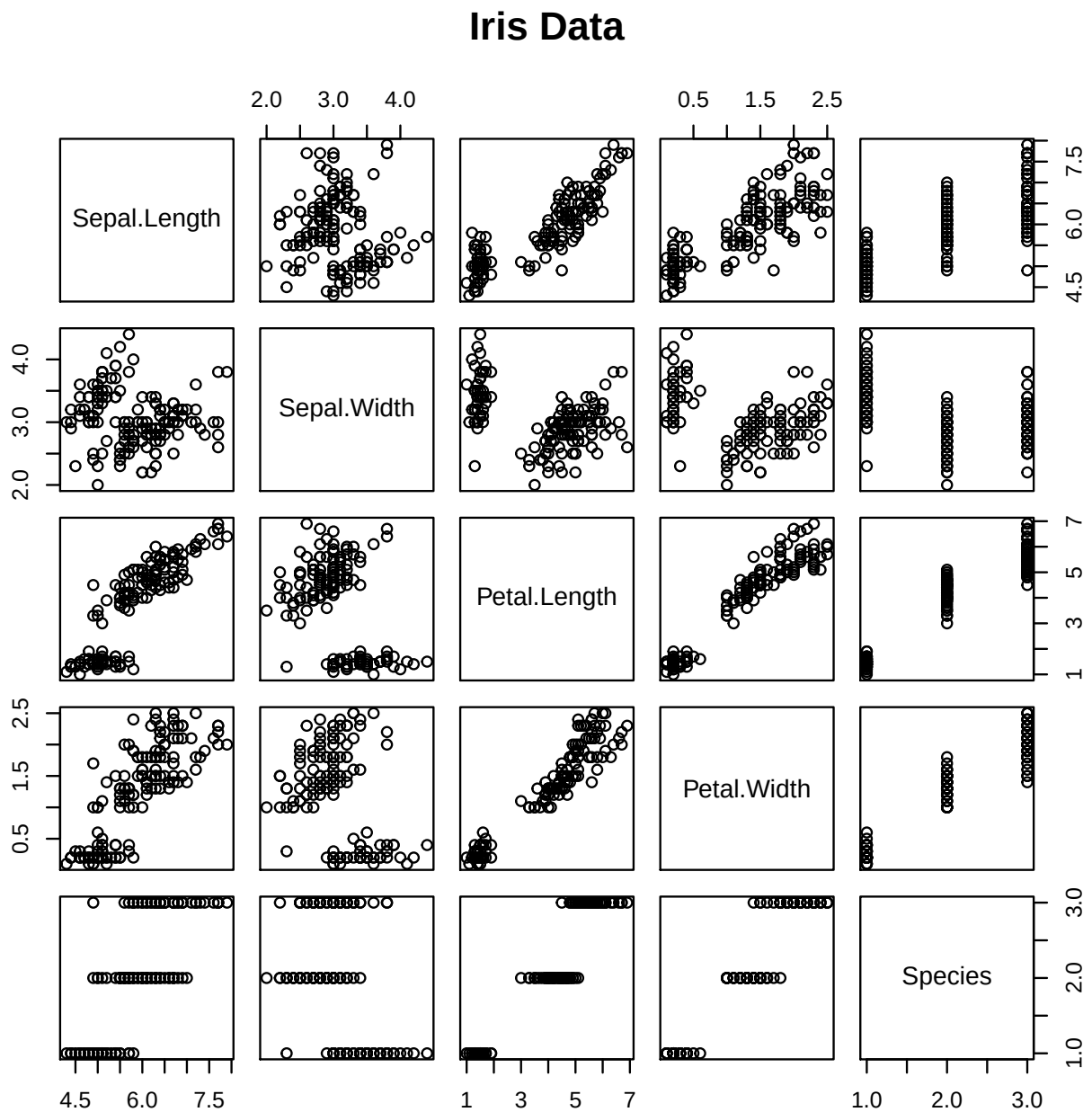
STATS 769

Classification

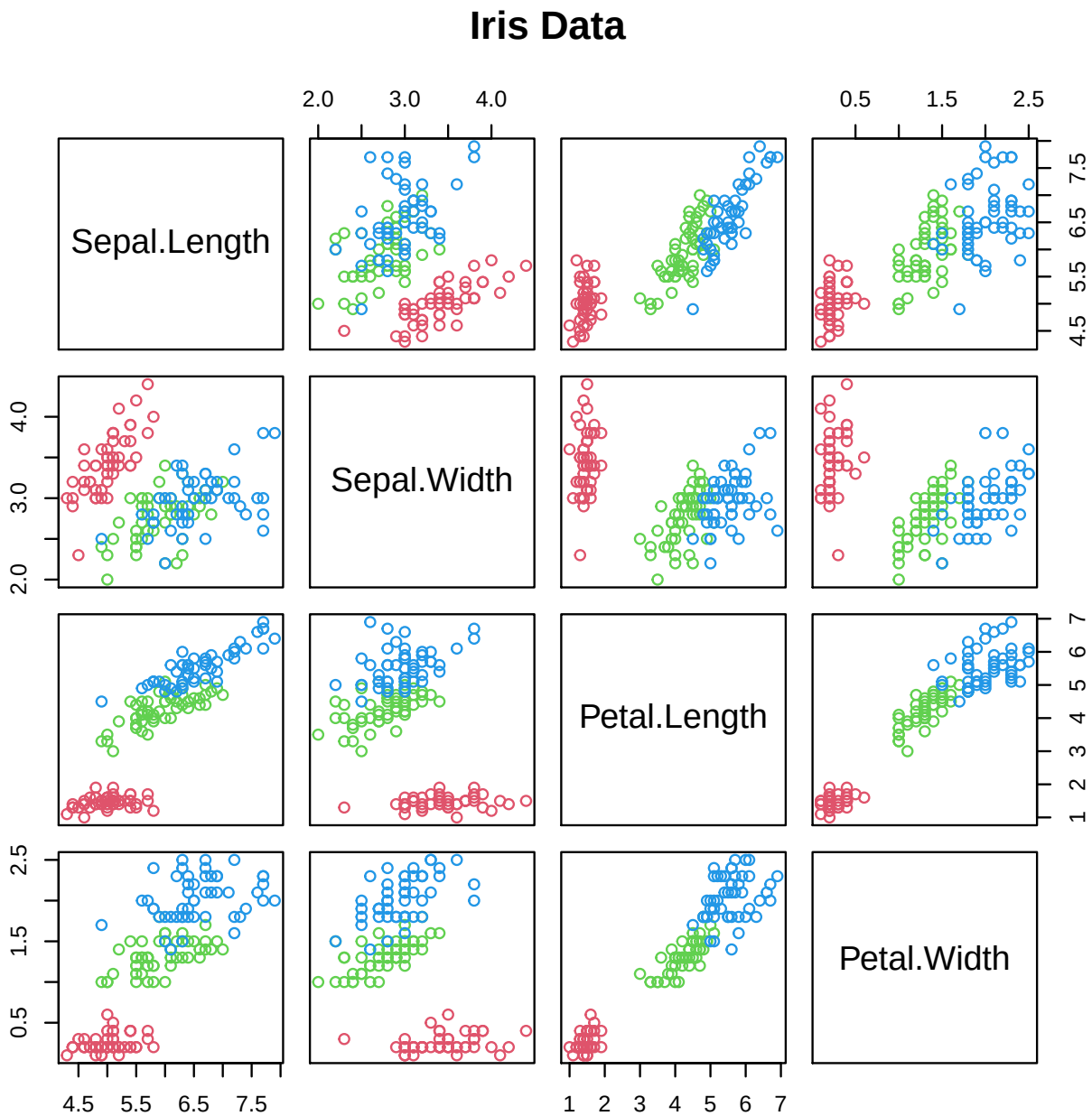
Yong Wang
Department of Statistics
The University of Auckland

2021

A Classification Problem



A Classification Problem II



Classification

- Classification is the most studied topic in data mining.
- It is somehow similar to regression, except that the response variable now is categorical.
  - The response Y (`Species`) has three classes:
    `setosa`, `versicolor`, `virginica`
- Classification is to build a model (classifier) $\widehat{Y} = \widehat{f}(X)$ from data to predict the value of Y.
- Many families of models can be used for both regression and classification, e.g., generalised linear models, tree-structured models, neural networks.

Basic Classification Methods

- Linear discriminant analysis
- Quadratic discriminant analysis
- Naive Bayes

- Logistic regression
- Generalised additive models
- K-nearest neighbours

Confusion Matrices

- A confusion matrix of a classifier for a data set of size n looks like:

Y \ $\widehat{Y}$	A	B	C
A	n_{11}	n_{12}	n_{13}
B	n_{21}	n_{22}	n_{23}
C	n_{31}	n_{32}	n_{33}

  where n_{ij} is the number of observations following in this cell.
- In this table, the number of the correctly-classified observations is $n_c = n_{11} + n_{22} + n_{33}$.
- The classification accuracy is $A = n_c / n$.
  - The higher, the better.
  - This is what is often reported.
- The misclassification rate is $R = 1 - n_c / n$.

Prediction Errors

- The misclassification rate is the most commonly-used prediction error (PE) for classification.
- It is a sample mean:

$$R = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \neq \widehat{Y}_i).$$

- It hence approaches a theoretical mean (as $n \to \infty$):

$$\mathbb{E}[I(Y_i \neq \widehat{Y}_i)] = \mathbb{P}(Y_i \neq \widehat{Y}_i),$$

  i.e., the misclassification probability.

How to Classify?

- A fundamental rule to classify an observation x is to use conditional probabilities:

$$\mathbb{P}(Y = j|X = x), \quad j = 1, \ldots, J,$$

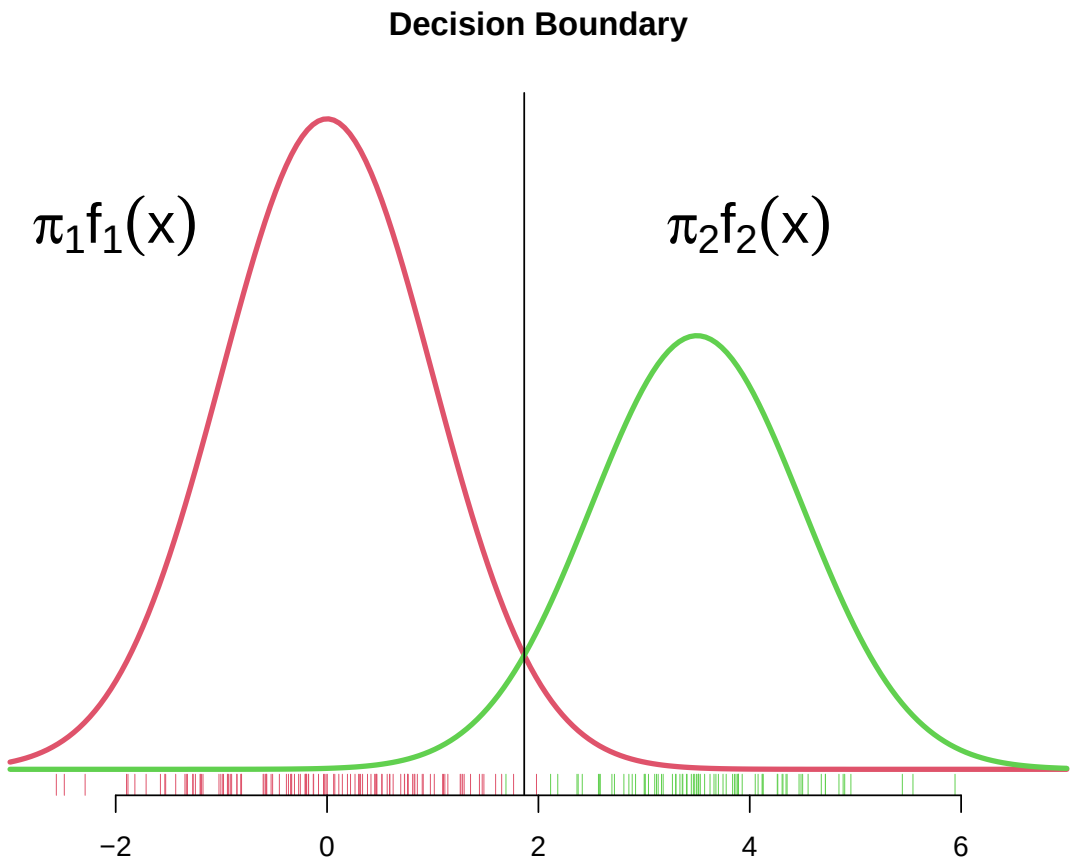
  which is also known as posterior probabilities.
- One should classify x into the most probable class, i.e., class j^* if $\mathbb{P}(Y = j^*|X = x)$ is the largest for all $j = 1, \ldots, J$.
- If class j is known to have a density function $f_j(x)$, then

$$\mathbb{P}(Y = j|X = x) = \frac{\pi_j f_j(x)}{\sum_{l=1}^{J} \pi_l f_l(x)}.$$

  where $\pi_j = \mathbb{P}(Y = j)$ is the proportion of class j observations in the population and is also known as the prior probability for class j.
- To find j^* is to find which $\pi_j f_j(x)$ is the largest.

With Known Densities

- Where to separate the two classes?



- The boundary separating the two classes is known as the decision boundary.

Linear Discriminant Analysis

- A d-dimensional multivariate normal density:

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},$$

  where μ is the mean vector and Σ is the covariance matrix.
- Fit a multivariate normal distribution to the observations in each class (using maximum likelihood or unbiased estimators).
- However, all multivariate normal distributions have the same covariance matrix estimate $\widehat{\Sigma}$ (but each has its own mean vector estimate $\widehat{\mu}_j$).
- Then classification is carried out with density estimates.

Linear Discriminant Analysis II

- For class j, let

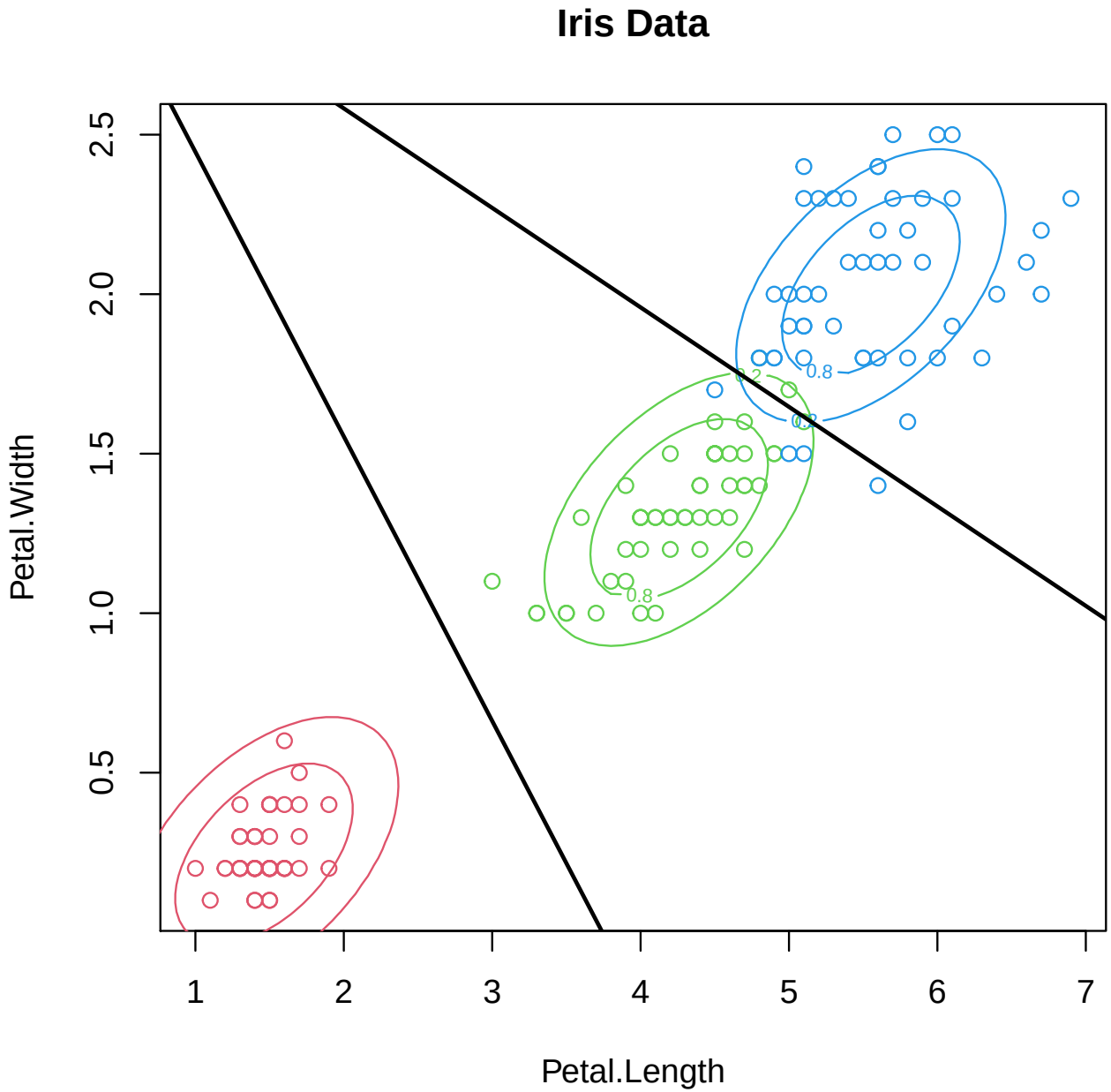
$$\delta_j(x) = \log[\pi_j f_j(x)].$$

- With a common variance matrix, it can reduce to

$$\delta_j(x) = x^T \mathbf{\Sigma}^{-1} \mu_j - \frac{1}{2} \mu_j^T \mathbf{\Sigma}^{-1} \mu_j + \log(\pi_j).$$

- Denote by $\delta_{jk} = \delta_j(x) - \delta_k(x)$, which is the log-odds and is a discriminant function.
  - Between two classes, one should classify observation x as class j if $\delta_{jk}(x) > 0$, or class k if $\delta_{jk} < 0$.
  - The decision boundary is where $\delta_{jk}(x) = 0$.
- $\delta_{jk}(x)$ is a linear function of $x \in \mathbb{R}^d$, hence linear discriminant analysis (LDA).
- Any decision boundary determined by $\delta_{jk}(x) = 0$ is a linear equation of x, which corresponds to a hyperplane in $\mathbb{R}^d$.

Decision Boundaries of LDA



Quadratic Discriminant Analysis

- Fit a multivariate normal distribution to the observations in each class (using the maximum likelihood method)
- Each multivariate normal distributions has its own covariance matrix estimate $\widehat{\Sigma}_j$ (and its mean vector estimate $\widehat{\mu}_j$).
- Then go ahead with classification using density estimates.

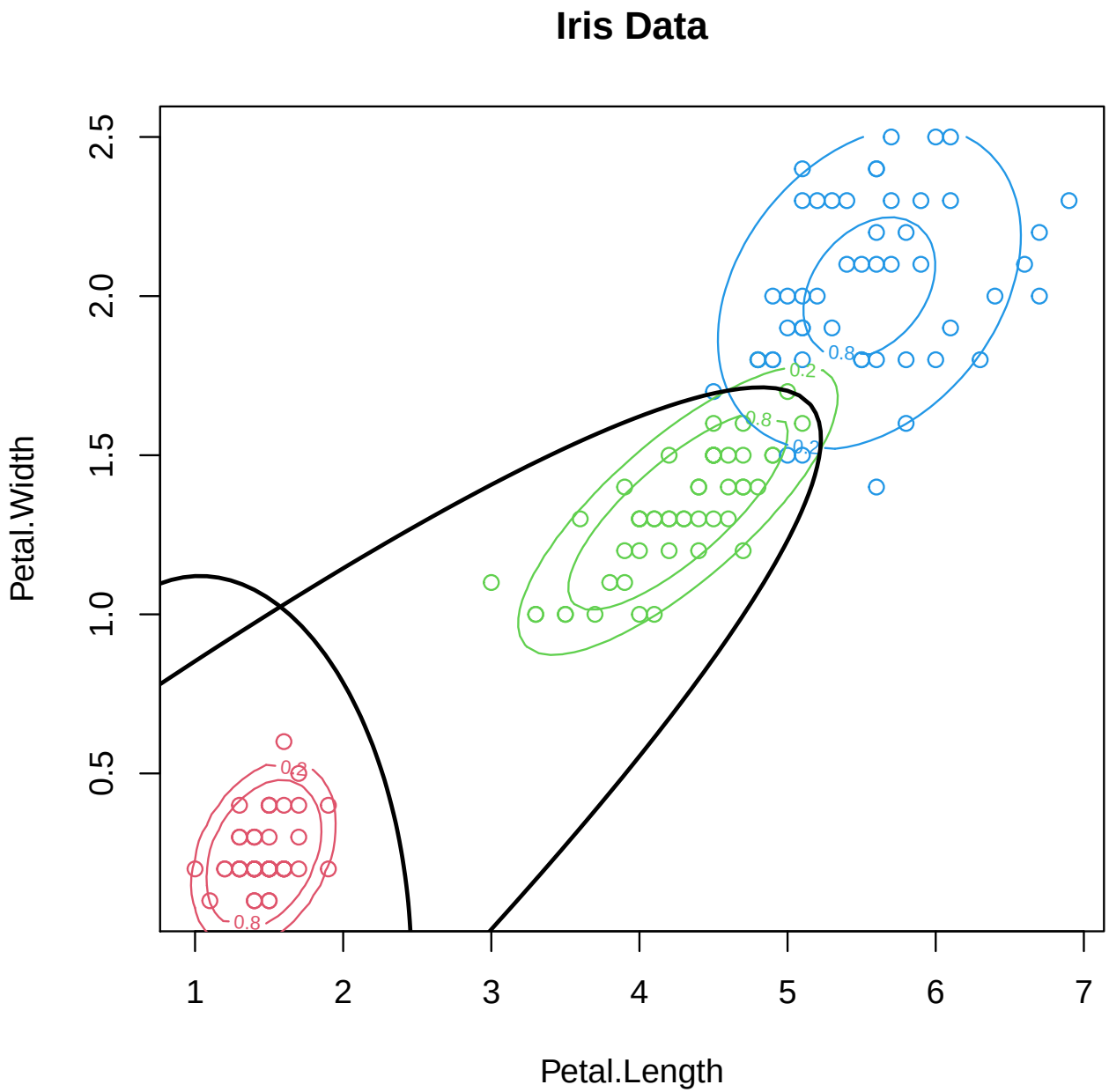
Quadratic Discriminant Analysis II

- Similarly let

$$\begin{aligned}\delta_j(x) &= \log[\pi_j f_j(x)] \\ &= -\frac{1}{2}x^T\Sigma_j^{-1}x + x^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j \\ &\quad - \frac{1}{2}\log(|\Sigma_j|) + \log(\pi_j).\end{aligned}$$

- This is a quadratic function of $x \in \mathbb{R}^d$, hence quadratic discriminant analysis (QDA).
- The decision boundary determined by $\delta_{jk}(x) = 0$ corresponds to a quadratic surface in $\mathbb{R}^d$.
- QDA can be more accurate than LDA, if there are many observations and the covariance matrices are far from being similar.

Decision Boundaries of QDA



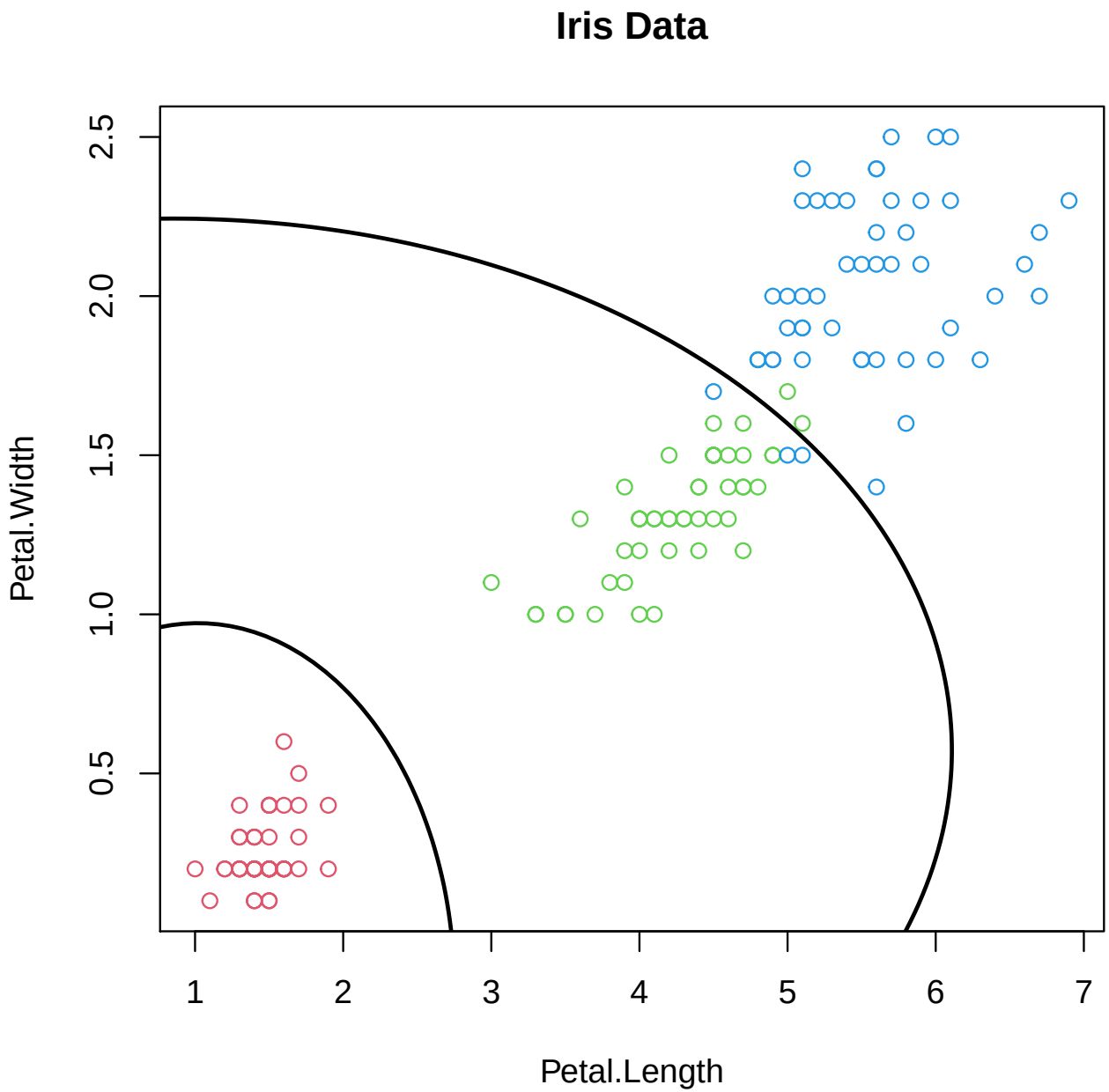
Naive Bayes

- The idea of Naive Bayes is pretty naive/simple.
- Performance is sometimes reasonable.
- Perhaps good for high-dimensional data, mixed with quantitative and qualitative variables.
- Estimate the density of class j by the product of one-dimensional (marginal) density estimates:

$$f_j(x) = f_{j1}(x_1) \cdots f_{jp}(x_p),$$

  as if assuming independence among predictor variables.
- Each $f_{jk}(x_k)$ is a density estimate from the values of x_k only.
  - For a quantitative variable, it can be a normal density estimate or a kernel density estimate (KDE) (to be studied later).
  - For a categorical variable, it is the sample proportions for each level of x_k.
- Same classification rule: Choose the class with the largest $\pi_j f_j(x)$.

Decision Boundaries of Naive Bayes (Marginal Normal)



Recommended Readings

ISLv2 (basics):

- Sections 4.1, 4.4,
- Labs: Sections 4.7.3–4.7.4

ESL (advanced):

- Section 4.3