CLASSIFICATION

Part II

Outline

- Probabilistic generative models
  - Linear discriminant analysis
  - Quadratic discriminant analysis
  - Naive Bayes

Drawbacks of logistic regression

- When the classes are well-separated, the parameter estimates are unstable
- If the sample size is small and the distribution of the predictors is approximately normal in each of the classes, there are more accurate approaches than logistic regression

Probabilistic generative models

The Bayes rule

- Assign an observation to the most likely class, given its predictor values

$$p_k(x) = \Pr(Y = k|X = x)$$

Bayes' theorem

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- π_k is the *prior* probability that a randomly chosen observation comes from the kth class
- $f_k(x) = \Pr(X = x|Y = k)$ is the density function of X for an observation that comes from the kth class

- $p_k(x)$ is called the posterior probability
- Assign an observation $X = x$ to the class for which $p_k(x)$ or $\pi_k f_k(x)$ is largest

Linear discriminant analysis for $p = 1$

- Assume that $f_k(x)$ is *normal* or *Gaussian*

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

  - μ_k and σ_k^2 are the mean and variance parameters

- Further assume that

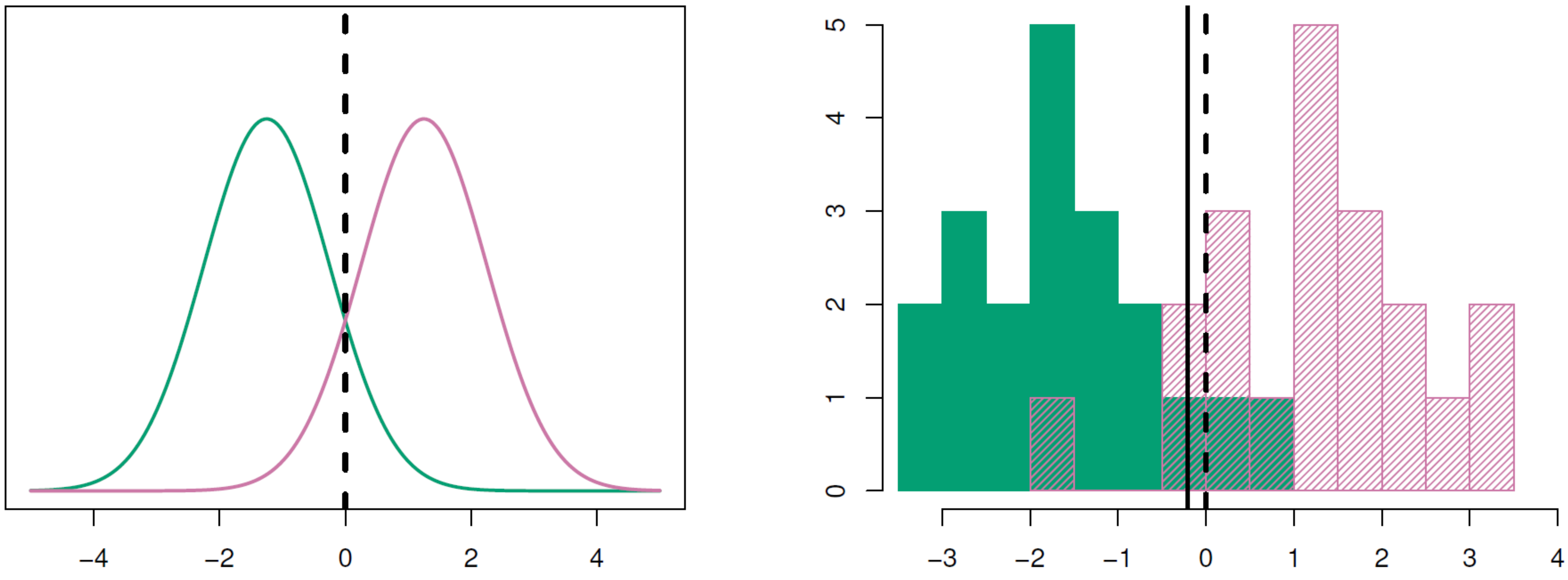
$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2$$

- Define

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

- The Bayes classifier is equivalent to assigning the observation $X = x$ to the class for which $\delta_k(x)$ is largest

- If $K = 2$ and $\pi_1 = \pi_2$, the Bayes decision boundary is
$$x = \frac{\mu_1 + \mu_2}{2}$$



$$K = 2, \pi_1 = \pi_2 = 0.5, \mu_1 = -1.25, \mu_2 = 1.25, \sigma_1 = \sigma_2 = 1$$

- Linear discriminant analysis (LDA) approximates the Bayes classifier by plugging estimates for π_k, μ_k, and σ^2

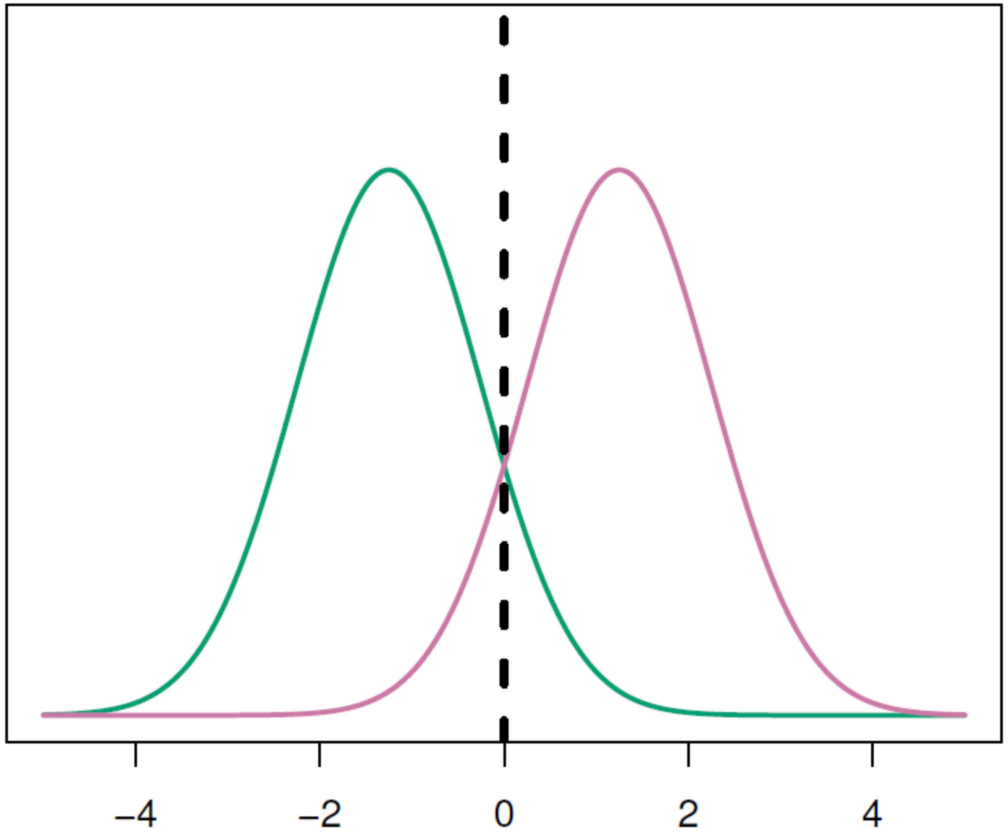
$$\hat{\pi}_k = \frac{n_k}{n}$$

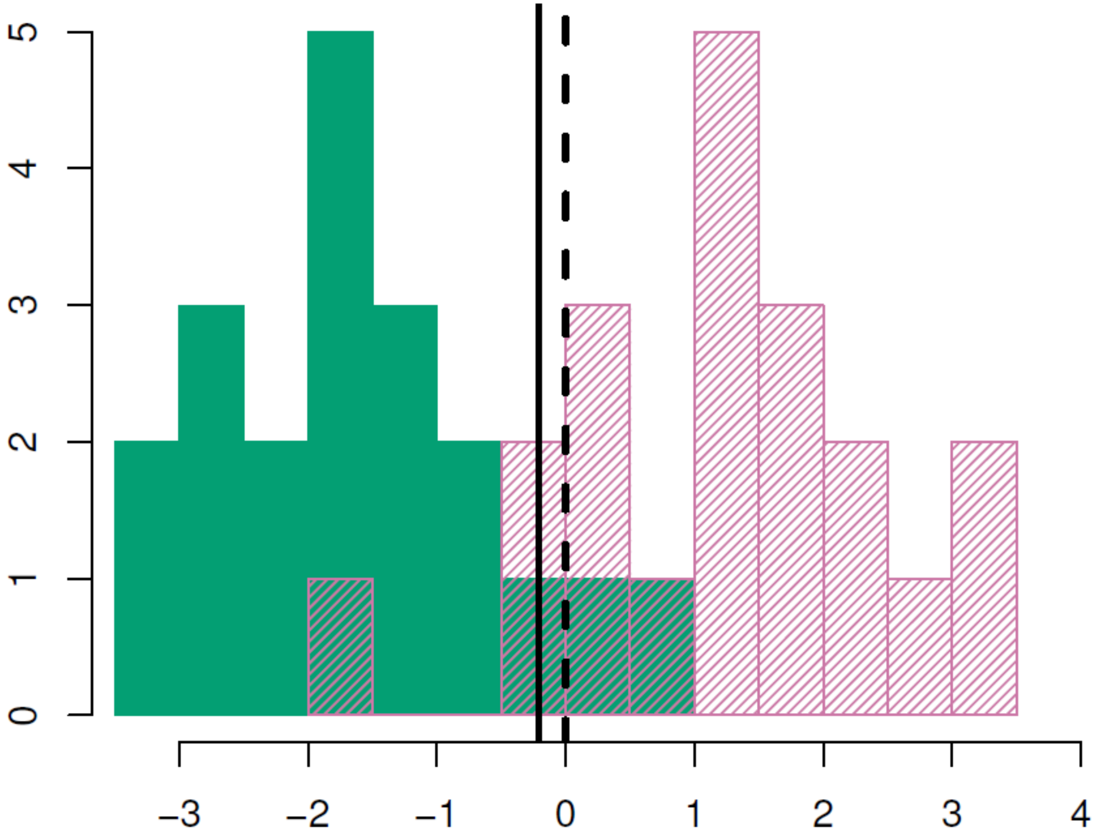
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

- The discriminant functions

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$





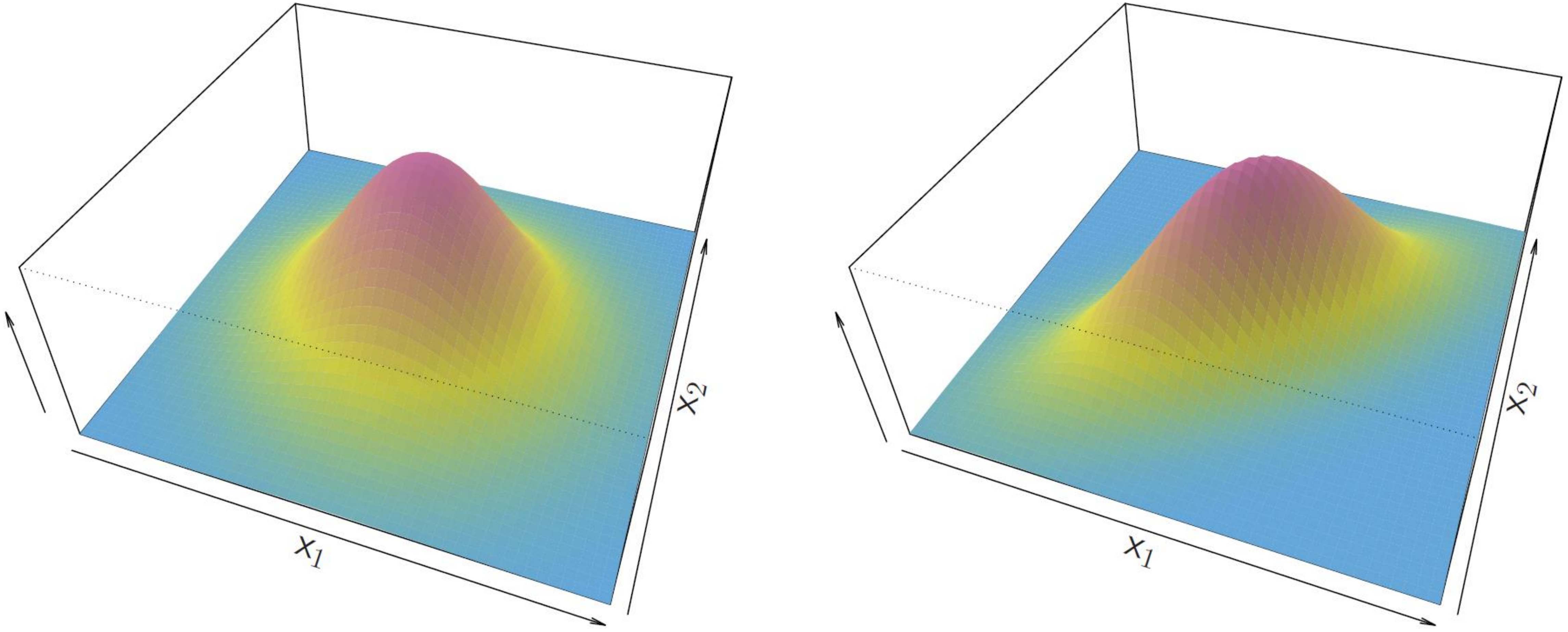
The Bayes error rate and the LDA test error rate are 10.6% and 11.1%, respectively

Linear discriminant analysis for $p > 1$

- Given $Y = k, X = (X_1, X_2, \ldots, X_p)^T$ is multivariate normal

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \times$$
$$\exp\left\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right\}$$

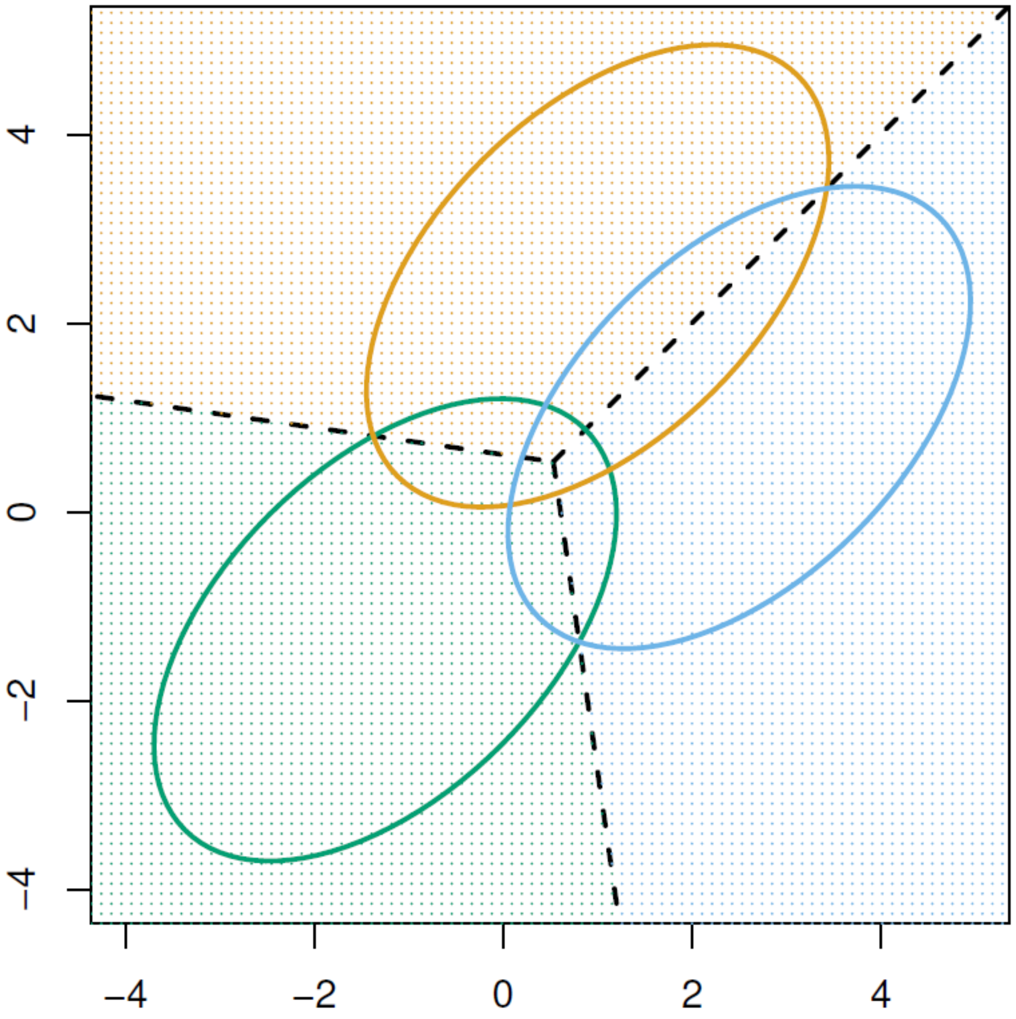
- $X|Y = k \sim N(\mu_k, \Sigma)$
- μ_k is a class-specific mean vector and Σ is a common covariance matrix

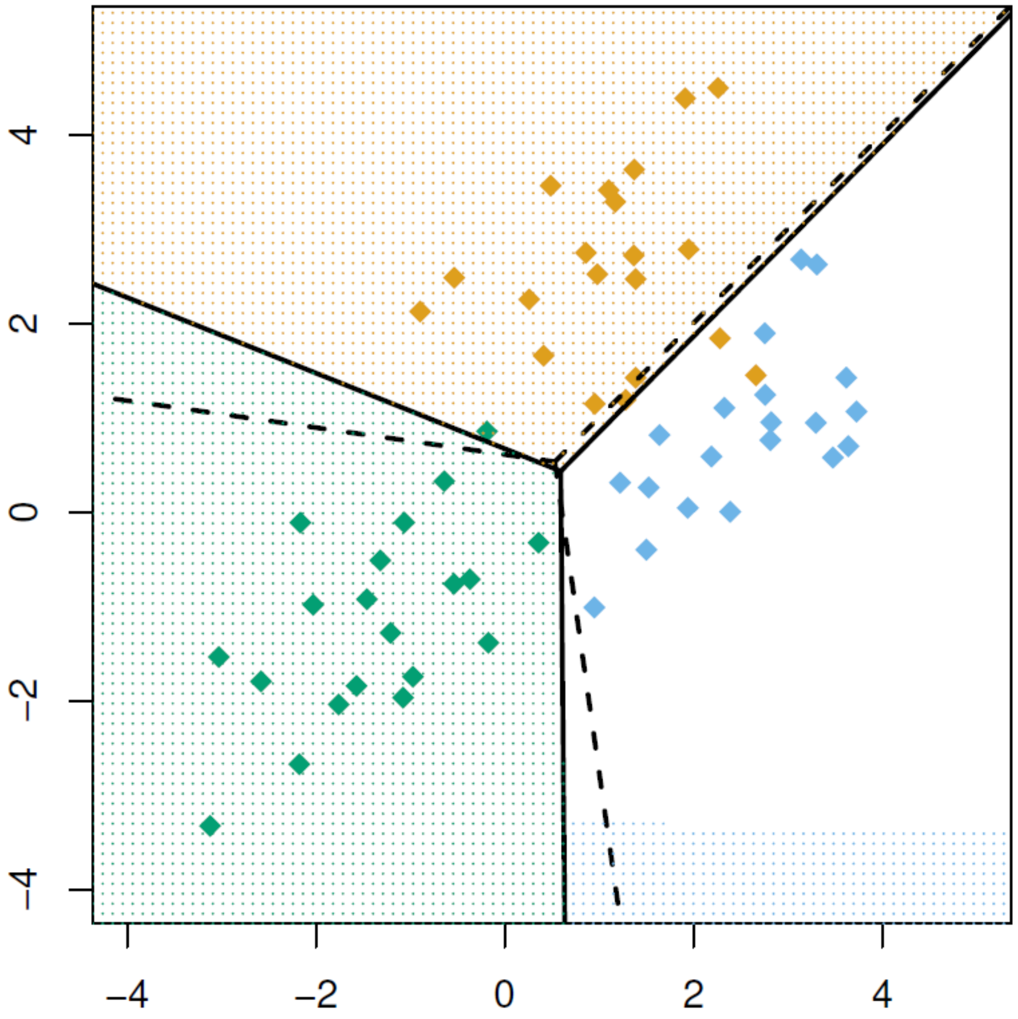


The two predictors are uncorrelated (left panel) and have a correlation of 0.7 (right panel)

- The Bayes classifier assigns an observation $X = x$ to the class for which $\delta_k(x)$ is largest

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





An illustrative example. The test error rates for the Bayes and LDA classifiers are 0.0746 and 0.0770, respectively

Default data

- Use LDA to predict whether an individual will default on the basis of balance and student
- The prediction rule

$$\Pr(default = Yes|X = x) > 0.5$$

- The *training* error rate is 2.75%

Class-specific performance

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

TABLE 4.4. *A confusion matrix compares the LDA predictions to the true default statuses for the 10,000 training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.*

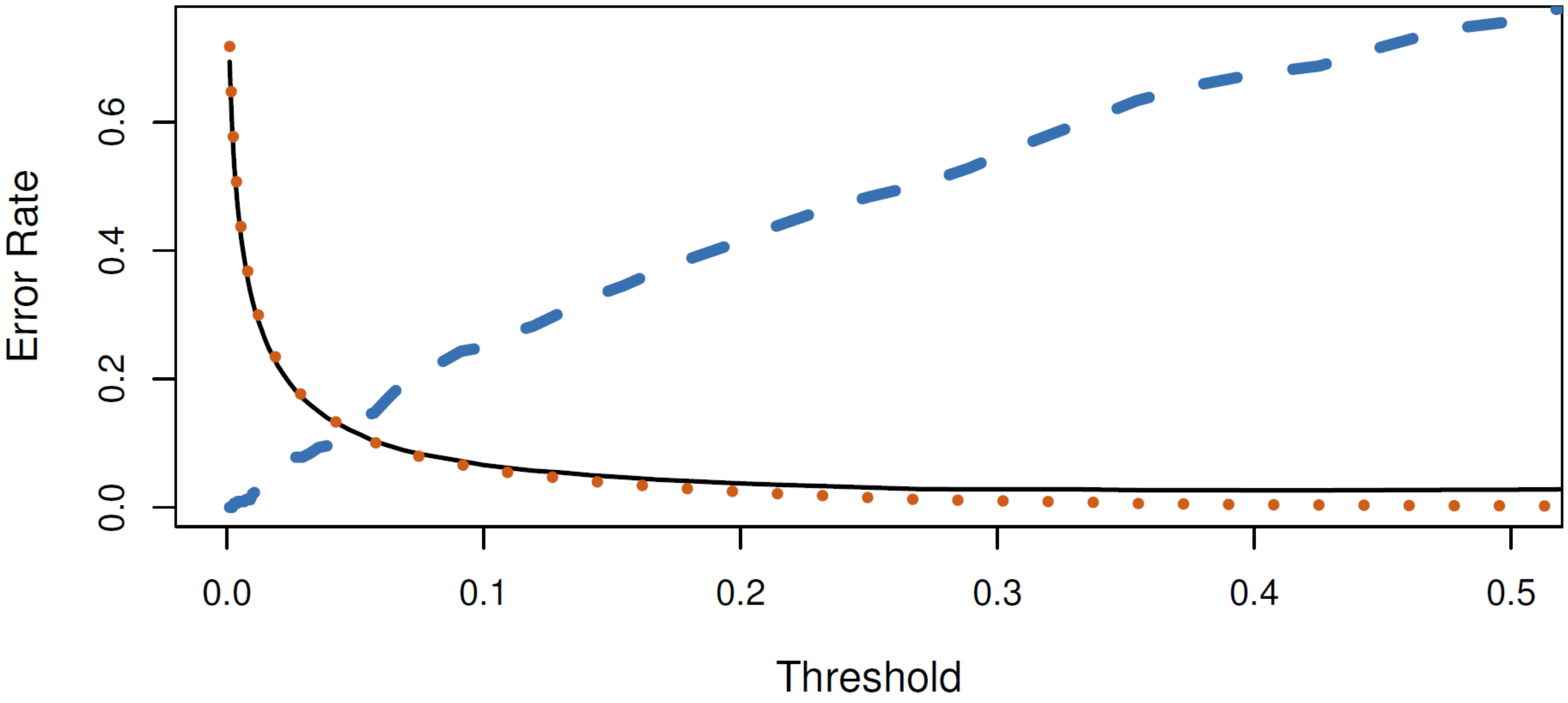
$\Pr(default = Yes | X = x) > 0.2$

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

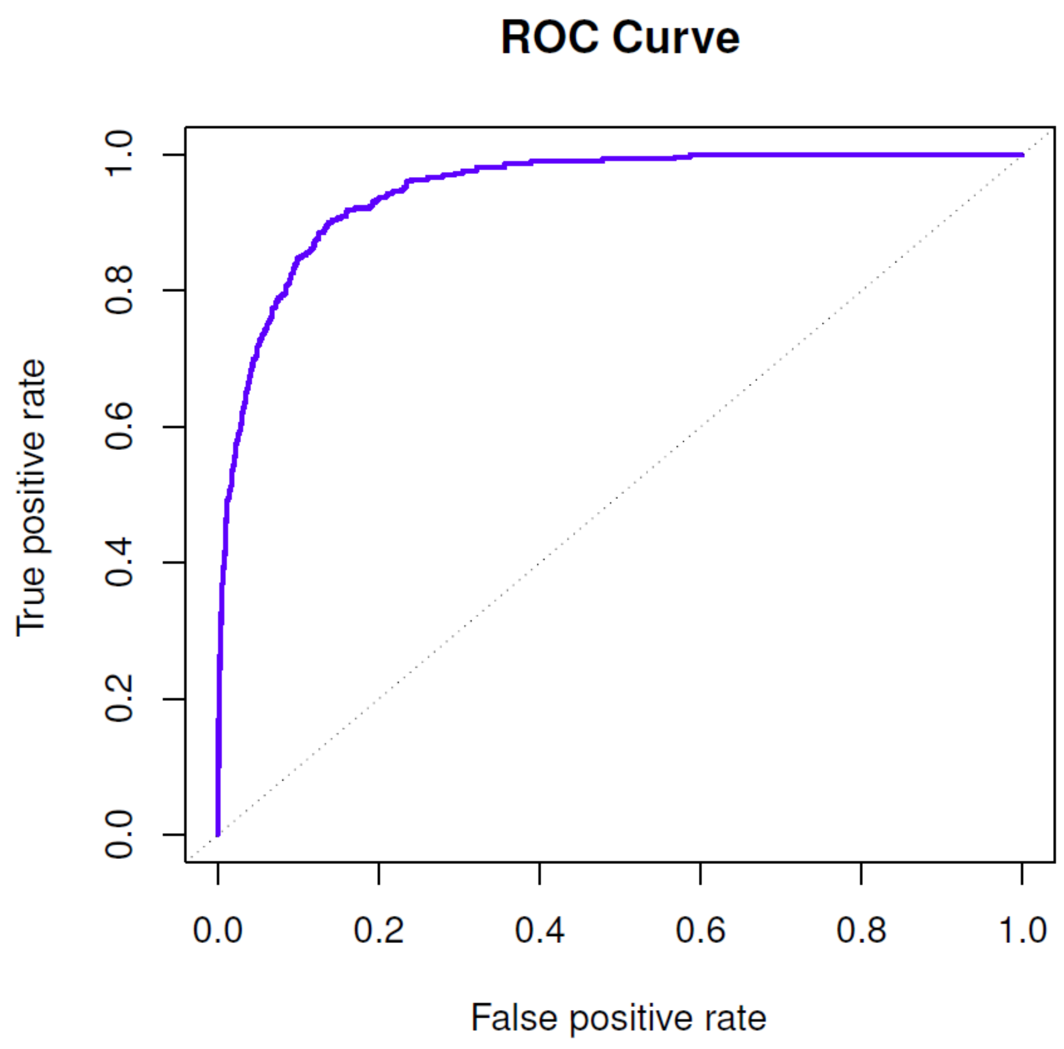
TABLE 4.5. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds* 20 %.

Sensitivity and specificity

- Sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value
- Specificity: the fraction of non-defaulters that are correctly identified, using that same threshold value



The Default data. Error rates are shown as a function of the threshold value for the posterior probability

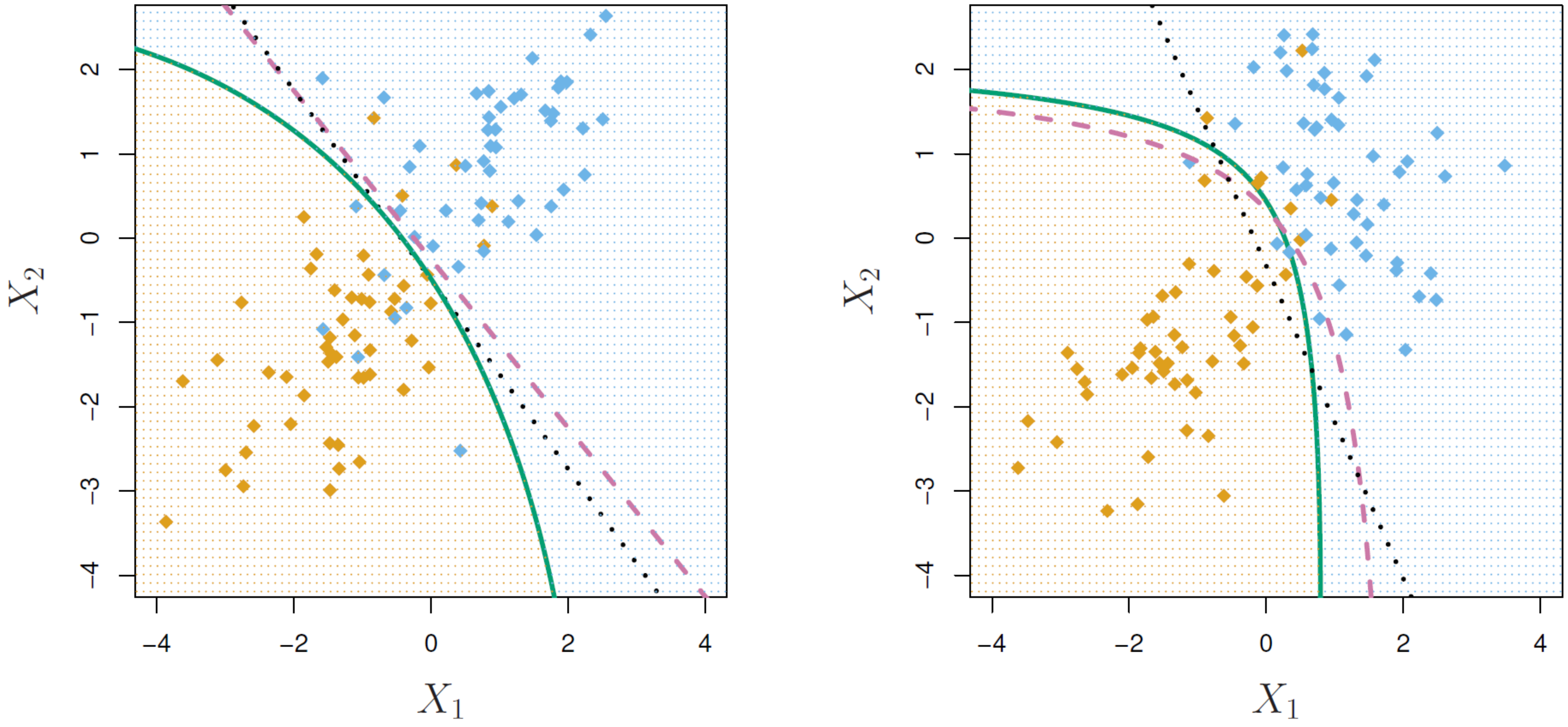


The true positive rate is the $sensitivity$, and the false positive rate is $1 - specificity$

Quadratic discriminant analysis

$$X|Y = k \sim N(\mu_k, \Sigma_k)$$

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



Simulated data. Left: $\Sigma_1 = \Sigma_2$. Right: $\Sigma_1 \neq \Sigma_2$

Naive Bayes

- $f_k(x)$ is the p-dimensional density function for an observation in the kth class
- Estimating a p-dimensional density function is in general challenging

- LDA and QDA replace the problem of estimating K p-dimensional density functions with the much simpler problem of estimating K p-dimensional mean vectors and one or K $(p \times p)$-dimensional covariance matrices

- Naive Bayes assumes that, within the kth class, the p predictors are independent

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

- This leads to

$$p_k(x) = \frac{\pi_k f_{k1}(x_1) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l f_{l1}(x_1) \times \cdots \times f_{lp}(x_p)}$$

Choices of $f_{kj}(x_j)$

- If X_j is quantitative
  - We can assume that $X_j|Y = k \sim N\left(\mu_{jk}, \sigma_{jk}^2\right)$
  - We can also use a non-parametric estimate for $f_{kj}(x_j)$ such as a kernel density estimate
- If X_j is qualitative
  - Simply use the proportion of training observations for X_j corresponding to each class

Class-specific performance

		True default status		
		No	Yes	Total
Predicted default status	No	9615	241	9856
	Yes	52	92	144
	Total	9667	333	10000

TABLE 4.8. *Comparison of the naive Bayes predictions to the true default status for the* $10,000$ *training observations in the* `Default` *data set, when we predict default for any observation for which* $P(Y = \texttt{default}|X = x) > 0.5$.

$\Pr(default = Yes|X = x) > 0.2$

		True default status		
		No	Yes	Total
Predicted	No	9320	128	9448
default status	Yes	347	205	552
	Total	9667	333	10000

TABLE 4.9. *Comparison of the naive Bayes predictions to the true default status for the* $10,000$ *training observations in the* `Default` *data set, when we predict default for any observation for which* $P(Y = \texttt{default}|X = x) > 0.2$.

- The independence assumption introduces some bias, but reduces variance
  - We expect to see a greater pay-off to using naive Bayes relative to LDA or QDA in instances where p is larger or n is smaller