Classification

ddebarr@uw.edu

2017-01-26

"All models are wrong, but some are useful" – George Box

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Resampling Methods
5. Linear Model Selection and Regularization
6. Moving Beyond Linearity
7. Tree-Based Methods
8. Support Vector Machines
9. Unsupervised Learning
10. Neural Networks and Genetic Algorithms

Agenda

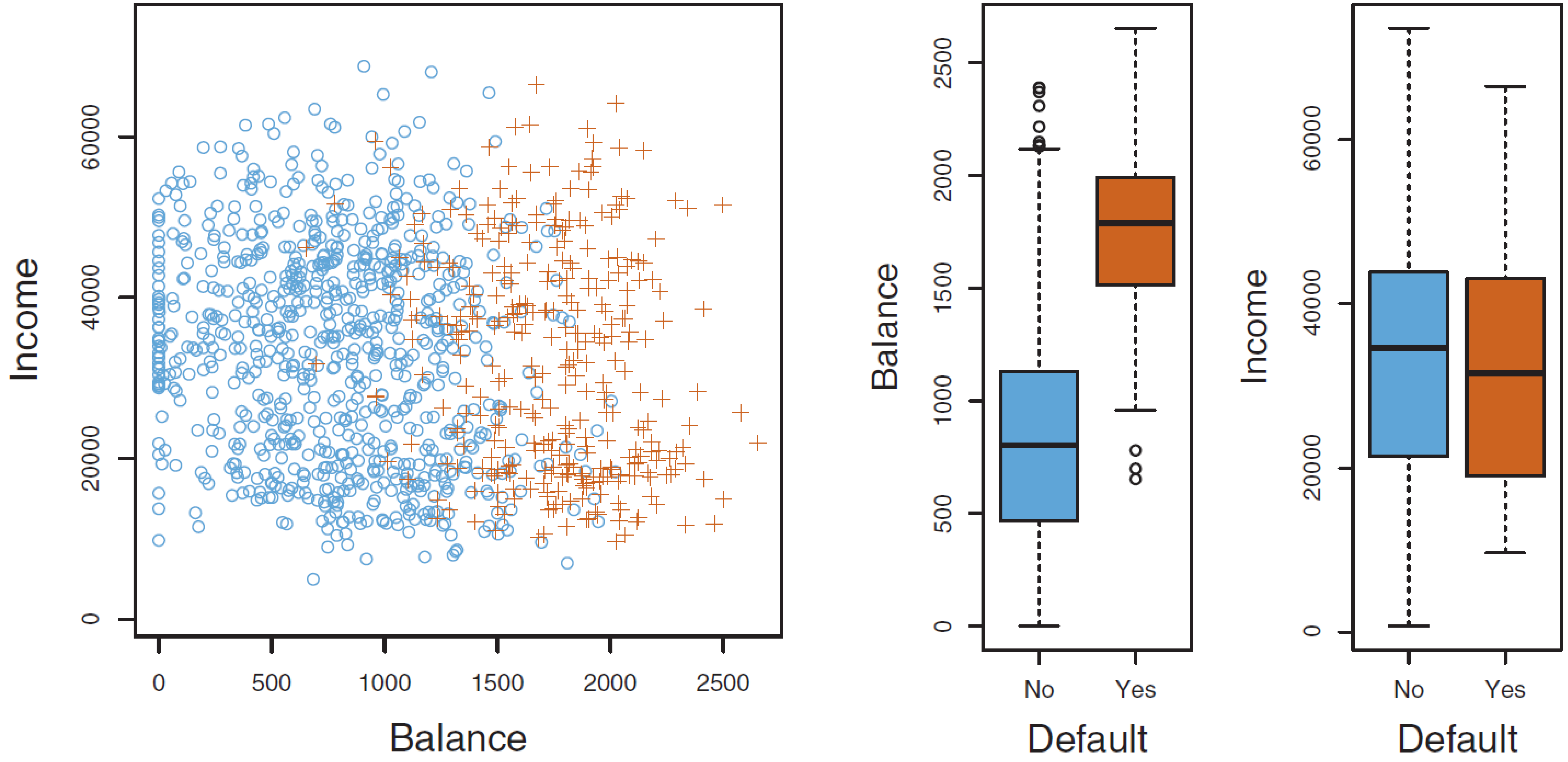
Homework Review

Chapter 4

Classification Examples

- Given a set of symptoms, diagnose medical condition: { Stroke, Drug Overdose, Epileptic Seizure }
- Determine whether an online transaction is fraudulent
- Determine which DNA mutations are associated with a disease

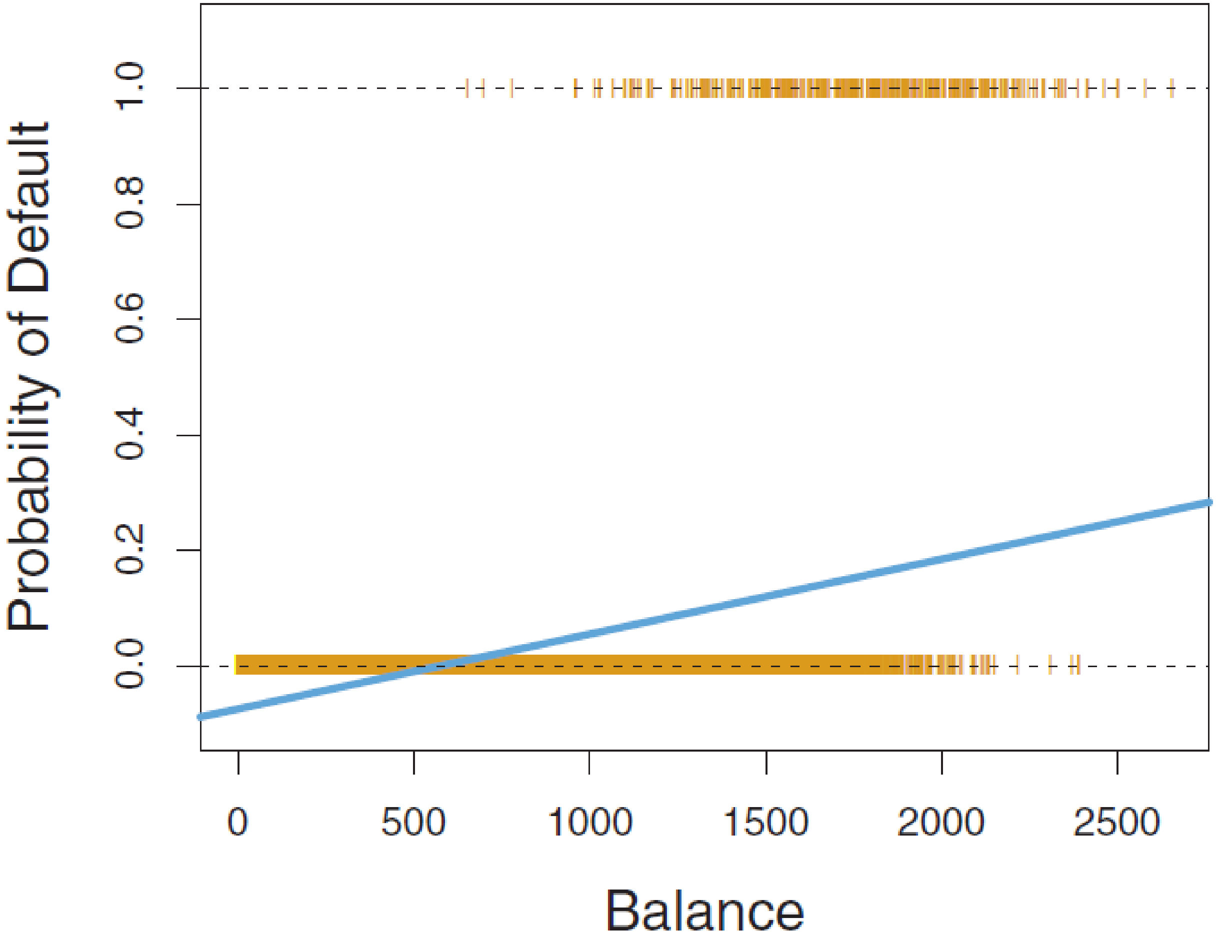
Default Data Set

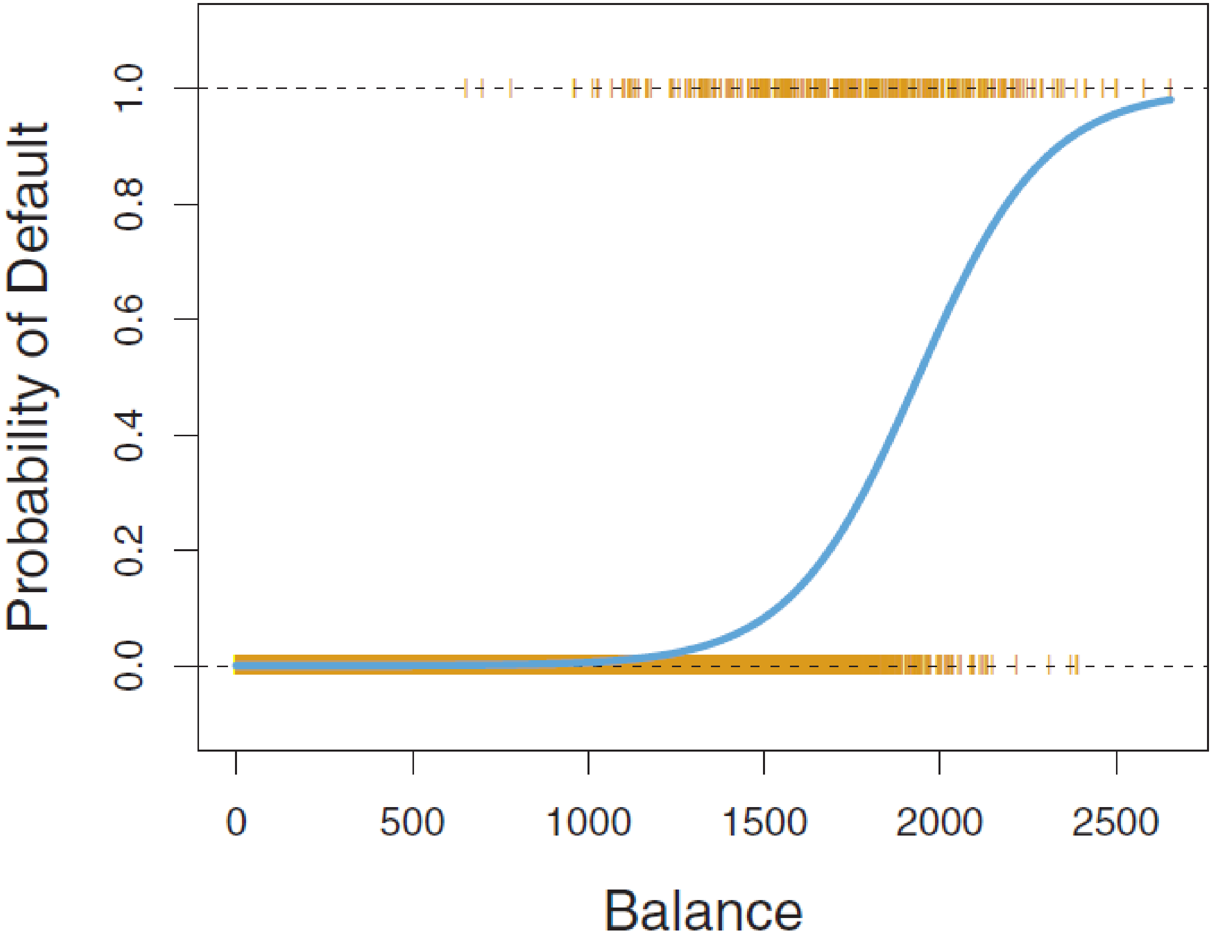


$$\Pr(\texttt{default} = \texttt{Yes} | \texttt{balance})$$

$$p(\texttt{balance})$$

Linear versus Logistic Regression





The Logistic Regression Model

logit function:

$$\log\left(\frac{\Pr(Y=1 \mid x_i)}{1-\Pr(Y=1 \mid x_i)}\right)$$

logistic function:

$$\frac{1}{1+\exp\left(-x_i{}^T\boldsymbol{\beta}\right)}$$

$$\log\left(\frac{\Pr(Y=1 \mid x)}{1-\Pr(Y=1 \mid x)}\right)=x^T\boldsymbol{\beta}$$

$$\frac{\Pr(Y=1 \mid x)}{1-\Pr(Y=1 \mid x)}=\exp\left(x^T\boldsymbol{\beta}\right)$$

$$\Pr(Y=1 \mid x)=\exp\left(x^T\boldsymbol{\beta}\right)\left(1-\Pr(Y=1 \mid x)\right)$$

$$\Pr(Y=1 \mid x)=\exp\left(x^T\boldsymbol{\beta}\right)-\exp\left(x^T\boldsymbol{\beta}\right)\Pr(Y=1 \mid x)$$

$$\Pr(Y=1 \mid x)+\exp\left(x^T\boldsymbol{\beta}\right)\Pr(Y=1 \mid x)=\exp\left(x^T\boldsymbol{\beta}\right)$$

$$\Pr(Y=1 \mid x)\left(1+\exp\left(x^T\boldsymbol{\beta}\right)\right)=\exp\left(x^T\boldsymbol{\beta}\right)$$

$$\Pr(Y=1 \mid x)=\frac{\exp\left(x^T\boldsymbol{\beta}\right)}{1+\exp\left(x^T\boldsymbol{\beta}\right)}$$

$$\Pr(Y=1 \mid x)=\frac{\dfrac{\exp\left(x^T\boldsymbol{\beta}\right)}{\exp\left(x^T\boldsymbol{\beta}\right)}}{\dfrac{1}{\exp\left(x^T\boldsymbol{\beta}\right)}+\dfrac{\exp\left(x^T\boldsymbol{\beta}\right)}{\exp\left(x^T\boldsymbol{\beta}\right)}}$$

$$\Pr(Y=1 \mid x)=\frac{1}{1+\exp\left(-x^T\boldsymbol{\beta}\right)}$$

Log Loss Function

We want to maximize the likelihood function …

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

… which is the same as minimizing the log loss …

$$-\log\left(\Pr(y_i^* = 1 \mid x_i; \boldsymbol{\beta}\right) = -\log\left(\left(\frac{1}{1+\exp\left(-x_i^T\boldsymbol{\beta}\right)}\right)^{y_i^*}\left(1-\frac{1}{1+\exp\left(-x_i^T\boldsymbol{\beta}\right)}\right)^{\left(1-y_i^*\right)}\right)$$

$$= -\log\left(\frac{1}{1+\exp\left(-y_i x_i^T\boldsymbol{\beta}\right)}\right)$$

$$= \log\left(1+\exp\left(-y_i x_i^T\boldsymbol{\beta}\right)\right)$$

$y_i = \{-1, +1\}$ $\qquad y_i^* = \dfrac{y_i + 1}{2}$

Model for Default: Simple Logistic Regression

Example of Simple Logistic Regression (only one predictor):

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Example of a prediction:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 \times 1{,}000}}{1 + e^{-10.6513 + 0.0055 \times 1{,}000}} = 0.00576$$

Model for Default: Simple Logistic Regression

Model:

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

Prediction:

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Model for Default: Multiple Logistic Regression

Model:

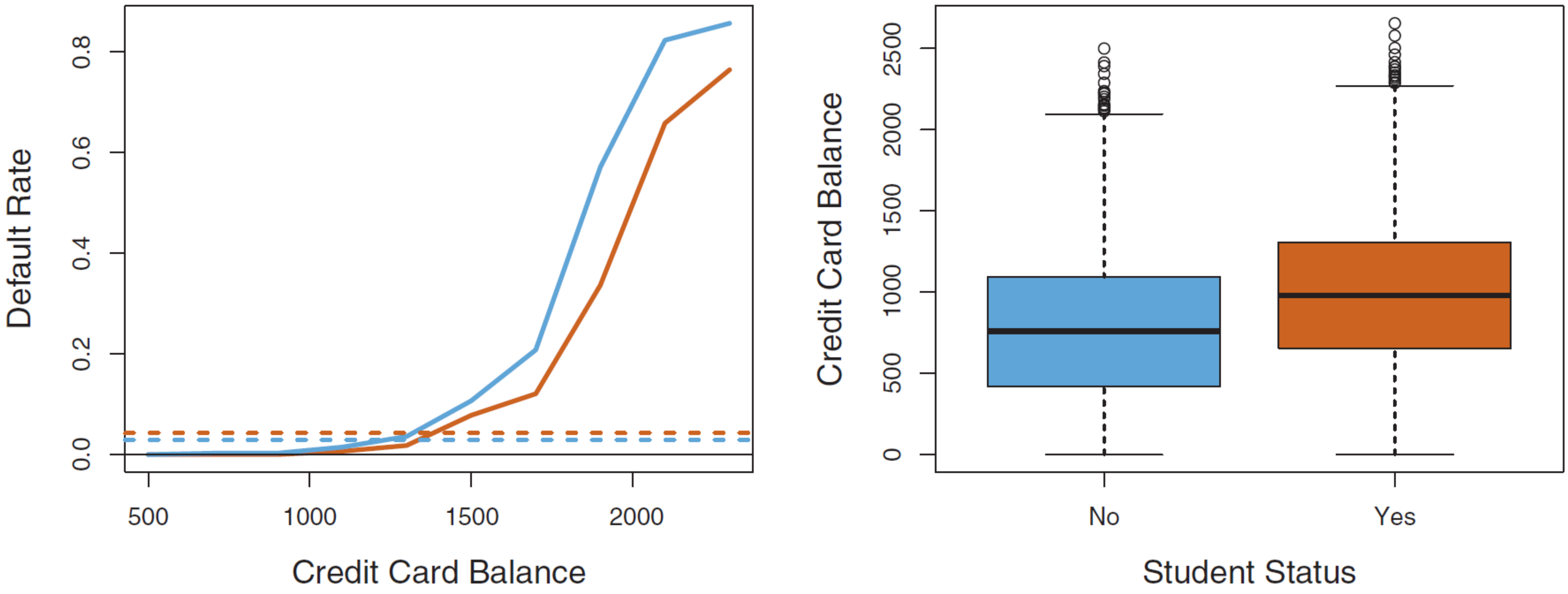
	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Predictions:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105$$

Confounding in the Default Data



In Table 4.2 of your book, we see that the coefficient for the student variable is positive (adds 0.4049 to the log odds); but in Table 4.3, we see that the coefficient for the student variable is negative (subtracts 0.6468 from the log odds). The left-hand side of Figure 4.3 shows students have a higher default rate [the dashed lines]; but for a fixed balance, students tend to have a lower default rate [the solid lines]. The right-hand side shows that students tend to have higher balances.

Iteratively Reweighted Least Squares

```
mydata = read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
model = glm(admit ~ ., data = mydata, family = binomial)
model$coefficients

X = as.matrix(cbind(rep(1, nrow(mydata)), mydata[,2:ncol(mydata)]))
y = mydata$admit
# logistic regression using Iteratively Reweighted Least Squares (IRLS)
beta = as.vector(array(0, ncol(X)))
for (i in 1:25) {
   predictions = 1 / (1 + exp(- (X %*% beta)))
   gradient = t(X) %*% (predictions - y)
   Hessian = t(X) %*% diag(as.vector(predictions * (1 - predictions))) %*% X
   beta = beta - solve(Hessian, diag(ncol(X))) %*% gradient
}
beta
```

See example code at http://cross-entropy.net/ML210/logistic_regression.txt

See example code at bottom of http://cross-entropy.net/ML210/logistic_regression.txt

Gradient Descent for Log Loss

$$-\frac{\partial}{\partial f(x_i)}\log\left(1+\exp\left(-y_i\hat{f}(x_i)\right)\right)=-\frac{1}{1+\exp\left(-y_i\hat{f}(x_i)\right)}\left(\frac{\partial}{\partial\hat{f}(x_i)}1+\frac{\partial}{\partial\hat{f}(x_i)}\exp\left(-y_i\hat{f}(x_i)\right)\right)$$

$$=-\frac{1}{1+\exp\left(-y_i\hat{f}(x_i)\right)}\left(0+\exp\left(-y_i\hat{f}(x_i)\right)\frac{\partial}{\partial\hat{f}(x_i)}\left(-y_i\hat{f}(x_i)\right)\right)$$

$$=-\frac{1}{1+\exp\left(-y_i\hat{f}(x_i)\right)}\left(0+\exp\left(-y_i\hat{f}(x_i)\right)(-y_i)\right)$$

$$=y_i\frac{\exp\left(-y_i\hat{f}(x_i)\right)}{1+\exp\left(-y_i\hat{f}(x_i)\right)}$$

$$=y_i\frac{1}{1+\exp\left(y_i\hat{f}(x_i)\right)}$$

$$=y_i\left(1-\frac{1}{1+\exp\left(-y_i\hat{f}(x_i)\right)}\right)$$

$$=y_i^*-\frac{1}{1+\exp\left(-\hat{f}(x_i)\right)}$$

$$y_i=\{-1,+1\}$$

$$y_i^*=\frac{y_i+1}{2}$$

Logistic Regression for > 2 Response Classes

- Construct K-1 models for K response classes

- For the diagnosis problem { stroke, drug overdose, epileptic seizure } …

$$\Pr(Y = \texttt{stroke}|X)$$

$$\Pr(Y = \texttt{drug overdose}|X)$$

$$1 - \Pr(Y = \texttt{stroke}|X) - \Pr(Y = \texttt{drug overdose}|X)$$

Why Cover More Than Logistic Regression?

- Estimates for the regression coefficients are “surprisingly” unstable when the classes are well separated
- If ‘n’ is small and the distribution of predictors is approximately Gaussian, the linear discriminant model is more stable
- Linear Discriminant Analysis (LDA) is popular when we have more than two classes

Using Bayes Theorem for Classification

$$\text{Posterior} = \frac{\text{Prior * Likelihood}}{\text{Evidence}}$$

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

Linear Discriminant Analysis (LDA) for p=1

- The term "discriminant" is just another name for a classifier; however, the term "Linear Discriminant Analysis" refers to the use of a Gaussian density function for estimating likelihood values

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

- The Linear Discriminant Analysis model is considered to be a <u>generative</u> classifier because it uses p(**x** | y) to estimate p(y | **x**), while the Logistic Regression model is considered to be a <u>discriminative</u> classifier because it does not use p(**x** | y) to estimate p(y | **x**)

Linear Discriminant Analysis

For p=1, the posterior probability is computed as follows

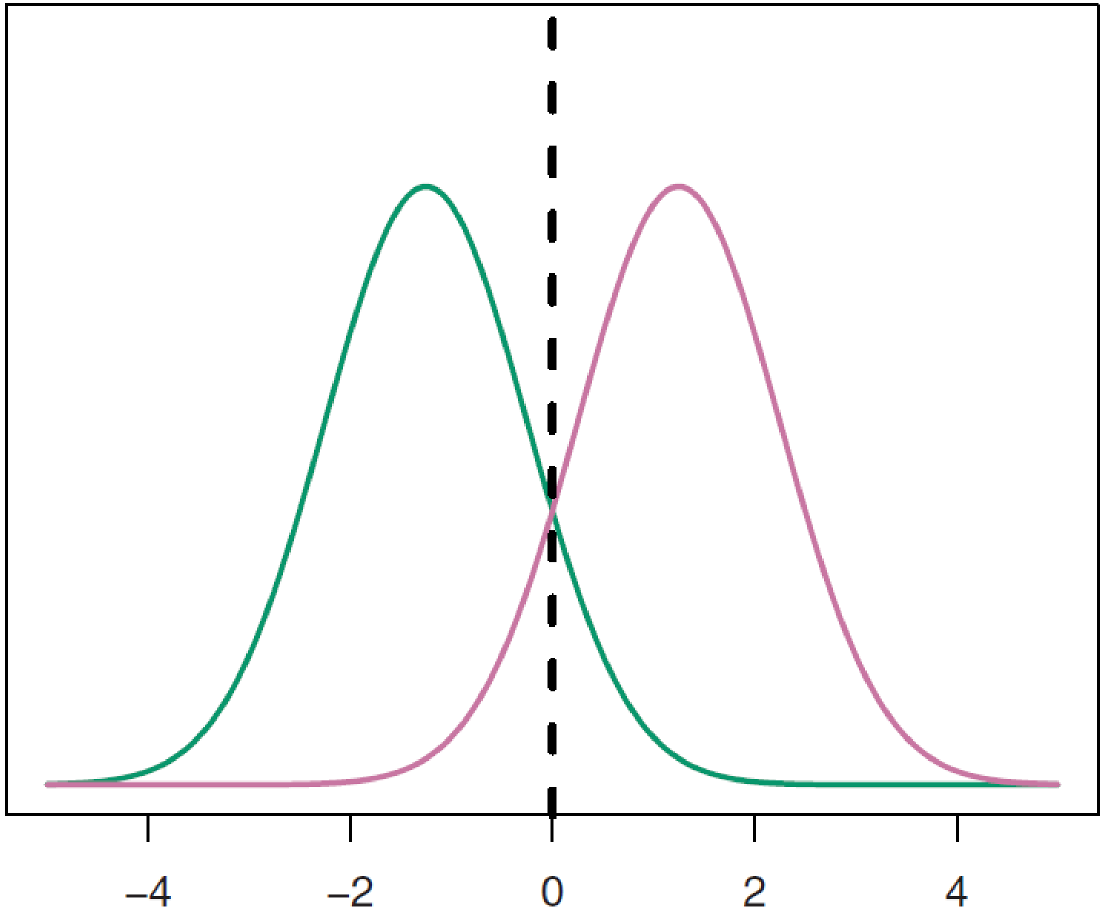
$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

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

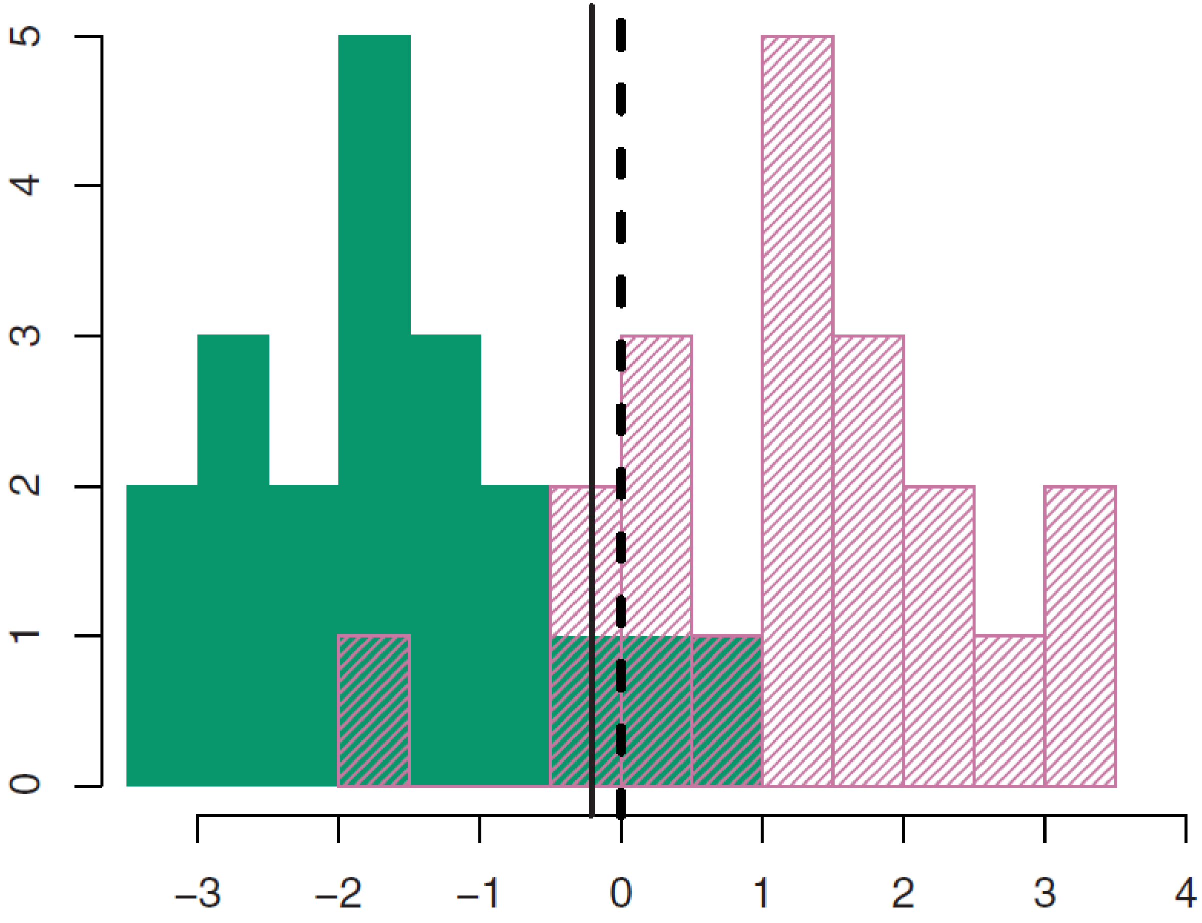
$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

decision boundary for a binary classifier
iff the priors are equal

Theory versus Practice



Bayes Classifier



Classifier Based on Sample Data

Linear Discriminant Analysis for p=1

Parameter Estimates for Each Class:

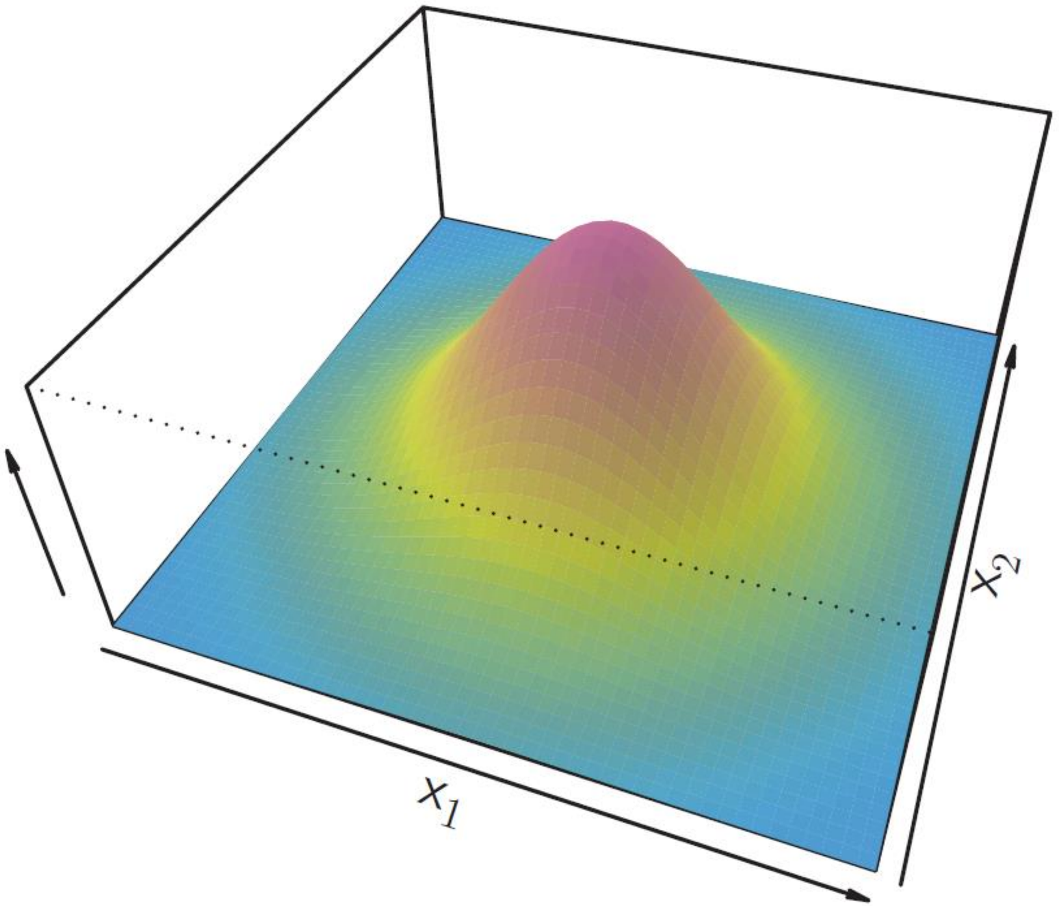
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

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

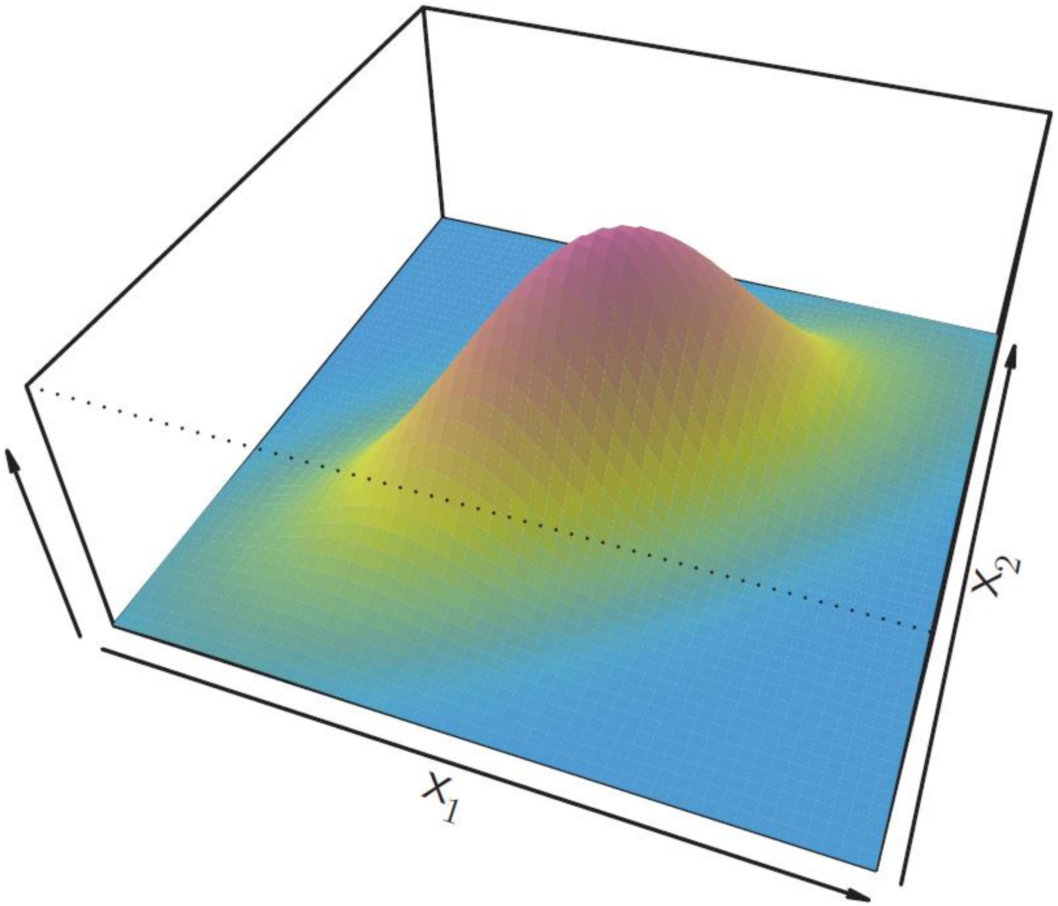
$$\hat{\pi}_k = n_k/n$$

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

Multivariate Gaussian Distribution Examples

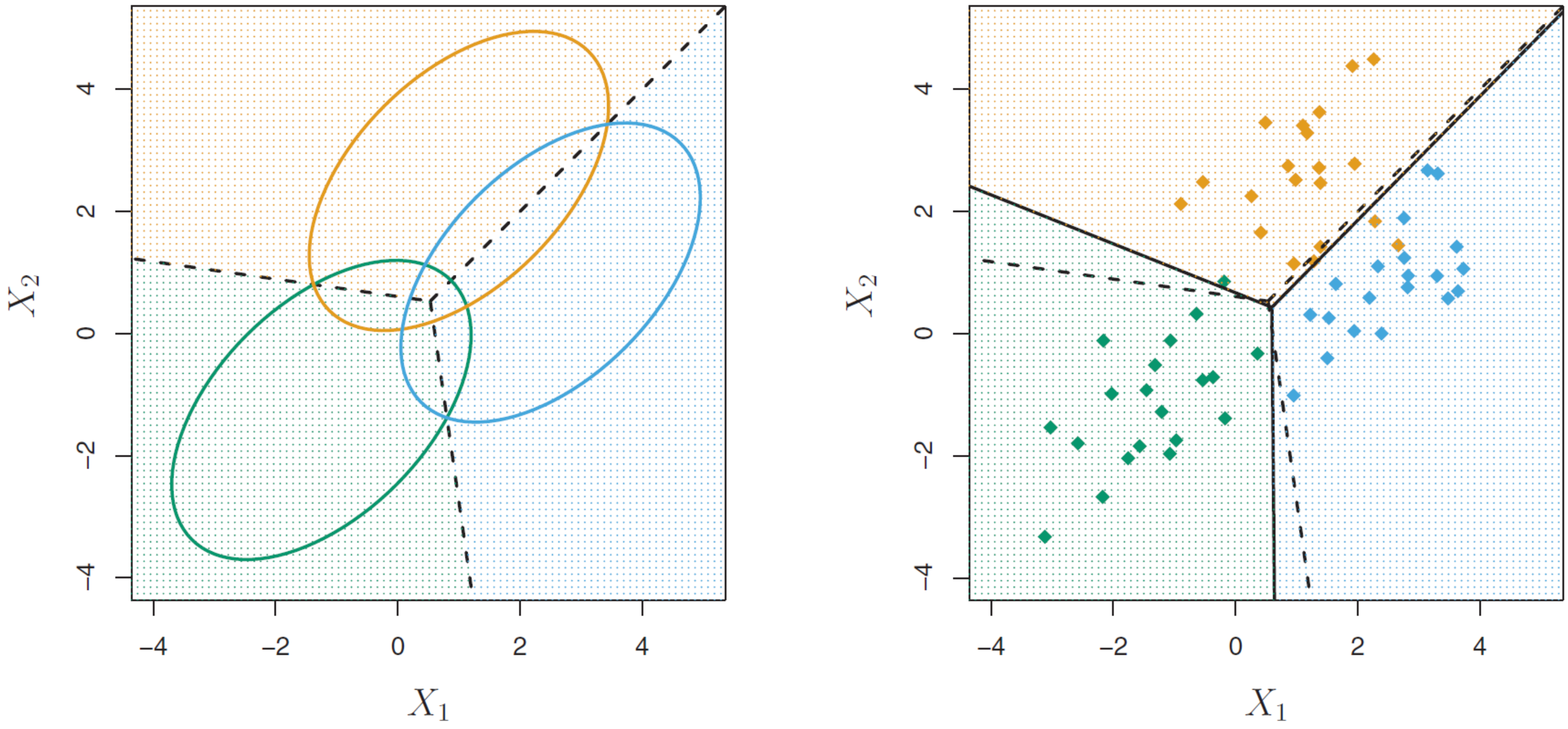


Correlation = 0



Correlation = 0.7

LDA with k=3 (classes) and p=2 (predictors)



Linear Discriminant Analysis with p > 1

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

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

decision boundary for a binary classifier *iff* their priors are equal …

$$x^T \boldsymbol{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1}\mu_k = x^T \boldsymbol{\Sigma}^{-1}\mu_l - \frac{1}{2}\mu_l^T \boldsymbol{\Sigma}^{-1}\mu_l$$

Example Confusion Matrices for LDA

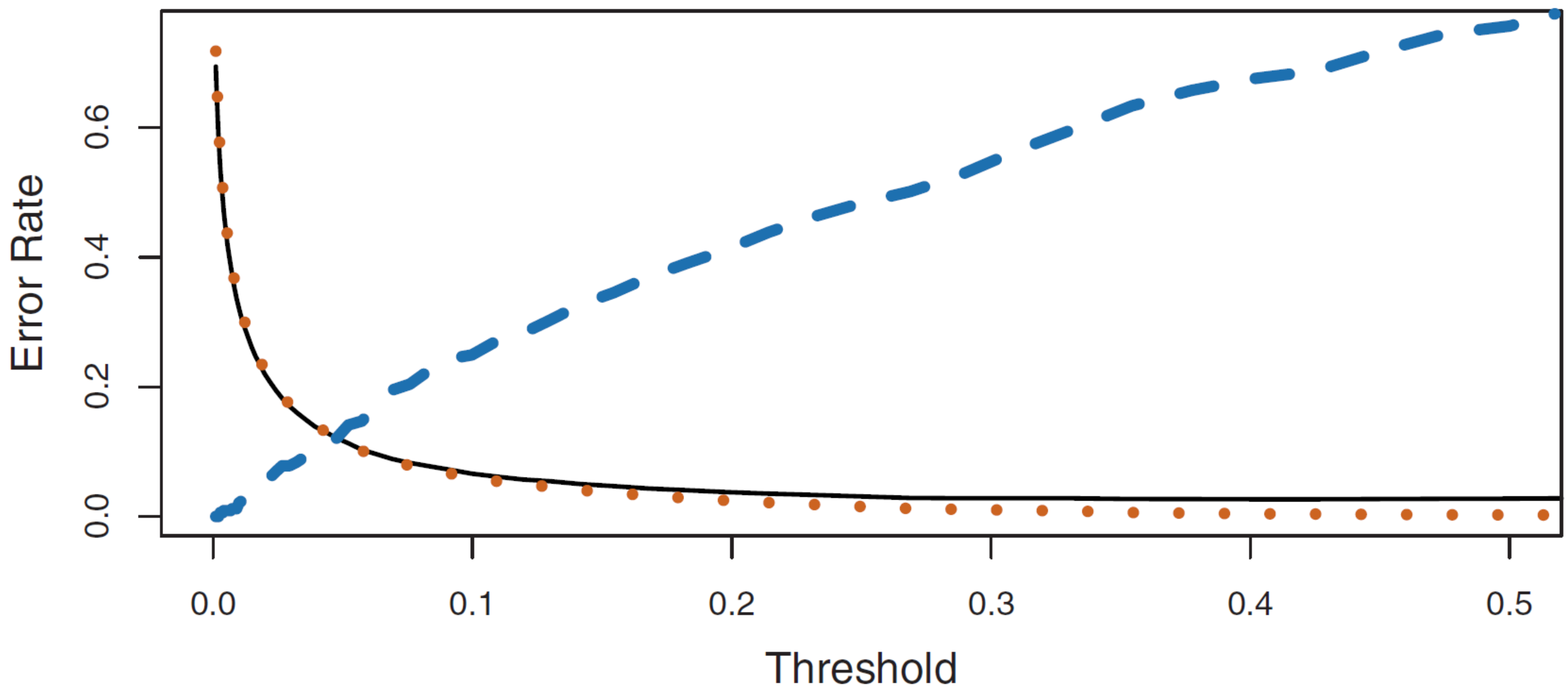
Classification threshold
Pr(Default = Yes | X = x) > 0.5

		True default status No	*True default status* Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Classification threshold
Pr(Default = Yes | X = x) > 0.2

		True default status No	*True default status* Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

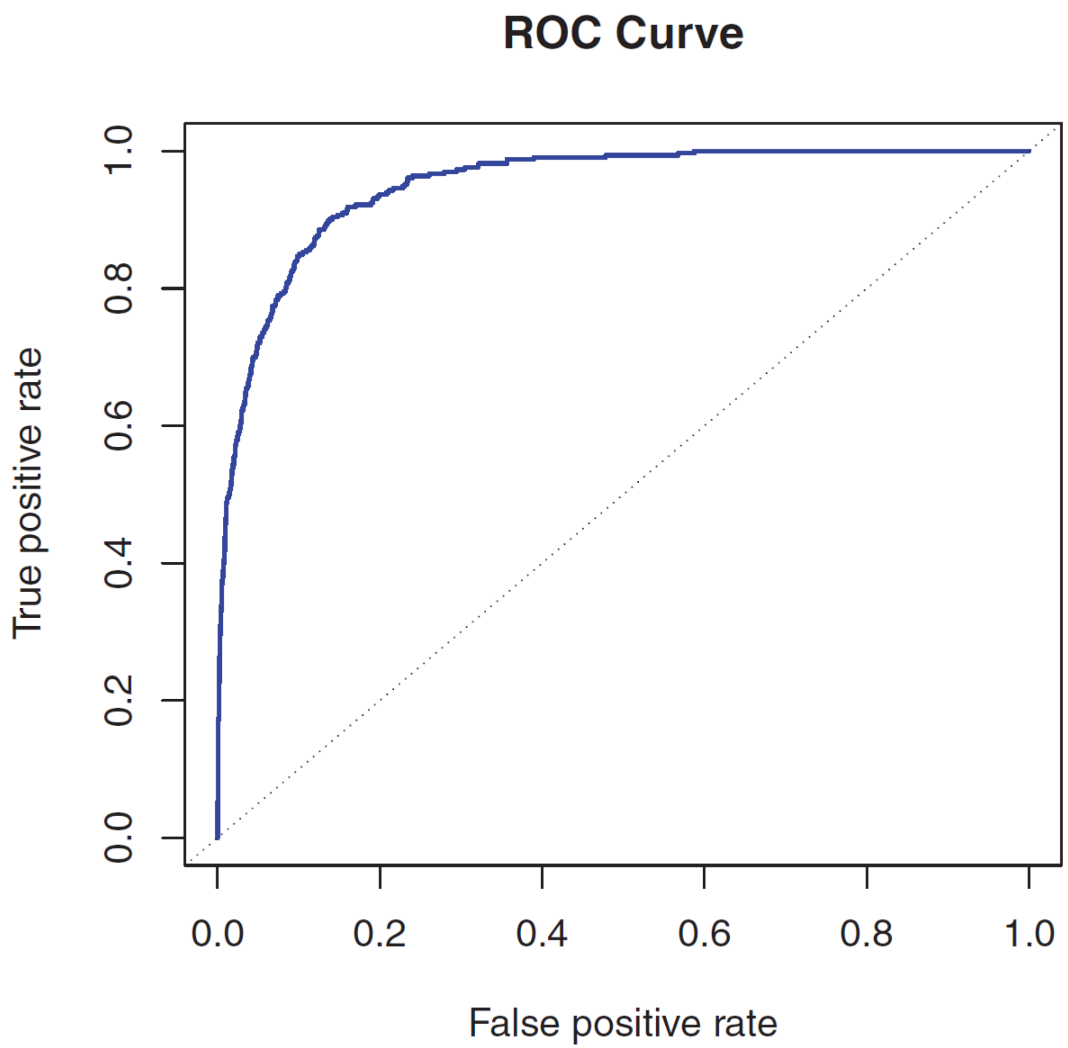
Error Rate as a Function of Threshold for LDA



Solid black line: overall error rate
Dotted red line: error rate for non-defaulting customers
Dashed blue lines: error rate for defaulting customers

Receiver Operating Characteristic (ROC) Curve for LDA

Area Under the Curve = 0.95



Notional Confusion Matrix for Binary Classification

Note: as shown below, I more commonly see the true class along the rows [suggestion: stick with the same format]

		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*	

Common Classification Metrics

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*	

When reporting metrics for a classification problem with more than two classes, either macro (unweighted) averages can be used or micro (weighted) averages can be used

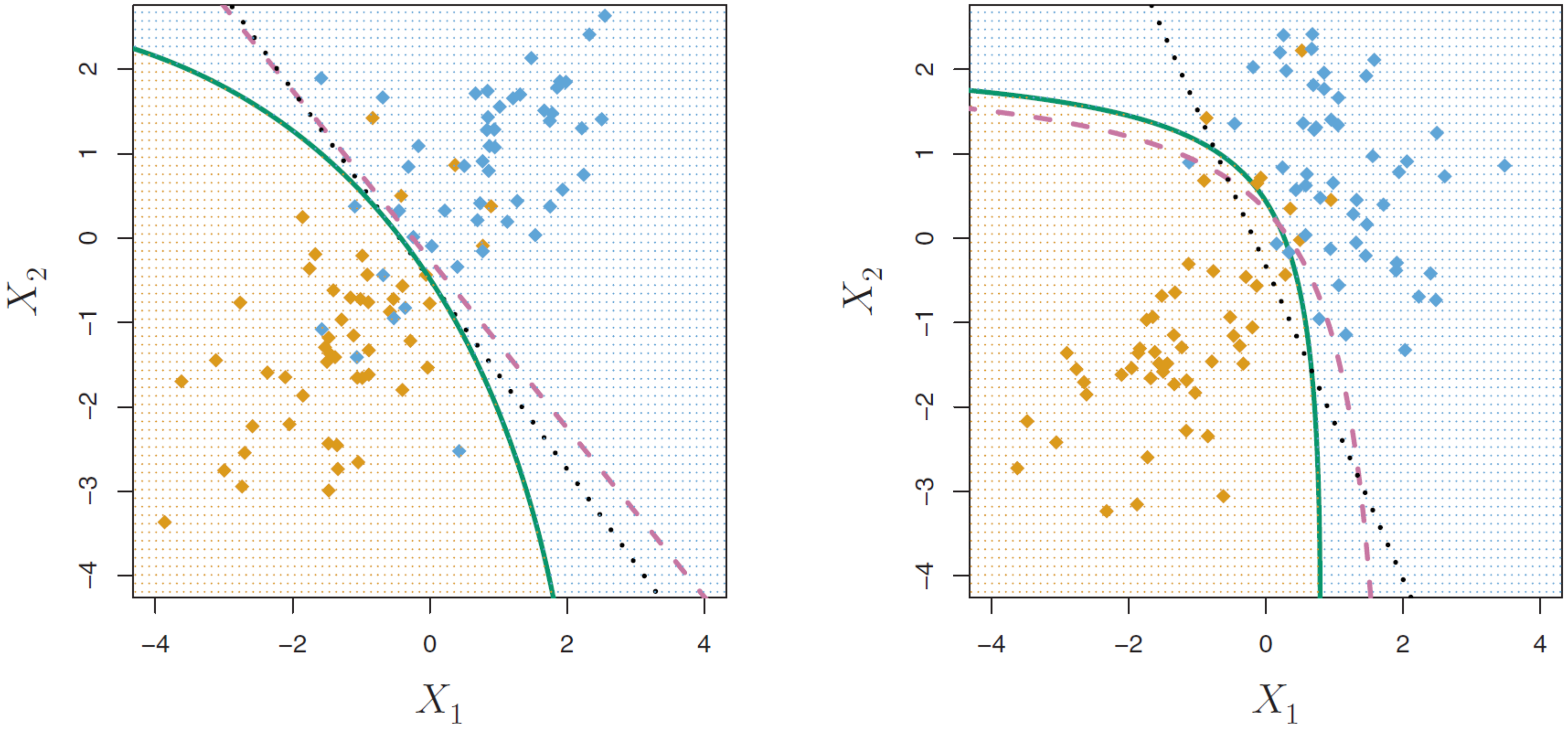
Quadratic Discriminant Analysis (QDA)

- QDA versus LDA
  - for LDA, a single covariance matrix is used for all classes
  - for QDA, a covariance matrix is estimated for each class [this allows for a non-linear boundary]

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

Bayes versus LDA versus QDA

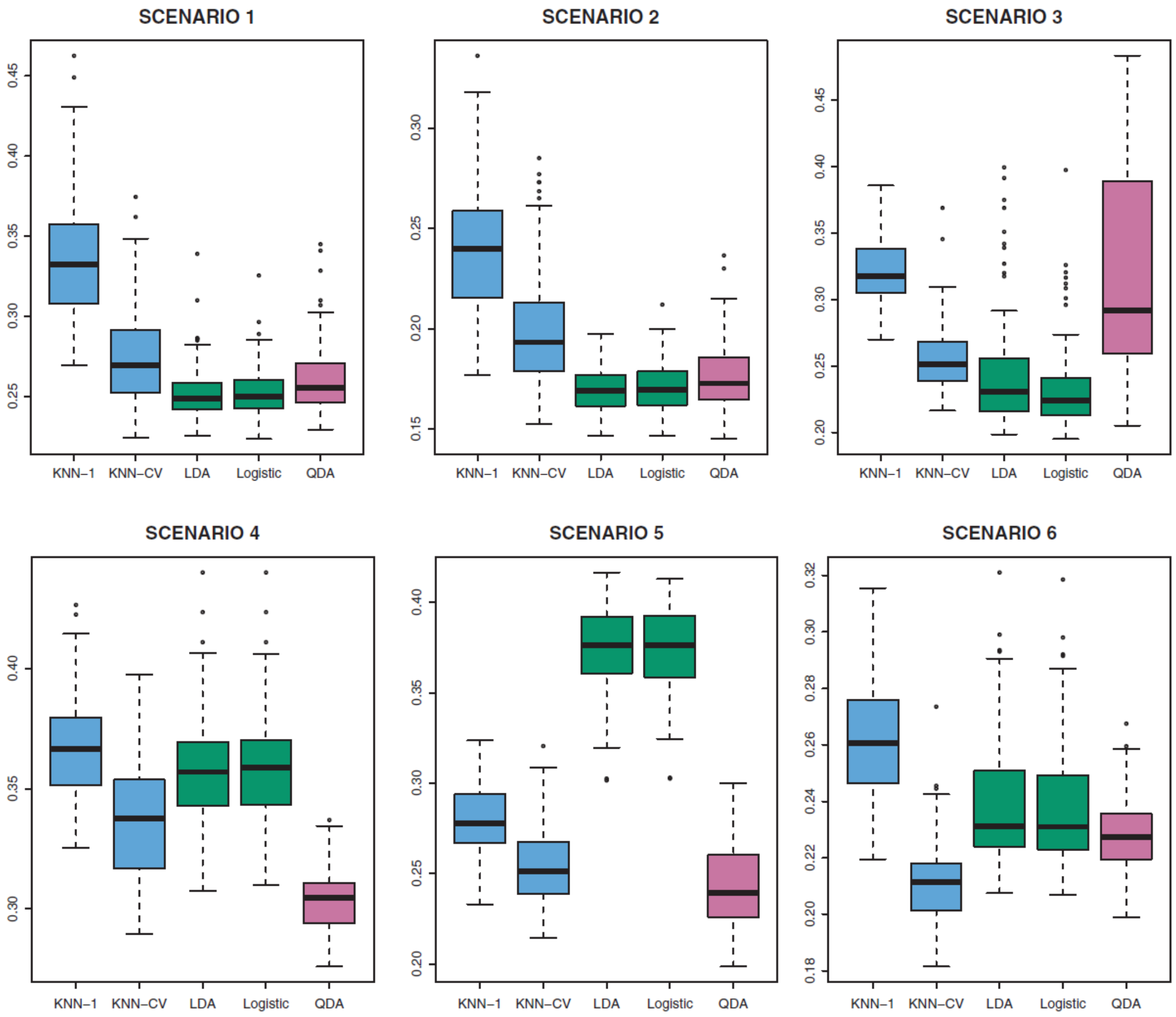
LDA is better



QDA is better

Dashed red line: Bayes (optimal) decision boundary
Dotted black line: LDA decision boundary
Solid green line: QDA decision boundary

Comparison of Classification Methods



Agenda

Homework Review

KNN for Regression (from last week)

Robust Regression

Gradient Descent

Chapter 4