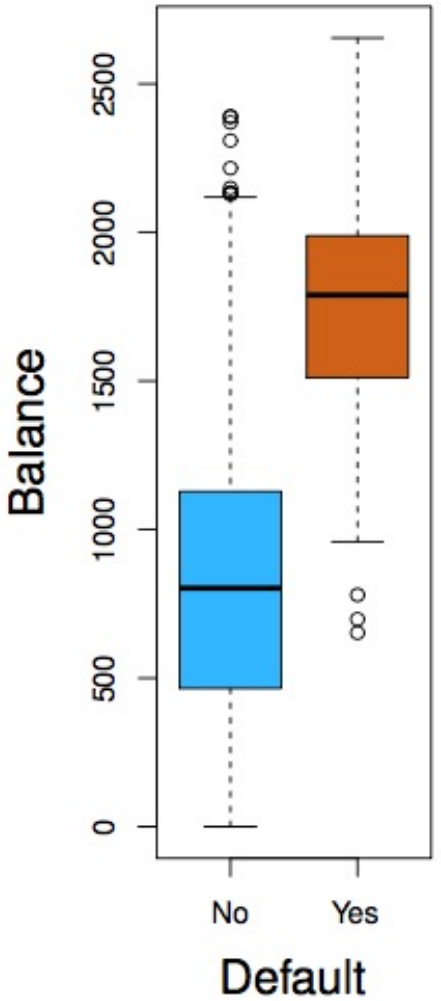
$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}.$$

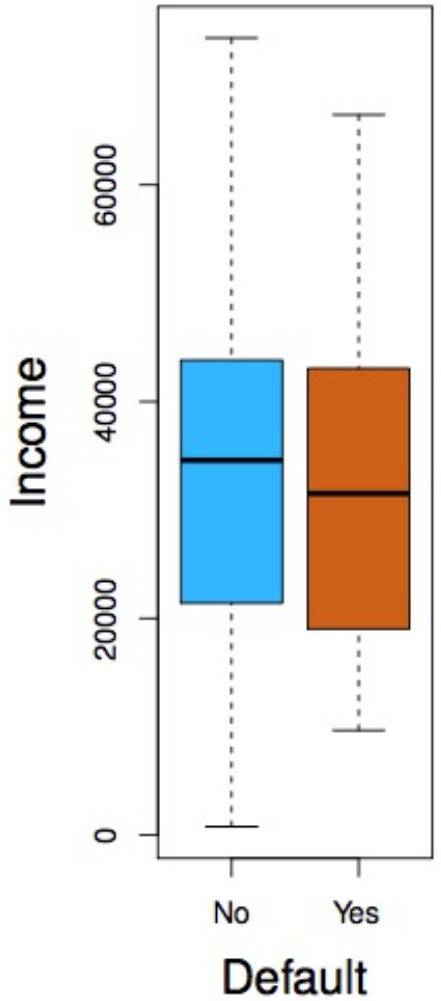
So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{\Pr}(Y = k|X = x)$ is largest.

When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2|X = x) \geq 0.5$, else to class 1.

Example: Credit Card Default







LDA on Credit Data

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

$(23 + 252)/10000$ errors — a 2.75% misclassification rate!

Some caveats:

- This is *training* error, and we may be overfitting. Not a big concern here since $n = 10000$ and $p = 2$!
- If we classified to the prior — always to class `No` in this case — we would make 333/10000 errors, or only 3.33%.
- Of the true `No`'s, we make $23/9667 = 0.2\%$ errors; of the true `Yes`'s, we make $252/333 = 75.7\%$ errors!

Classification Terminology

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

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 — 75.7% in example.

We produced this table by classifying to class `Yes` if

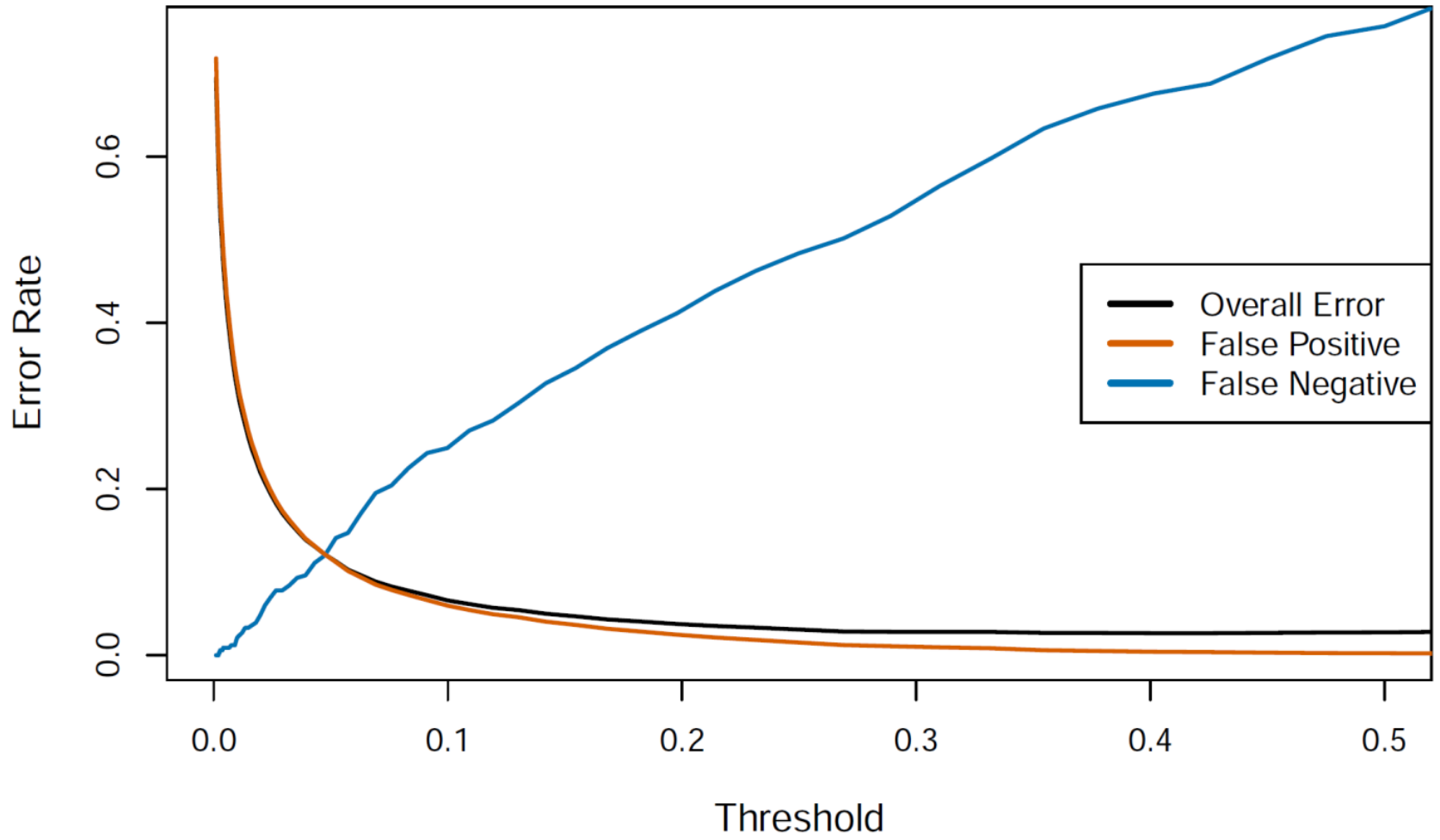
$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes} | \texttt{Balance}, \texttt{Student}) \geq 0.5$$

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

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

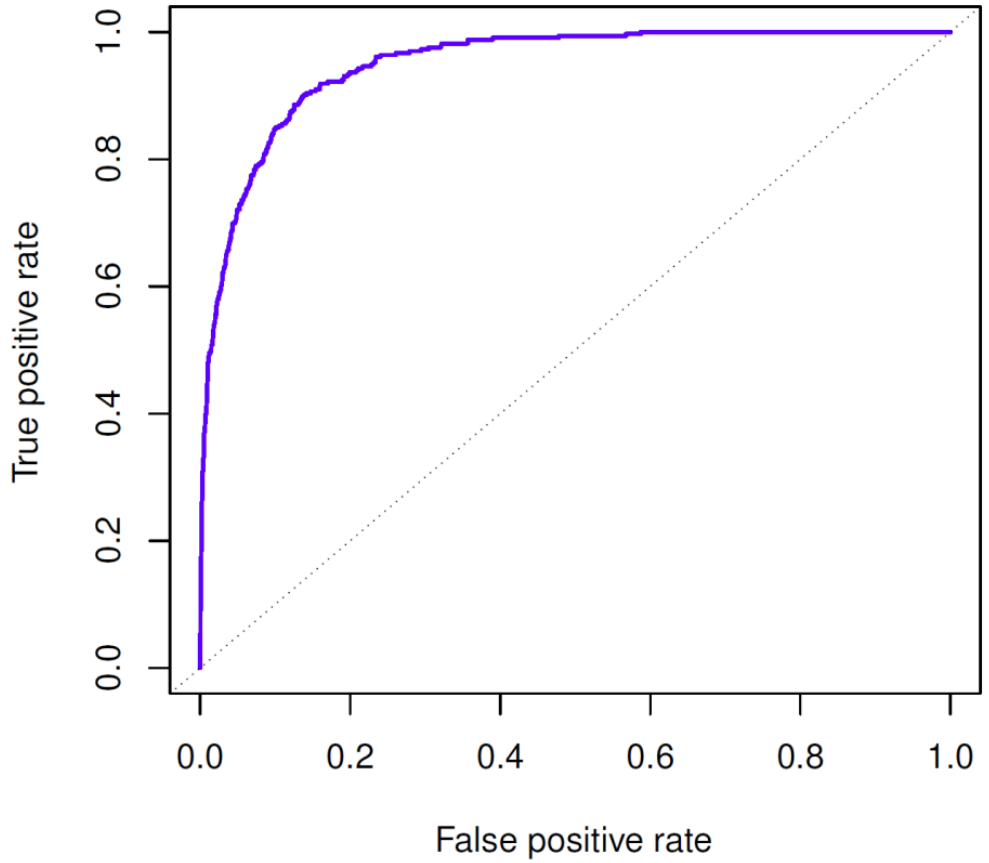
and vary *threshold*.

Varying the Threshold



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

The Receiver Operating Characteristics (ROC) Curve



- The ROC plot
  - False positive rate: 1-Specificity; healthy identified as not
  - True positive rate: Sensitivity; sick identified as so
- Higher area under the curve (AUC) is better

Riding Mowers

- R illustration

Other Forms of Discriminant Analysis

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

When $f_k(x)$ are Gaussian densities, with the same covariance matrix $\mathbf{\Sigma}$ in each class, this leads to linear discriminant analysis. By altering the forms for $f_k(x)$, we get different classifiers.

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

Quadratic Discriminant Analysis

With different covariance matrix Σ_k, the Bayes classifier assigns an observation $X = x$ to the class with the largest

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k\end{aligned}$$

The decision boundary is a quadratic function of x.

Quadratic Discriminant Analysis

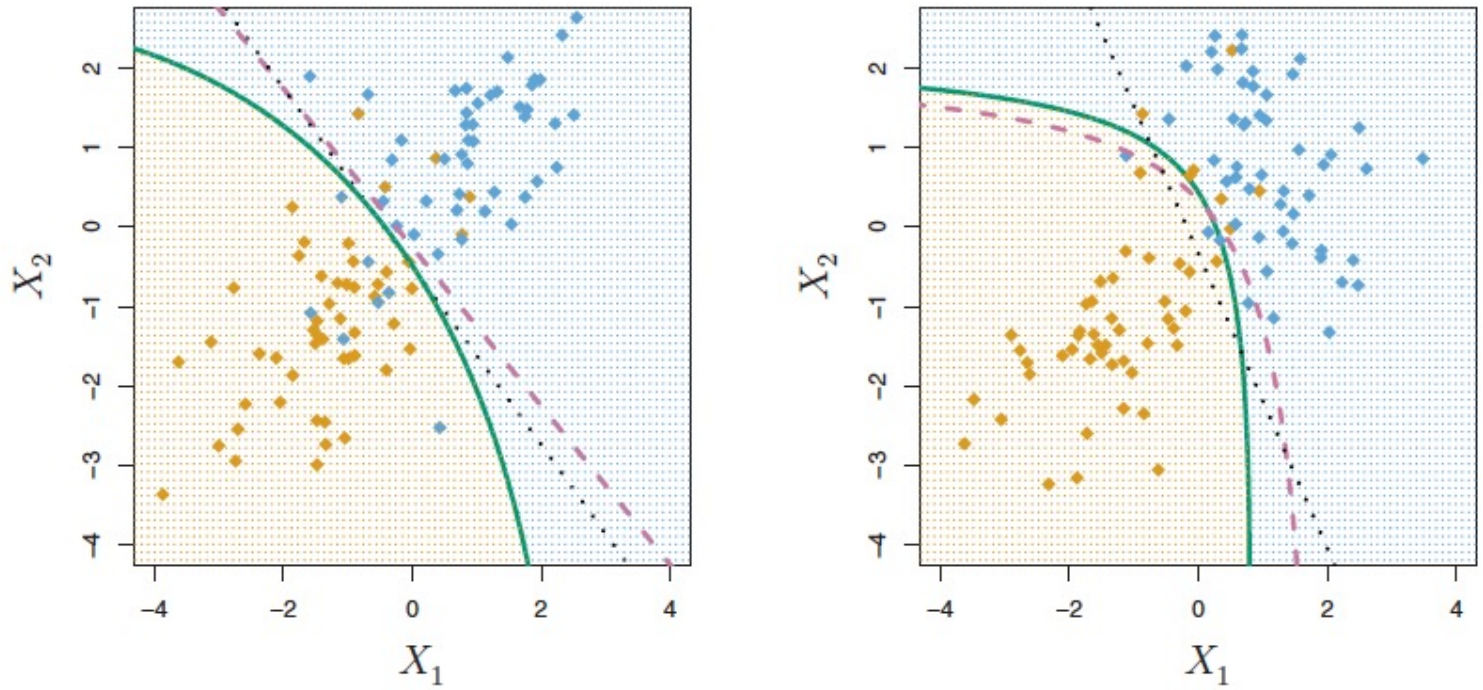


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

Naive Bayes

- Bayes Theorem gives $\Pr(Y = k|X = x) = \dfrac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$
- QDA: each $f_k(x)$ is p-dim multivariate Gaussian(μ_k, Σ_k).
- LDA: further assume $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma$.
- Naive Bayes: assume features are independent

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p),$$

where $f_{kj}(x_j)$ can be either Gaussian or a general unknown distribution.

	LDA	QDA	NB
Gaussian $f_{kj}(x_j)$?	Yes	Yes	Not necessarily
Diagonal Σ_k?	No	No	Yes
Shared $\Sigma_k = \Sigma$?	Yes	No	No

Naive Bayes

Assumes features are independent in each class.
Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Gaussian naive Bayes assumes each $\mathbf{\Sigma}_k$ is diagonal:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k$$

- can use for *mixed* feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.

Despite strong assumptions, naive Bayes often produces good classification results.

Comparison of LDA, QLA, NB, Logistic

- Always assign an observation to the class with maximal $P(Y = k|X = x)$.
- LDA:

$$\begin{aligned}\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) &= \log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right)\\ &= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\boldsymbol{\Sigma}^{-1}(\mu_k-\mu_K)\\ &\quad + x^T\boldsymbol{\Sigma}_p^{-1}(\mu_k-\mu_K)\\ &= a_k + \sum_{j=1} b_{kj}x_j,\end{aligned}$$

- QDA:

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl}x_j x_l,$$

- LDA is a special case of QDA with $c_{jkl} = 0$.

Comparison of LDA, QLA, NB, Logistic

- Naive Bayes:

$$\begin{aligned}
\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) &= \log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right) \\
&= \log\left(\frac{\pi_k \prod_{j=1}^{p} f_{kj}(x_j)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) + \sum_{j=1}^{p} \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right) \\
&= a_k + \sum_{j=1}^{p} g_{kj}(x_j),
\end{aligned}$$

- Any linear classifier (e.g. LDA) is a special case of NB with $g_{kj}(x_j) = b_{kj}x_j$.
- If we model $f_{kj}(x_j)$ in NB using one-dim Gaussian, then $g_{kj}(x_j) = b_{kj}x_j$.

Comparison of LDA, QLA, NB, Logistic

- Logistic:

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)=\beta_{k0}+\sum_{j=1}^{p}\beta_{kj}x_j.$$

  - Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).
  - LDA uses the full likelihood based on $\Pr(X,Y)$ (known as *generative learning*).
  - Despite these differences, in practice the results are often very similar.

Footnote: logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model.

Why Discriminant Analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis (LDA) does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, LDA is again more stable than the logistic regression model.
- LDA is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

Summary

- Logistic regression is popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.
- There are other classification methods e.g. K-nearest neighbors (not covered in this course).
- See Section 4.5 for comparisons of logistic regression, LDA, and KNN.

Linear Scenarios

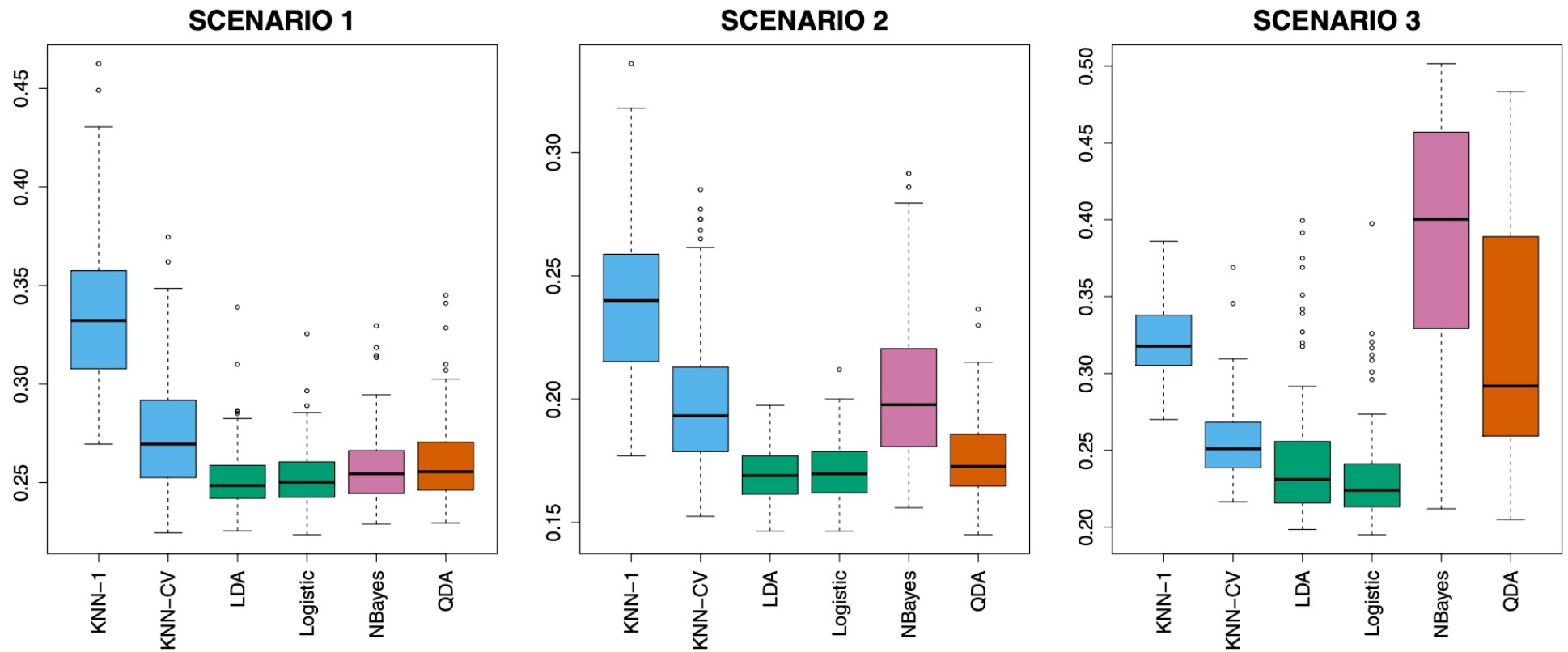


FIGURE 4.11. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

Non-linear Scenarios

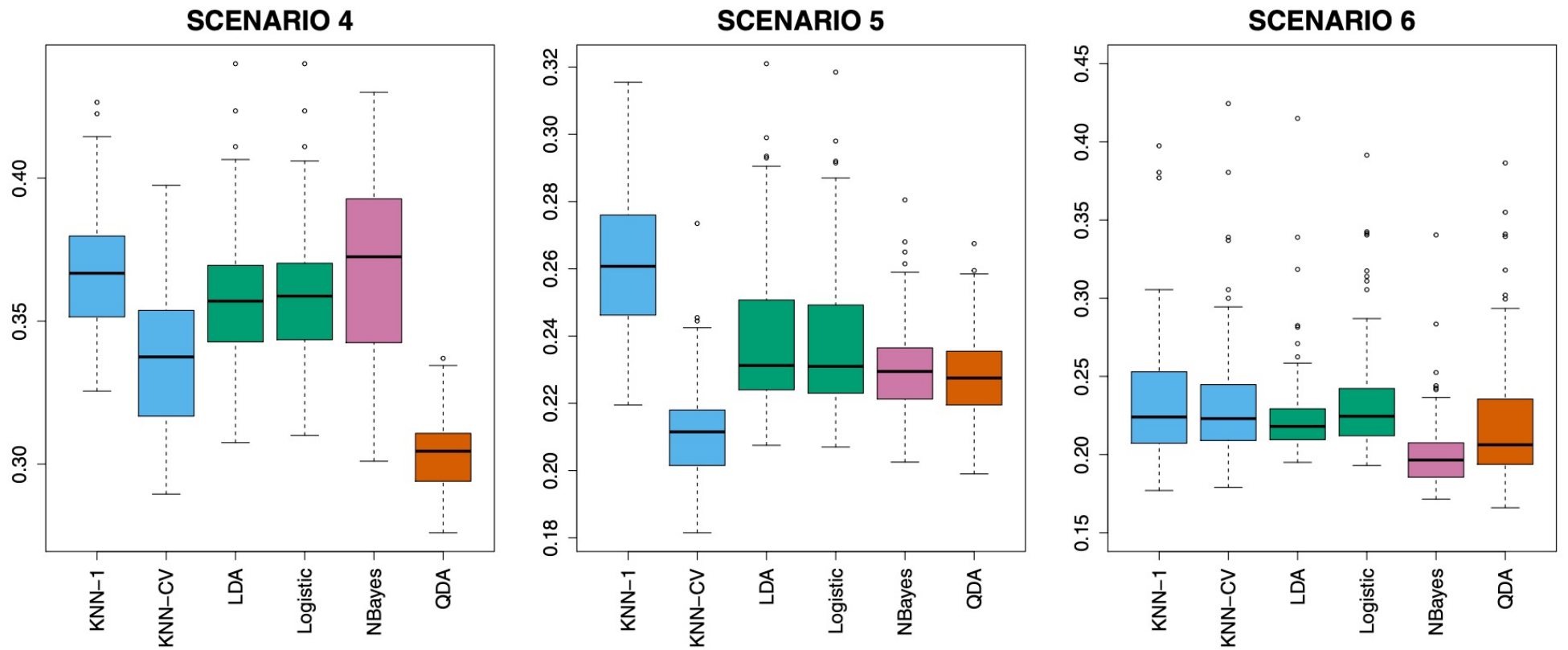


FIGURE 4.12. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*