Classification

Parametric approach

- The linear regression model assumes that the response variable Y is quantitative. But in many situations, the response variable is instead qualitative.
- For example, eye color is qualitative, taking on values blue, brown, or green.
- Often qualitative variables are referred to as categorical; we will use these terms interchangeably.
- We study approaches for predicting qualitative responses, a process that is known as classification.
- Predicting a qualitative response for an observation can be referred to as classifying that observation.
- On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.

An Overview of Classification

- There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response.
- We discuss three of the most widely-used classifiers:
  1. logistic regression,
  2. linear discriminant analysis,
  3. K-nearest neighbors.

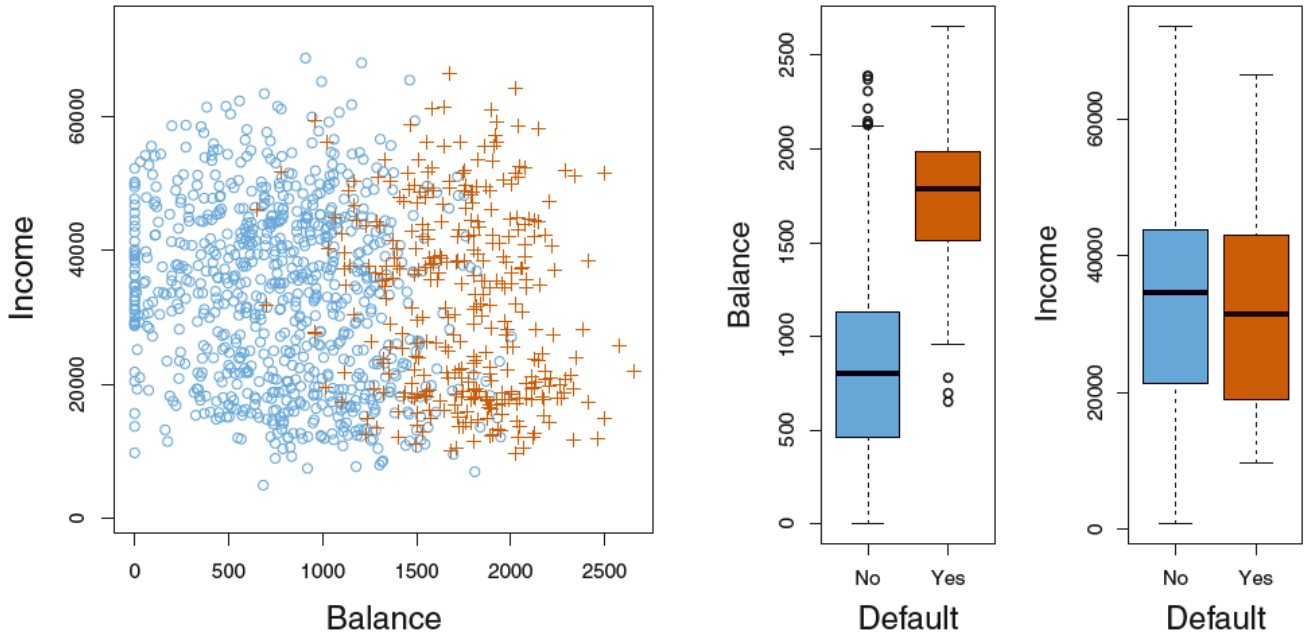
- We discuss more computer-intensive methods later

An Overview of Classification

- Classification problems occur often, perhaps even more so than regression problems. Some examples include:
    1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
    2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
    3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

Classification: some details

- Just as in the regression setting, in the classification setting we have a set of training observations $(x_1, y_1),\ldots, (x_n, y_n)$ that we can use to build a classifier.
- We want our classifier to perform well not only on the training data, but also on test observations that were not used to train the classifier.
- It is worth noting that the figure below displays a very pronounced relationship between the predictor balance and the response default. In most real applications, the relationship between the predictor and the response will not be nearly so strong. However, for the sake of illustrating the classification procedures discussed in this chapter, we use an example in which the relationship between the predictor and the response is somewhat exaggerated.



Why Not Linear Regression?

- Linear regression is not appropriate in the case of a qualitative response. Why not?
- Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms. There are three possible diagnoses: stroke, drug overdose, and epileptic seizure. We could consider encoding these values as a quantitative response variable, Y, as follows:

$$Y = \begin{cases} 1 \text{ if stroke} \\ 2 \text{ if drug overdose} \\ 3 \text{ if epileptic seizure} \end{cases}$$

- Using this coding, least squares could be used to fit a linear regression model to predict Y
- Unfortunately, this coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure. In practice there is no particular reason that this needs to be the case.
- For instance, one could choose an equally reasonable coding,

$$Y = \begin{cases} 1 \text{ if epileptic seizure} \\ 2 \text{ if stroke} \\ 3 \text{ if drug overdose} \end{cases}$$

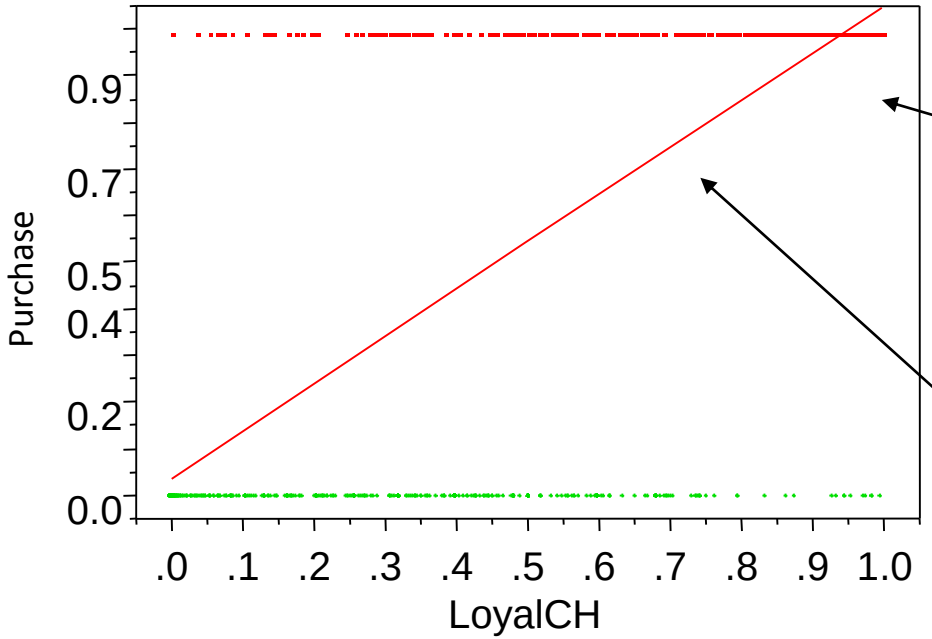
Why Not Linear Regression?

- If the response variable's values did take on a natural ordering, such as mild, moderate, and severe, and we felt the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable.
- Unfortunately, in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.
- For a binary (two level) qualitative response, the situation is better.
- For perhaps there are only two possibilities for the patient's medical condition: stroke and drug overdose. We could then code the response as follows: Y = 0 if stroke; 1 if drug overdose.
- We could then fit a linear regression to this binary response, and predict drug overdose if $\hat{Y} > 0.5$ and stroke otherwise.
- In the binary case it is not hard to show that even if we flip the above coding, linear regression will produce the same final predictions.
- However, the dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels.
- For these reasons, it is preferable to use a classification method that is truly suited for qualitative response values, such as the ones presented next.

The Logistic Model

- We would like to predict what customers prefer to buy: Citrus Hill or Minute Maid orange juice?
- The Y (Purchase) variable is categorical: 0 or 1
- The X (LoyalCH) variable is a numerical value (between 0 and 1) which specifies the how much the customers are loyal to the Citrus Hill (CH) orange juice
- Can we use Linear Regression when Y is categorical?
- The regression line $\beta_0+\beta_1 X$ can take on any value between negative and positive infinity
- In the orange juice classification problem, Y can only take on two possible values: 0 or 1.
- Therefore the regression line almost always predicts the wrong value for Y in classification problems

The Logistic Model



How do we interpret values greater than 1?

How do we interpret values of Y between 0 and 1?

A Fundamental Picture

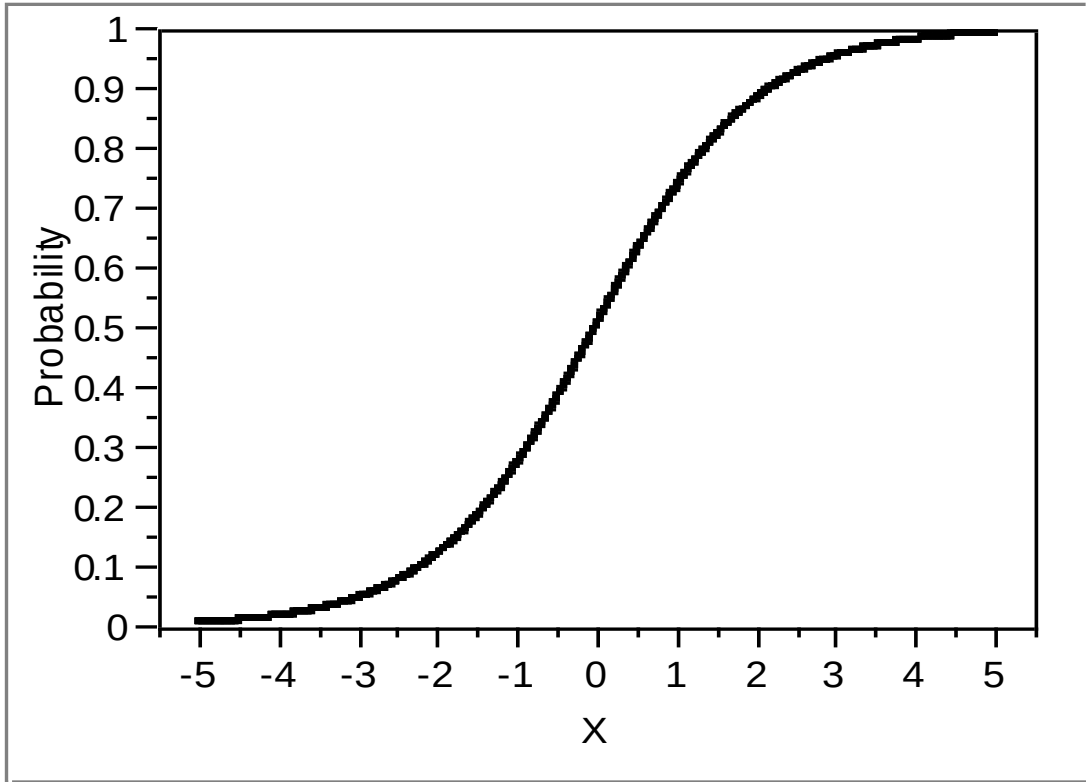
- The regression line $\beta_0+\beta_1 X$ can take on any value between negative and positive infinity
- In the orange juice classification problem, Y can only take on two possible values: 0 or 1.
- Therefore the regression line almost always predicts the wrong value for Y in classification problems .

Solution: Use Logistic Function

- Instead of trying to predict Y, let's try to predict P(Y = 1), i.e., the probability a customer buys Citrus Hill (CH) juice.
- Thus, we can model P(Y = 1) using a function that gives outputs between 0 and 1.
- We can use the logistic function
- Logistic Regression!

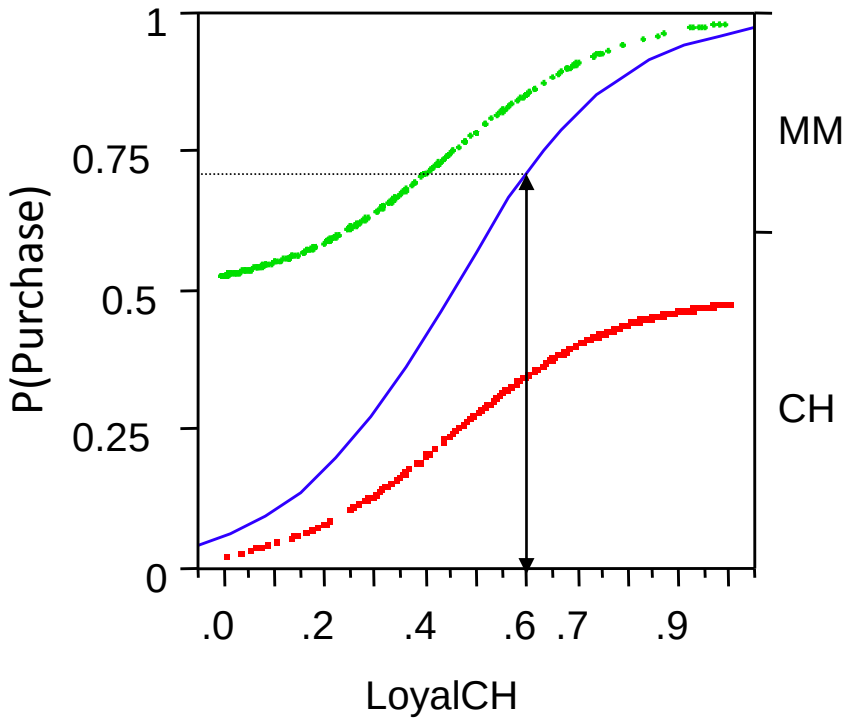
Logistic Function

$$p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$



Logistic Regression

- Logistic regression is very similar to linear regression
- We come up with b_0 and b_1 to estimate β_0 and β_1.
- We have similar problems and questions as in linear regression
  - e.g. Is β_1 equal to 0? How sure are we about our guesses for β_0 and β_1?



If LoyalCH is about .6 then Pr(CH) ≈ .7.

Maximum Likelihood

- In logistic regression, we use the *logistic function*

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- To fit the model, we use a method called *maximum likelihood*

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

- The quantity $p(X)/[1-p(X)]$ is called the *odds*, and can take on any value between 0 and ∞.

Odds

- Values of the odds close to 0 and ∞ indicate very low and very high probabilities of default, respectively.
- For the credit card default example, on average 1 in 5 people with an odds of 1/4 will default, since $p(X) = 0.2$ implies an odds of

$$\frac{0.2}{1-0.2} = 1/4$$

- Likewise on average nine out of every ten people with an odds of 9 will default, since $p(X) = 0.9$ implies an odds of 0.9

$$\frac{0.9}{1-0.9} = 9$$

- Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

Log-odds

- By taking the logarithm of both sides, we arrive at

$$\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- The left-hand side is called the *log-odds* or *logit*.
- We see that the logistic regression model has a logit that is linear in X.

Estimating the Regression Coefficients

- The coefficients β_0 and β_1 are unknown, and must be estimated based on the available training data.
- In Chapter 3, we used the least squares approach to estimate the unknown linear regression coefficients
- Although we could use (non-linear) least squares to fit the model, the more general method of *maximum likelihood* is preferred

likelihood function

- The basic intuition behind using maximum likelihood to fit a logistic regression model is as follows:
- we seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual, corresponds as closely as possible to the individual's observed default status.
- In other words, we try to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging these estimates into the model for $p(X)$, yields a number close to 1 for all individuals who defaulted, and a number close to 0 for all individuals who did not.

likelihood function

- We use a *likelihood function*:

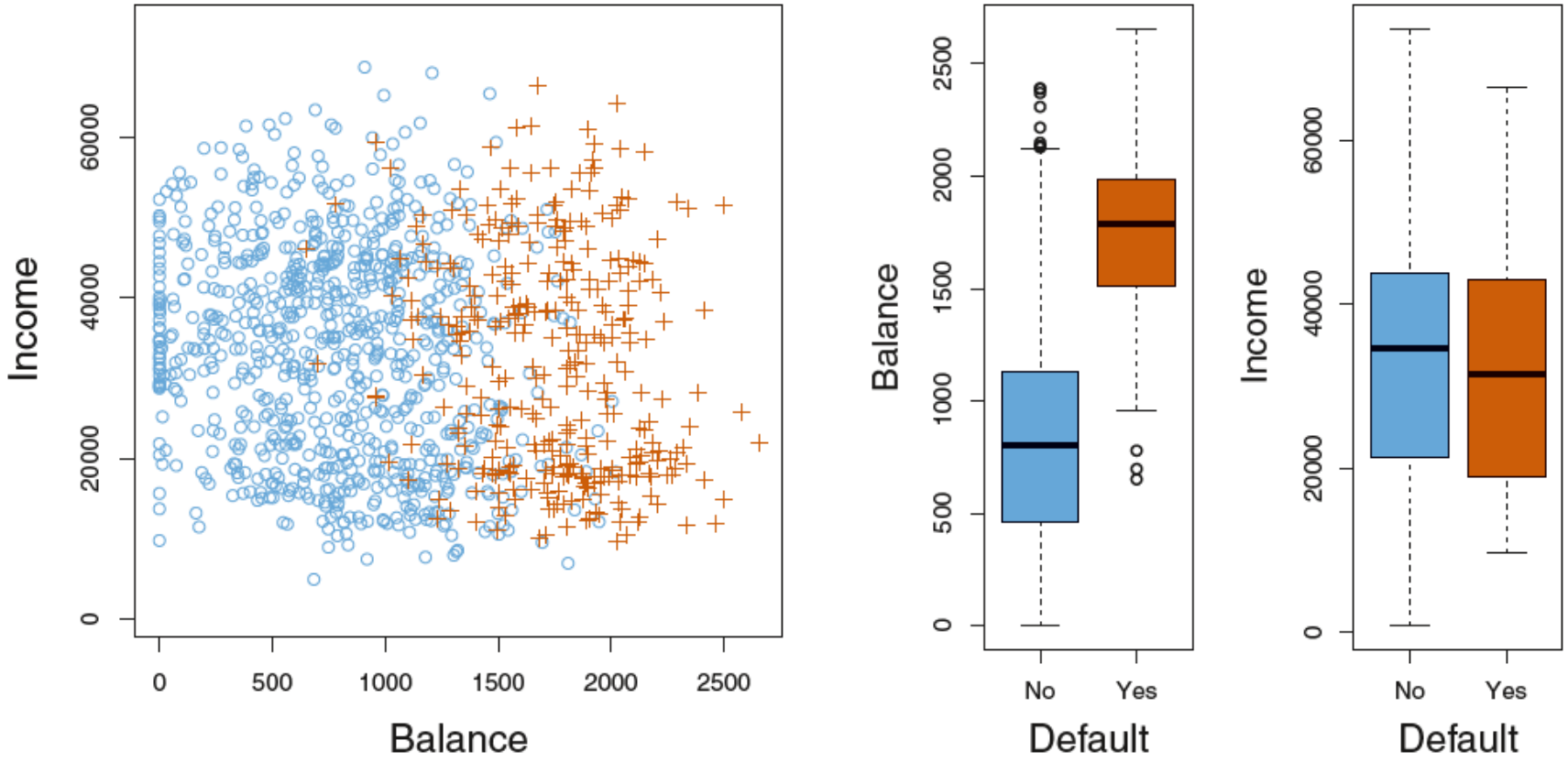
$$l(\beta_0, \beta_1) = \prod_{i:\, y_i=1} p(x_i) \prod_{i':\, y_{i'}=1} (1 - p(x_{i'}))$$

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to *maximize* this likelihood function.

Case 2: Credit Card Default Data

- We would like to be able to predict customers that are likely to default
- Possible X variables are:
  - Annual Income
  - Monthly credit card balance
- The Y variable (Default) is categorical: Yes or No
- How do we check the relationship between Y and X?

The Default Dataset

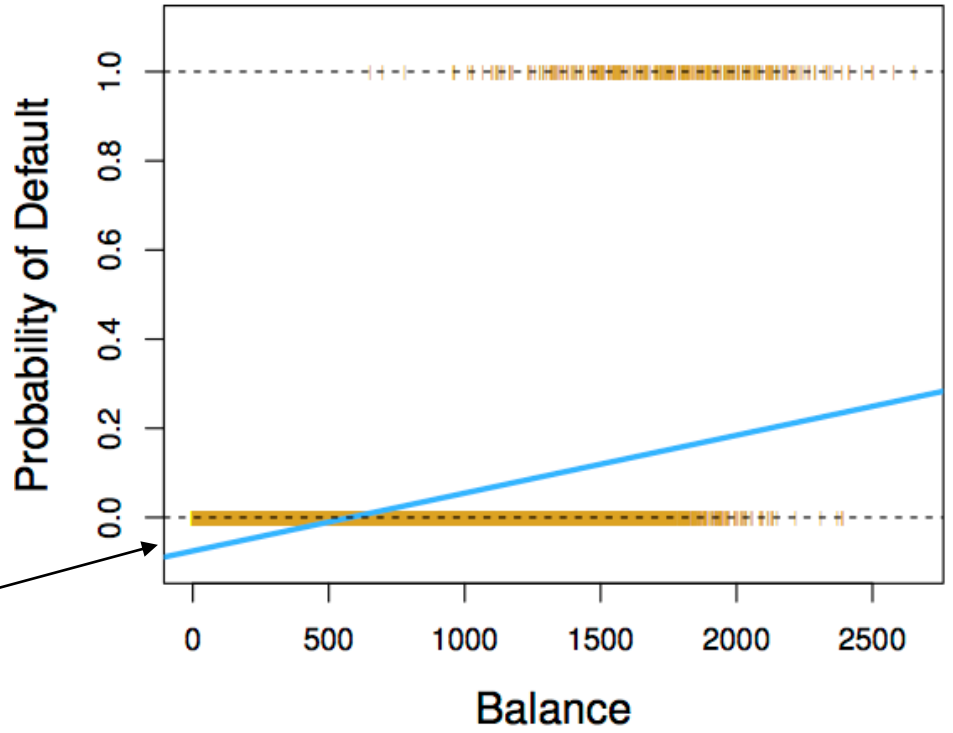


Default Dataset

If we fit a linear regression to the Default data, then for very low balances we predict a negative probability, and for high balances we predict a probability above 1!

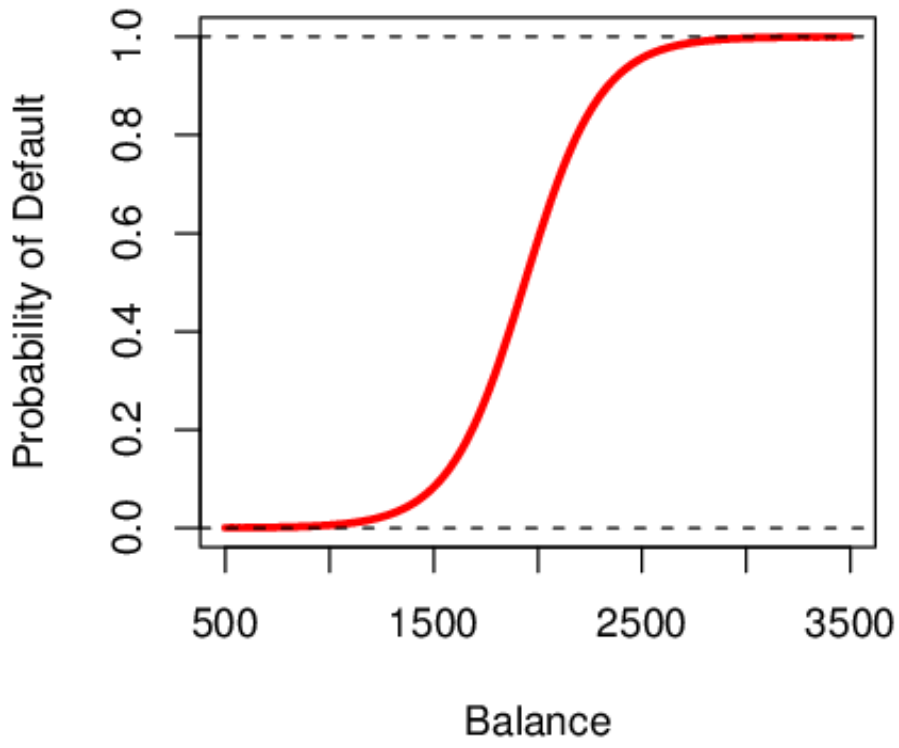
When Balance < 500,

Pr(default) is negative!



Logistic Function on Default Data

- Now the probability of default is close to, but not less than zero for low balances. And close to but not above 1 for high balances



Interpreting β_1

- Interpreting β_1 is not very easy with logistic regression, simply because we are predicting $P(Y)$ and not Y.
- If β_1 =0, this means that there is no relationship between Y and X.
- If β_1 >0, this means that when X gets larger so does the probability that Y = 1.
- If β_1 <0, this means that when X gets larger, the probability that Y = 1 gets smaller.
- But how much bigger or smaller depends on where we are on the slope

Are the coefficients significant?

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a Z test instead of a T test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and b_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

	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

Making Prediction

- Suppose an individual has an average balance of $1000. What is their probability of default?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$

- The predicted probability of default for an individual with a balance of $1000 is less than 1%.
- For a balance of $2000, the probability is much higher, and equals to 0.586 (58.6%).

Qualitative Predictors in Logistic Regression

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable “Student” coded as (Student = 1, Non-student =0).
- b_1 is positive: This indicates students tend to have higher default probabilities than non-students

	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

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

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

Multiple Logistic Regression

- We can fit multiple logistic just like regular regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

Multiple Logistic Regression- Default Data

- Predict Default using:
  - Balance (quantitative)
  - Income (quantitative)
  - Student (qualitative)

	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

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}}{1+e^{-10.869+0.00574\times1500+0.003\times40-0.6468\times1}} = 0.058.$$

An Apparent Contradiction

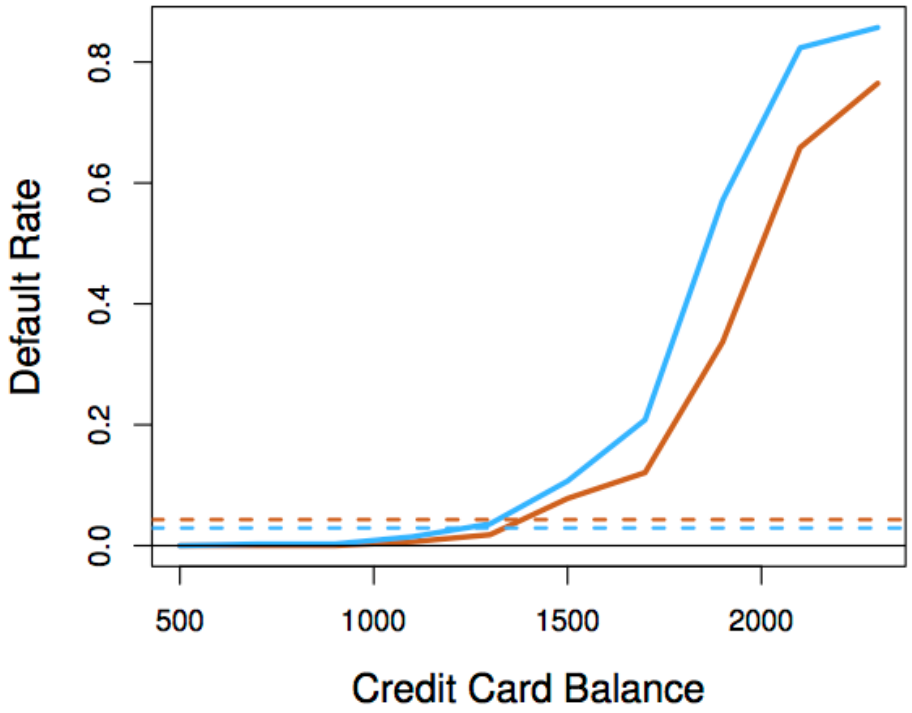
	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

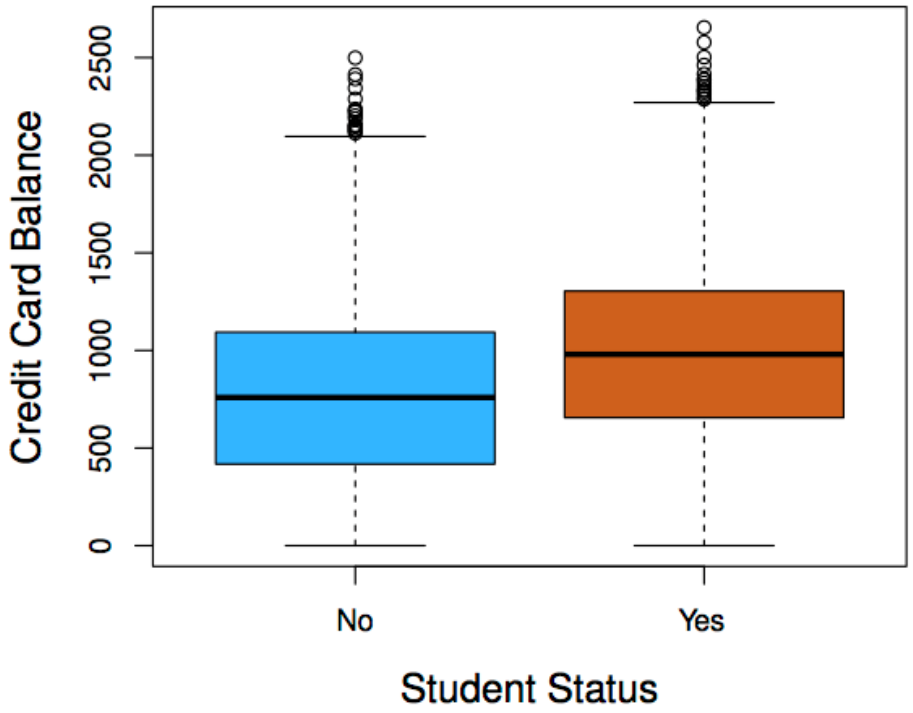
Positive

	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

Negative

Students (Orange) vs. Non-students (Blue)





To whom should credit be offered?

To whom should credit be offered?

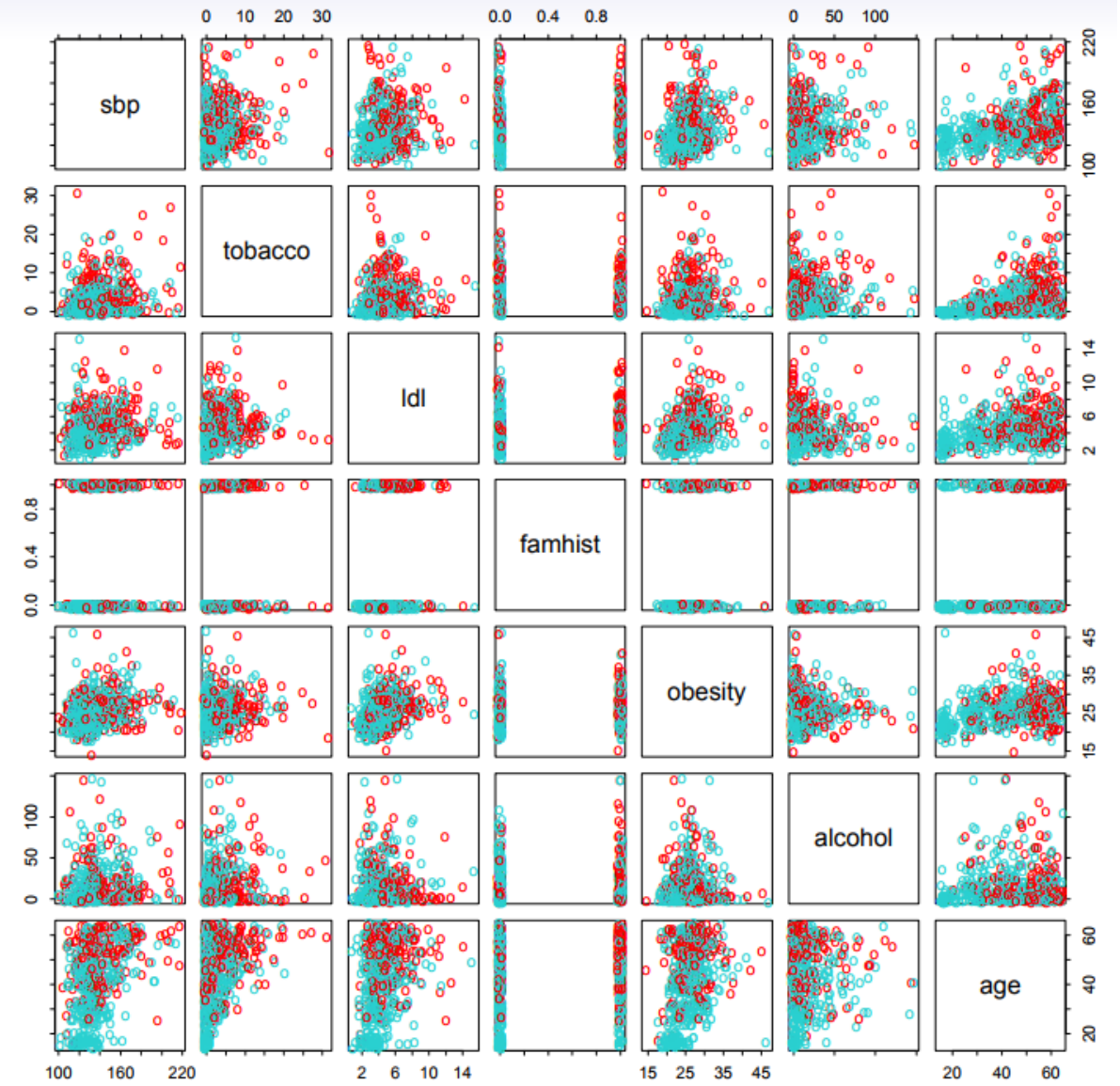
- A student is risker than non students if no information about the credit card balance is available

- However, that student is less risky than a non student with the same credit card balance!

Example: South African Heart Disease

- 160 cases of MI (myocardial infarction) and 302 controls (all male in age range 15-64), from Western Cape, South Africa in early 80s.
- Overall prevalence very high in this region: 5.1%.
- Measurements on seven predictors (risk factors), shown in scatterplot matrix.
- Goal is to identify relative strengths and directions of risk factors.
- This was part of an intervention study aimed at educating the public on healthier diets.

South African Heart Disease



Scatterplot matrix of the South African Heart Disease data. The response is color coded — The cases (MI) are red, the controls turquoise. famhist is a binary variable, with 1 indicating family history of MI.

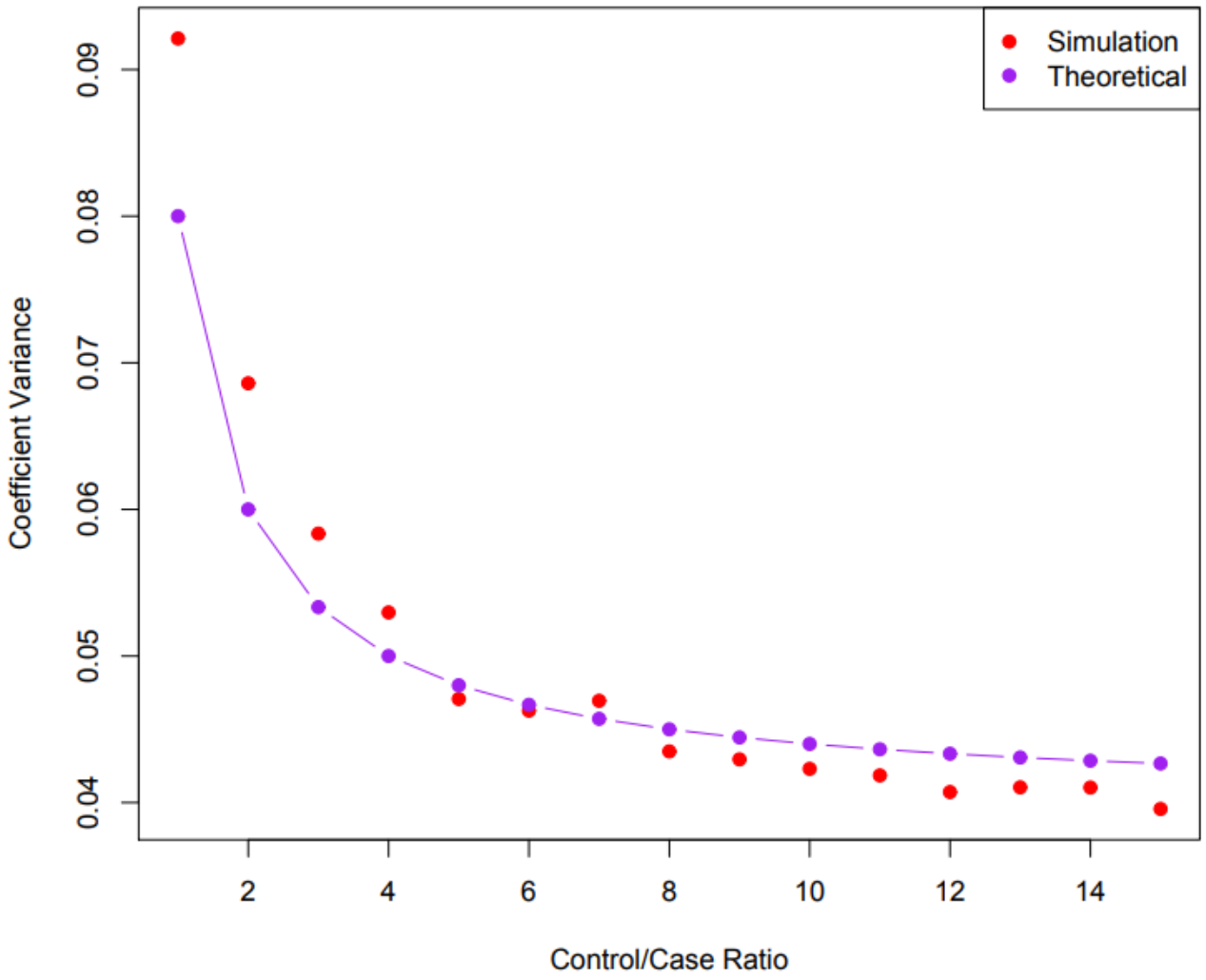
Case-control sampling and logistic regression

- In South African data, there are 160 cases, 302 controls — $\tilde{\pi}$ = 0.35 are cases. Yet the prevalence of MI in this region is π = 0.05.
- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect.
- We can correct the estimated intercept by a simple transformation

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\pi}$$

Often cases are rare and we take them all; up to five times that number of controls is sufficient.

Diminishing returns in unbalanced binary data



Sampling more controls than cases reduces the variance of the parameter estimates. But after a ratio of about 5 to 1 the variance reduction flattens out.

Linear Discriminant Analysis

- LDA undertakes the same task as Logistic Regression. It classifies data based on categorical variables
  - Making profit or not
  - Buy a product or not
  - Satisfied customer or not
  - Political party voting intention

Why not Logistic Regression??

There are several reasons:

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis 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, the linear discriminant model is again more stable than the logistic regression model.
- As mentioned earlier, linear discriminant analysis is popular when we have more than two response classes.

Why Linear? Why Discriminant?

- LDA involves the determination of linear equation (just like linear regression) that will predict which group the case belongs to.

$$D = v_1 X_1 + v_2 X_2 + \ldots + v_i X_i + a$$

- D: discriminant function
- v: discriminant coefficient or weight for the variable
- X: variable
- a: constant

Purpose of LDA

- Choose the v's in a way to maximize the distance between the means of different categories
- Good predictors tend to have large v's (weight)
- We want to discriminate between the different categories
- Think of food recipe. Changing the proportions (weights) of the ingredients will change the characteristics of the finished cakes. Hopefully that will produce different types of cake!

Assumptions of LDA

- The observations are a random sample

- Each predictor variable is normally distributed

Bayes' Classifier

- Bayes' classifier is the golden standard. Unfortunately, it is unattainable.

- So far, we have estimated it with two methods:
  - KNN classifier
  - Logistic Regression

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:

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

Estimating Bayes' Classifier

- Estimating Bayes' Classifier
- Bayes Theorem states $P(A|B)P(B) = P(B|A)P(A)$ or
- $\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)}$
- With Logistic Regression we modeled the probability of Y being from the k^{th} class as

$$p(X) = \Pr(Y = k|X = x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

- However, Bayes' Theorem states

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

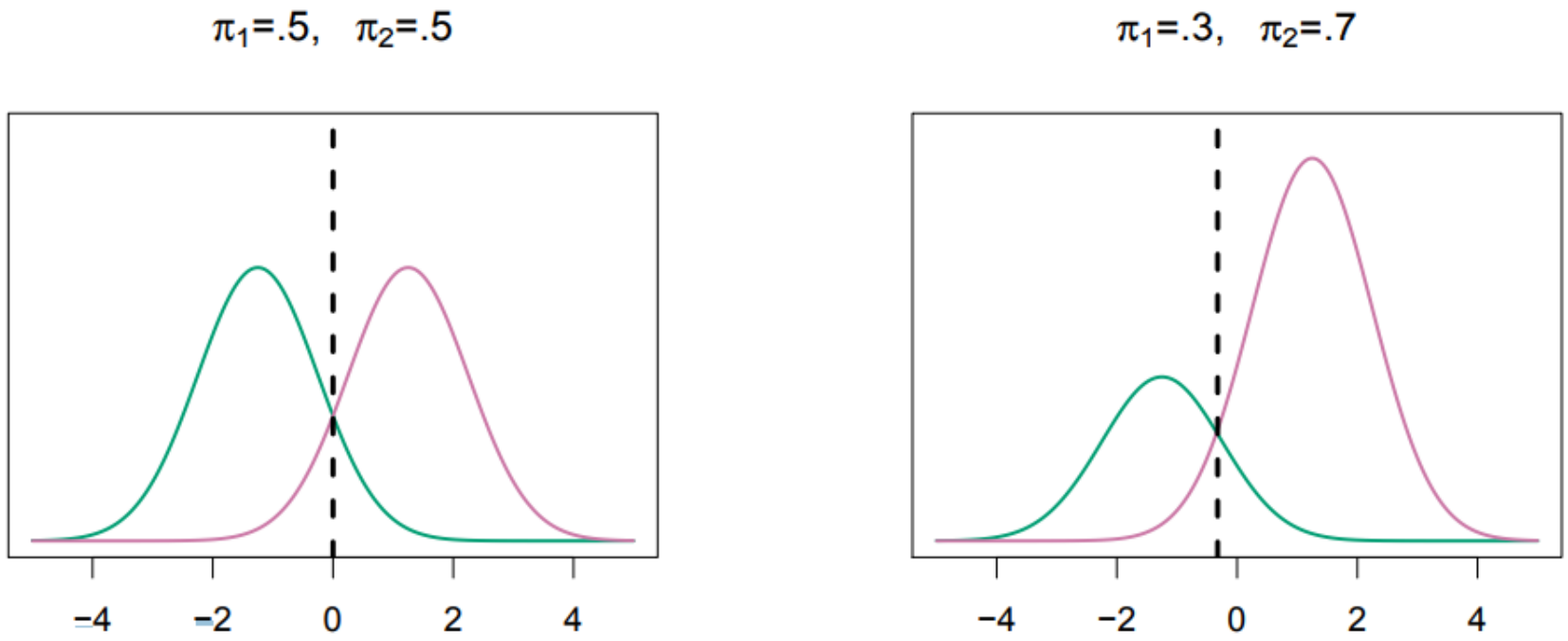
π_k : Probability of coming from class k (prior probability)

$f_k(X)$: Density function for X given that X is an observation from class k

Estimating Bayes' Classifier

- $f_k(x)$ = 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.
- π_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.

Estimate π_k and $f_k(x)$

- The most common model for $f_k(x)$ is the Normal Density

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

- where μ_k and σ_k^2 are the mean and variance parameters for the kth class.
- We will assume that all the $\sigma_k = \sigma$ are the same.

Linear Discriminant Analysis for $p = 1$

- For now, assume that $p = 1$—that is, we have only one predictor.
- let us further assume that $\sigma_1^2 = \ldots = \sigma_K^2$: that is, there is a shared
- variance term across all K classes, which for simplicity we can denote by σ^2
- $$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}$$
- Note that in (4.12), π_k denotes the prior probability that an observation belongs to the kth class, not to be confused with $\pi \approx 3.14159$, the mathematical constant.
- Happily, there are simplifications and cancellations.

Discriminant functions

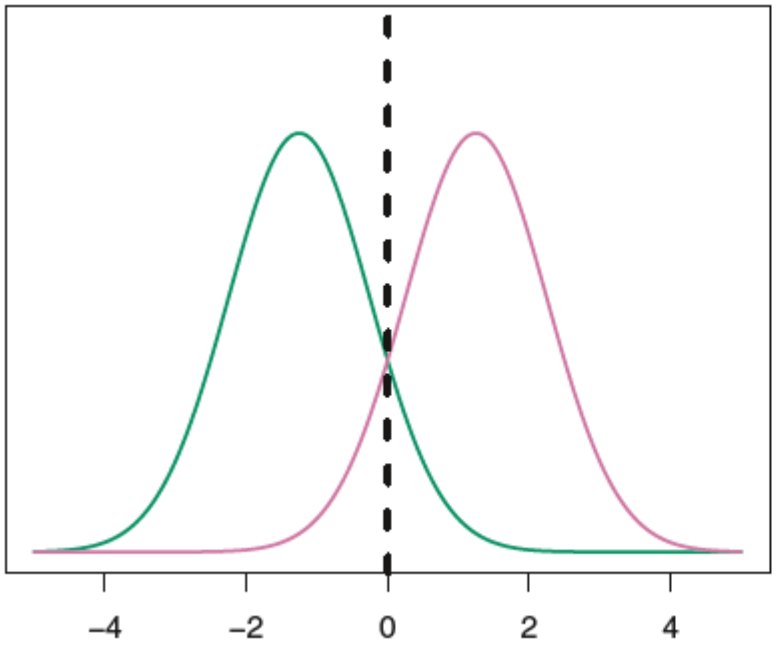
- The Bayes classifier involves assigning an observation $X = x$ to the class for which expression above is largest.
- Taking the log of (4.12) and rearranging the terms, it is not hard to show that this is equivalent to assigning the observation to the class for which

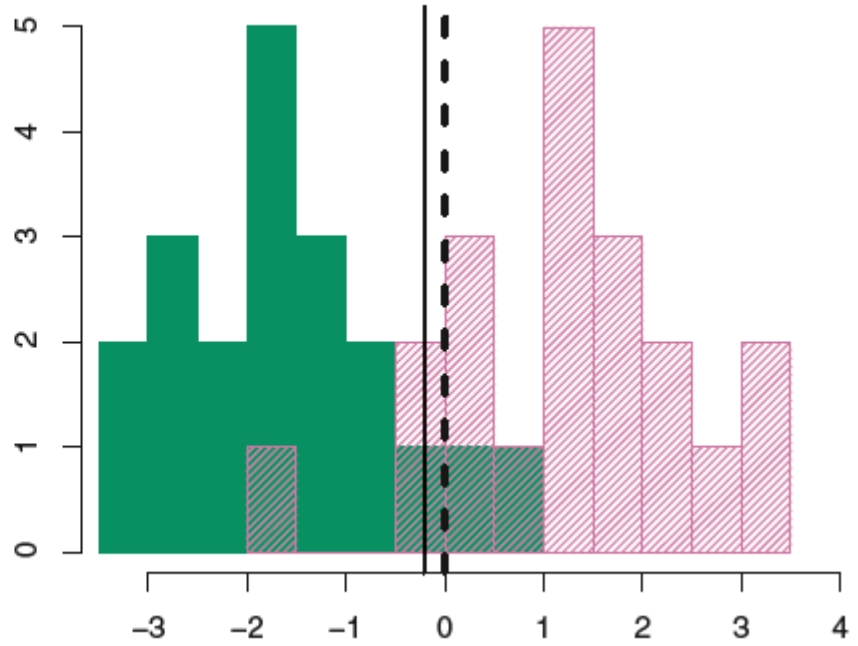
$$\delta_k(x) = x \cdot \frac{\mu^k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_x)$$

  is largest (*discriminant score*)
- Note that $\delta_k(x)$ is a linear function of x.
- For instance, if $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x\,(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise.

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

Example



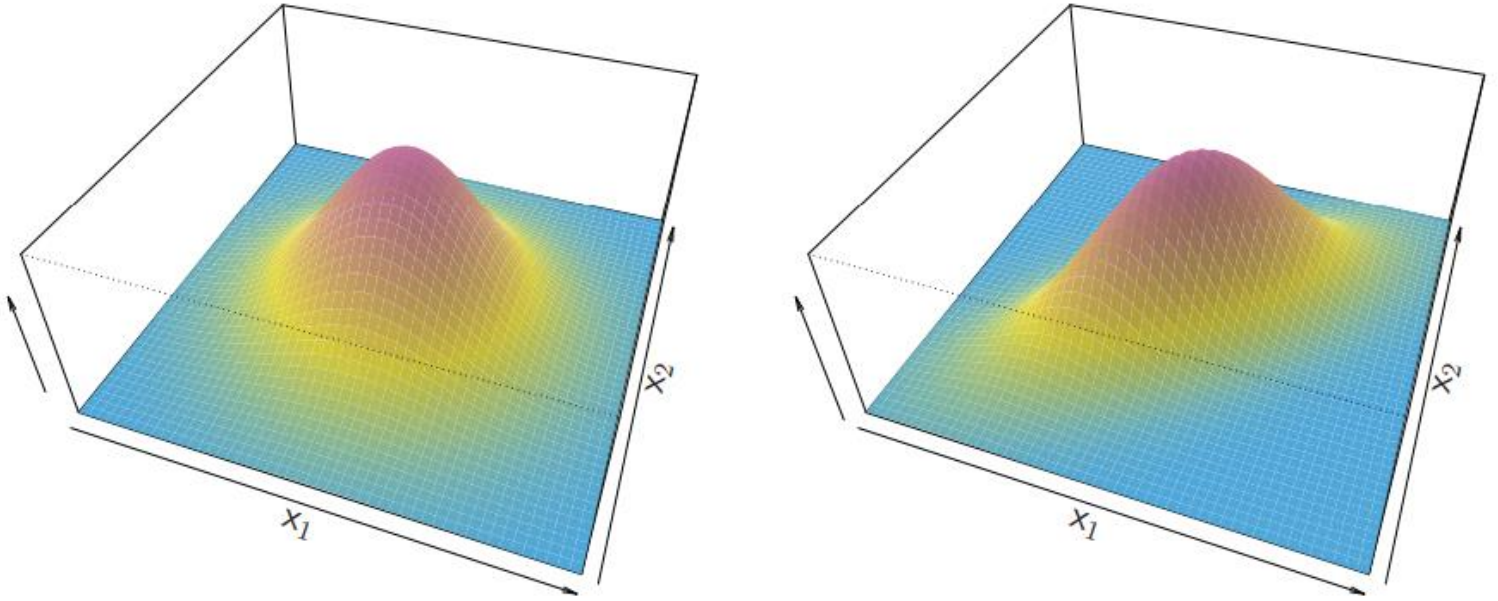


Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.*
Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

Estimating the parameters

- 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.
- $\hat{\pi}_k = \frac{n_k}{n}$
- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
- $\hat{\sigma}^2 = \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$
- $= \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$
- 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 *k*th class.

Linear Discriminant Analysis when $p > 1$



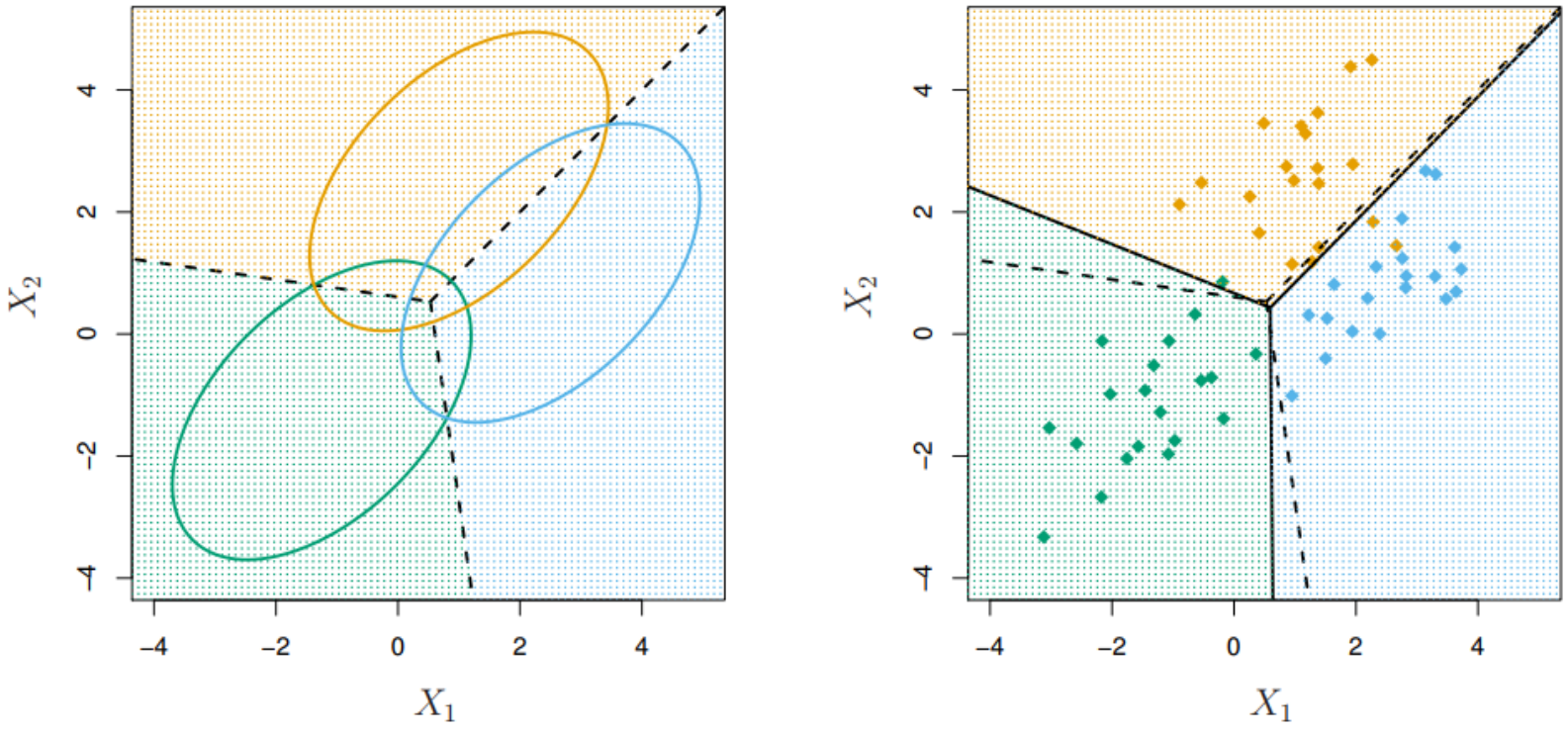
$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

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

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.

Illustration: $p = 2$ and $K = 3$ classes



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

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

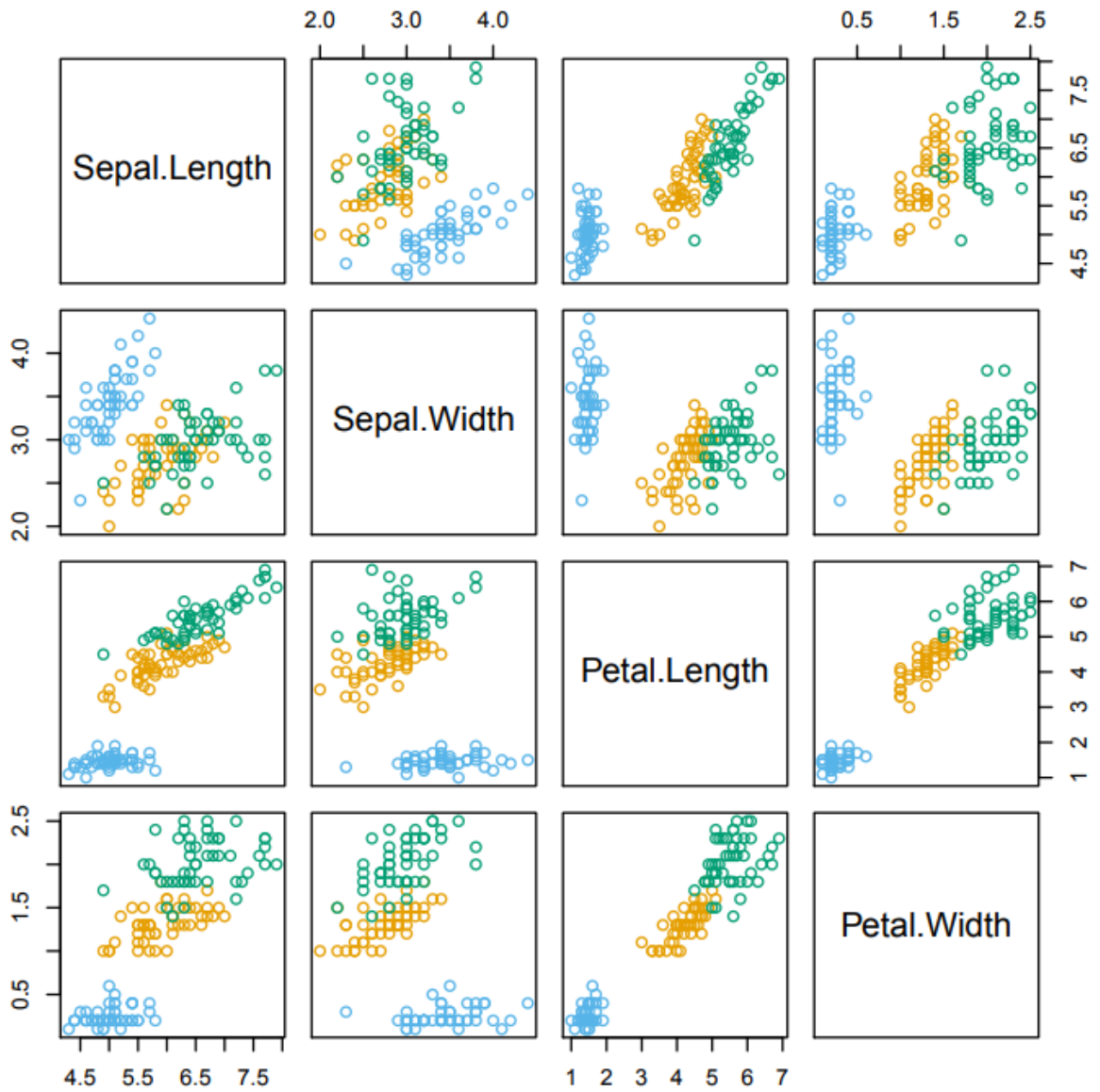
Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Fisher's Iris Data

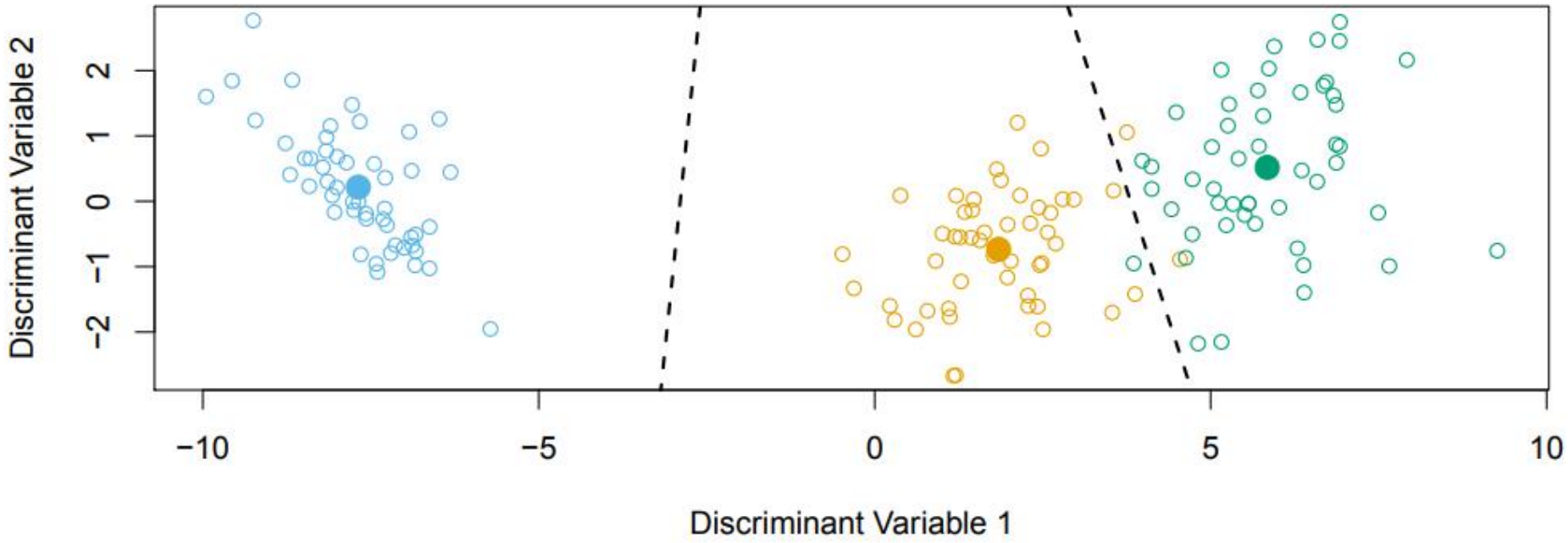
4 variables
3 species
50 samples/class

- Setosa
- Versicolor
- Virginica

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



Fisher's Discriminant Plot



When there are K classes, linear discriminant analysis can be viewed exactly in a $K-1$ dimensional plot.
Why? Because it essentially classifies to the closest centroid, and they span a $K-1$ dimensional plane.
Even when $K>3$, we can find the "best" 2-dimensional plane for vizualizing the discriminant rule.

From $\delta_k(x)$ to probabilities

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

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

- 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.

LDA on Credit Data

		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	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 = 4!
- 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!

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

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

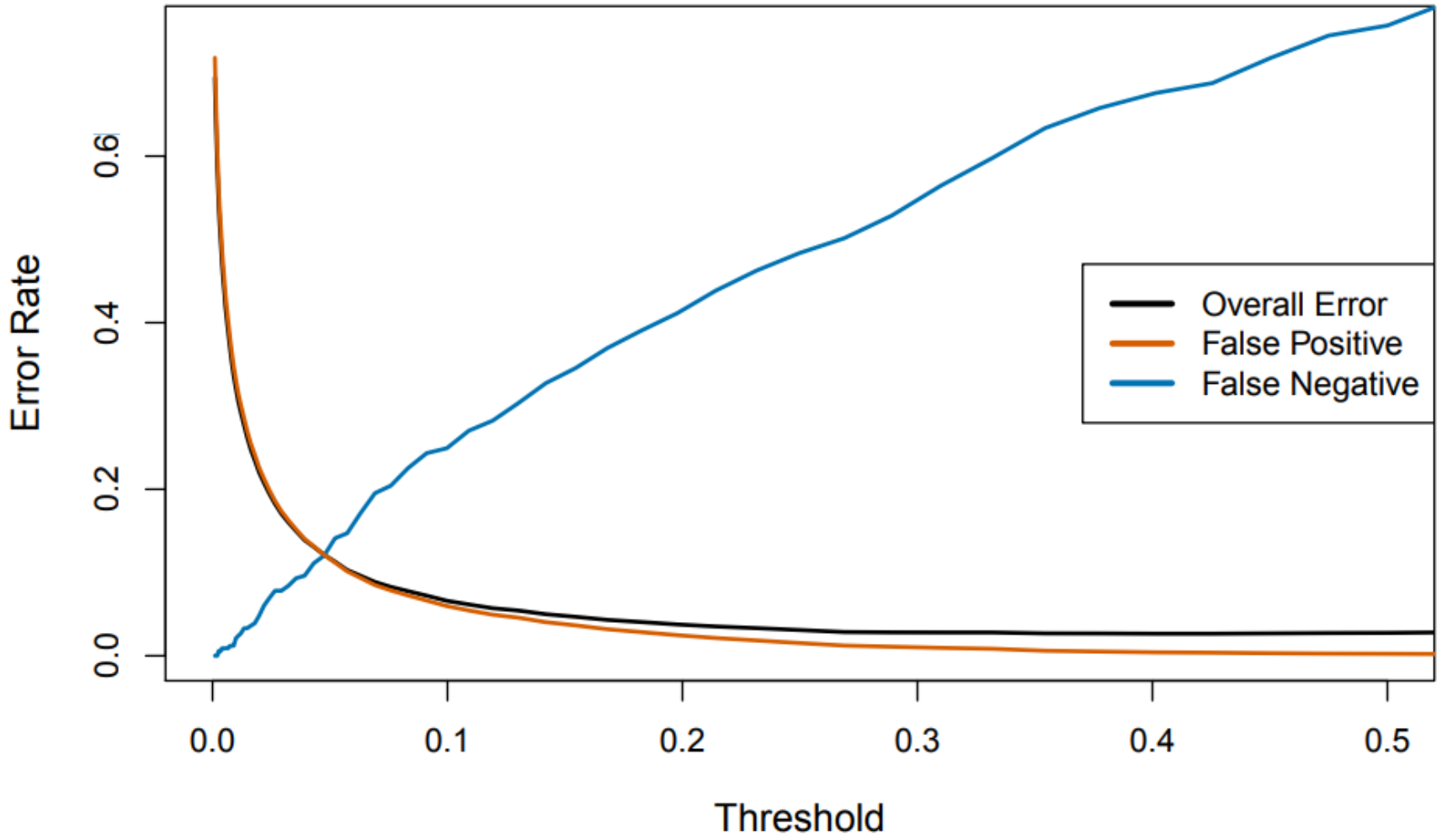
- We can change the two error rates by changing the threshold from 0.5 to some other value in [0, 1]:

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

and vary *threshold*

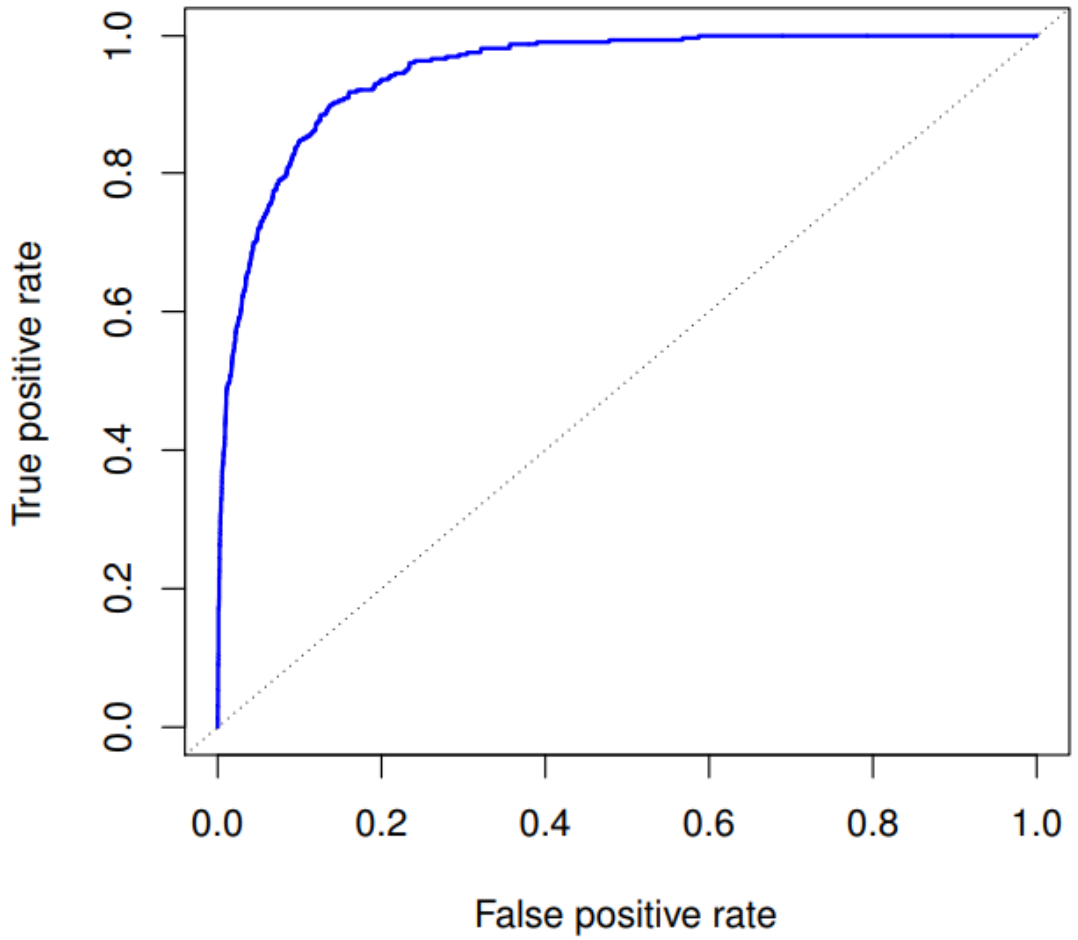
Varying the *threshold*

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



ROC Curve

- The *ROC plot* displays both simultaneously.
- Sometimes we use the *AUC* or *area under the curve* to summarize the overall performance. Higher *AUC* is good.
- As we have seen, varying the classifier threshold changes its true
- These are also called the *sensitivity* and one minus the *specificity* of our classifier.



Possible results when applying a classifier

		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*	

- The table shows the possible results when applying a classifier (or diagnostic test) to a population.
- To make the connection with the epidemiology literature, we think of "+" as the "disease" that we are trying to detect, and "–" as the "non-disease" state.
- To make the connection to the classical hypothesis testing literature, we think of "–" as the null hypothesis and "+" as the alternative (non-null) hypothesis.

Possible results when applying a classifier

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*	

- The denominators for the false positive and true positive rates are the actual population counts in each class.
- In contrast, the denominators for the positive predictive value and the negative predictive value are the total predicted counts for each class.

Quadratic Discriminant Analysis

- LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution
- A class specific mean vector and a covariance matrix is common to all *K* classes.
- *Quadratic discriminant analysis* (QDA) provides an alternative quadratic discriminant analysis approach.
- Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction.
- Unlike LDA, QDA assumes that each class has its own covariance matrix.

Quadratic Discriminant Analysis

- Assume that an observation from the *k*th class is of the form
- $X \sim N(\mu_k, \sum_k)$, where $\sum_k$ is a covariance matrix for the *k*th class.
- Under this assumption, the Bayes classifier assigns an observation $X = x$ to the class for which

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

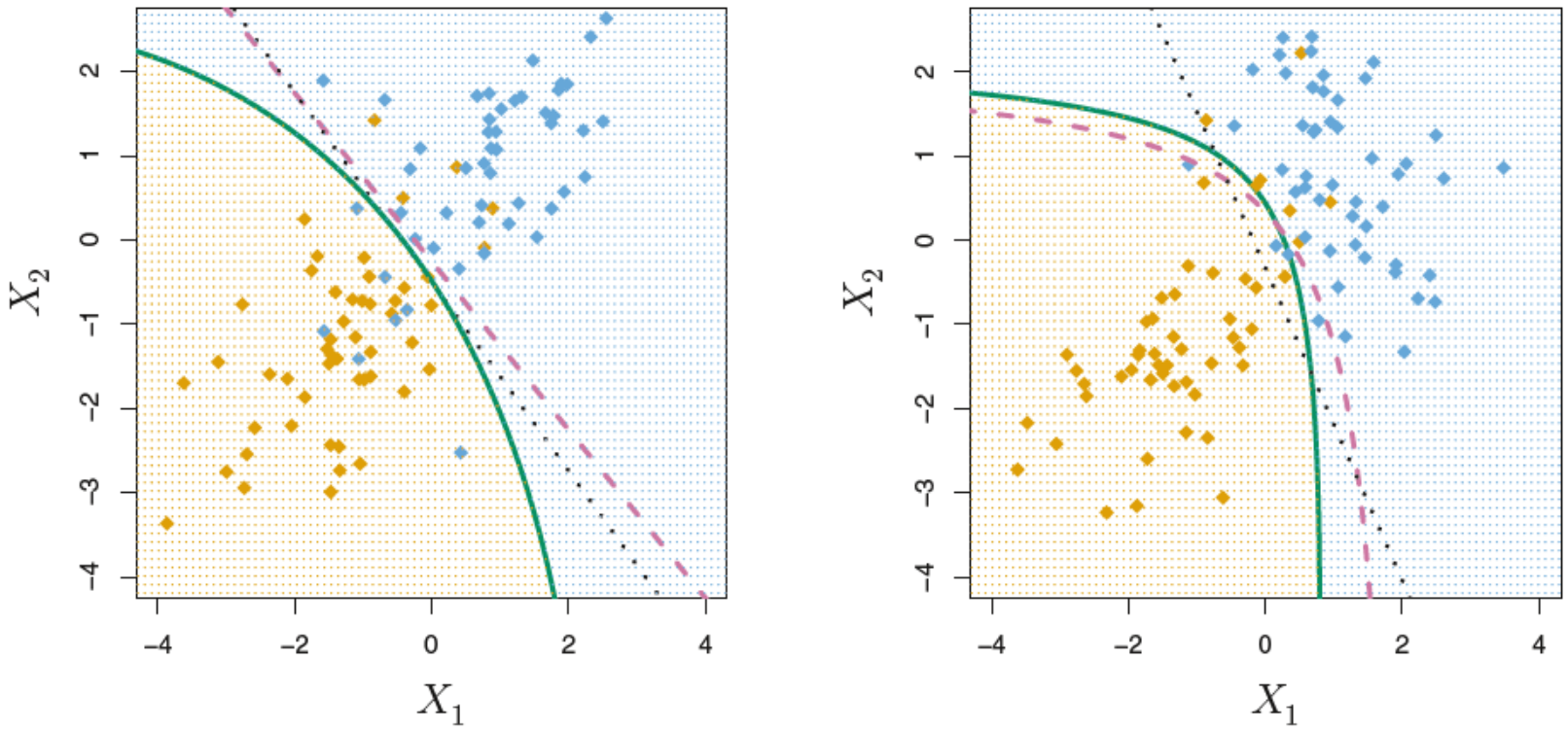
  is largest.

- So the QDA classifier involves plugging estimates for $\sum_k$, μ_k, and π_k into the above expression and then assigning an observation $X = x$ to the class for which this quantity is largest.

Quadratic Discriminant Analysis

- Unlike LDA, the quantity x appears as a *quadratic* function.
- Why would one prefer LDA to QDA, or vice-versa?
- The answer lies in the bias-variance trade-off.
- When there are p predictors, then estimating a covariance matrix requires estimating $p(p+1)/2$ parameters.
- QDA estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters
- With 50 predictors this is some multiple of 1,275.
- If the K classes share a common covariance matrix, the LDA model becomes linear in x, meaning Kp linear coefficients to estimate.
- LDA is a much less flexible classifier than QDA, and so has substantially lower variance.

LDA vs. QDA



Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *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* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

A Comparison of Classification Methods

We have considered three different classification approaches: KNN, logistic regression, LDA, and QDA.

- Logistic regression and LDA methods are closely connected.
- Consider the two-class setting with $p = 1$ predictor, and let $p1(x)$ and $p2(x) = 1-p1(x)$ be the probabilities that the observation $X = x$ belongs to class 1 and class 2, respectively.

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x,$$

- we know that in logistic regression,

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x.$$

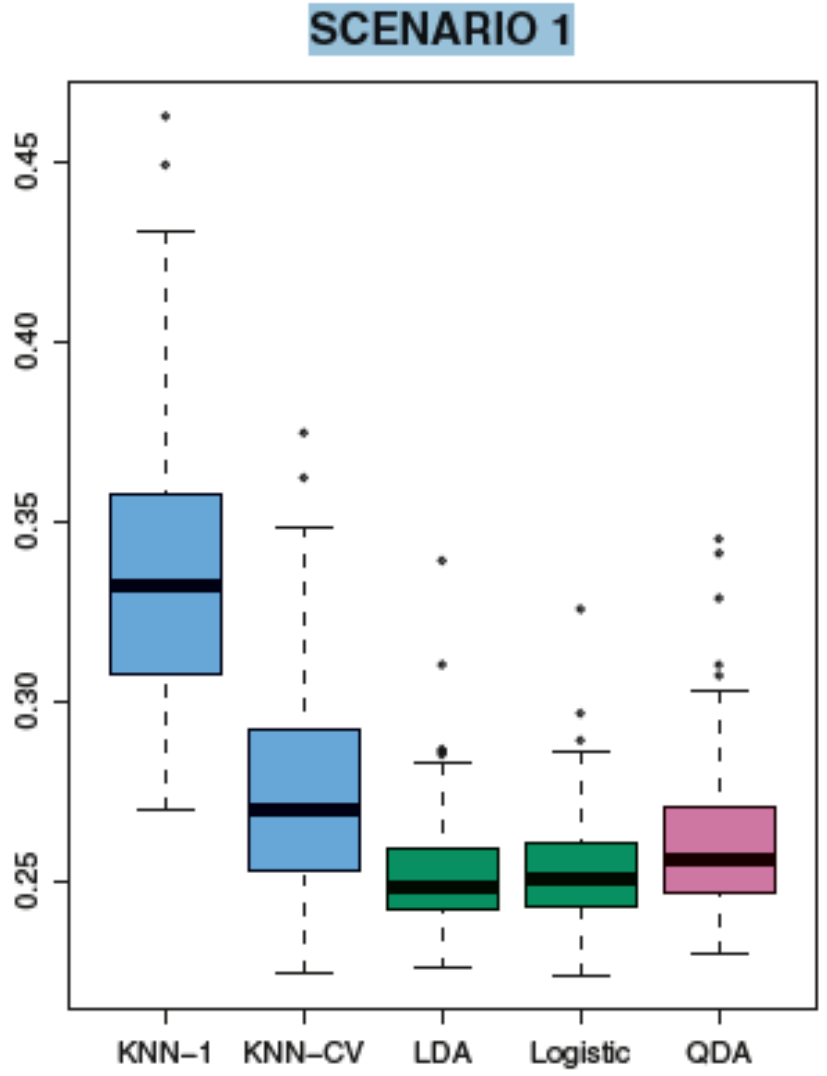
A Comparison of Classification Methods

- Both logistic regression and LDA produce linear decision boundaries.
- The only difference between the two approaches lies in the fact that β_0 and β_1 are estimated using maximum likelihood, whereas c_0 and c_1 are computed using the estimated mean and variance from a normal distribution.
- Recall from Chapter 2 that KNN takes a completely different approach from the classifiers seen in this chapter.
- In order to make a prediction for an observation $X = x$, the K training observations that are closest to x are identified. Then X is assigned to the class to which the plurality of these observations belong.
- Hence KNN is a completely non-parametric approach: no assumptions are made about the shape of the decision boundary. Therefore, we can expect this approach to dominate LDA and logistic regression when the decision boundary is highly non-linear.

SCENARIO 1

Scenario 1: There were 20 training observations in each of two classes.

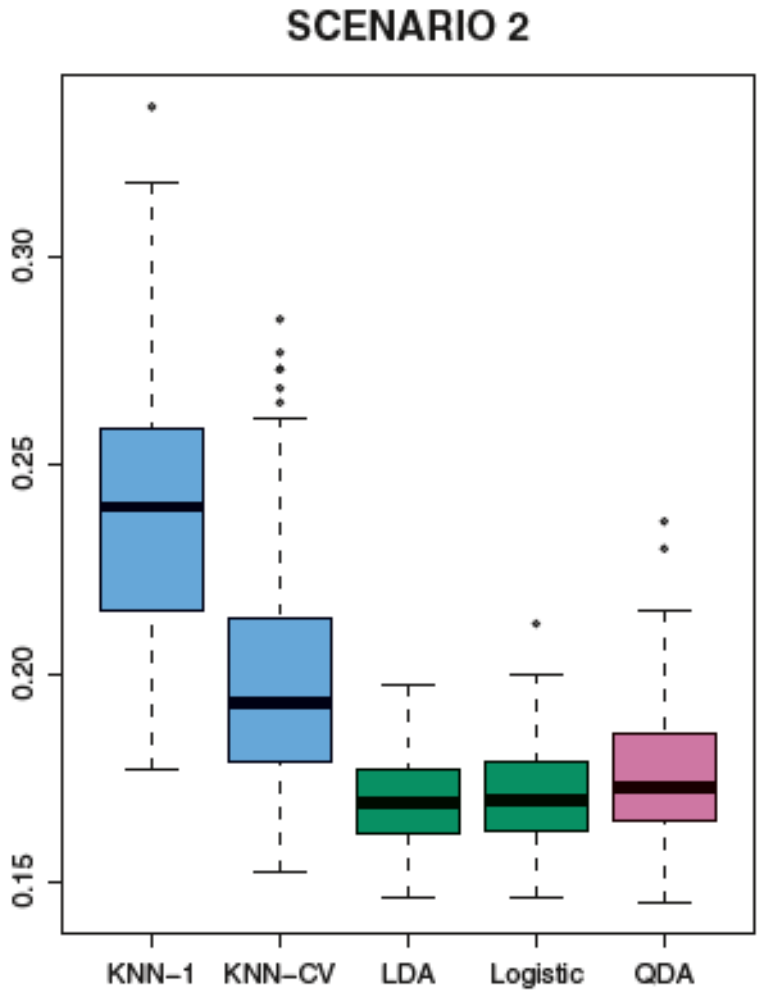
- The observations within each class were uncorrelated random normal variables with a different mean in each class. The Figure shows that LDA performed well in this setting, as one would expect since this is the model assumed by LDA.
- KNN performed poorly because it paid a price in terms of variance that was not offset by a reduction in bias.
- QDA also performed worse than LDA, since it fit a more flexible classifier than necessary.
- Since logistic regression assumes a linear decision boundary, its results were only slightly inferior to those of LDA.



Scenario 2:

Scenario 2: Details are as in Scenario 1, except that within each class, the two predictors had a correlation of −0.5.

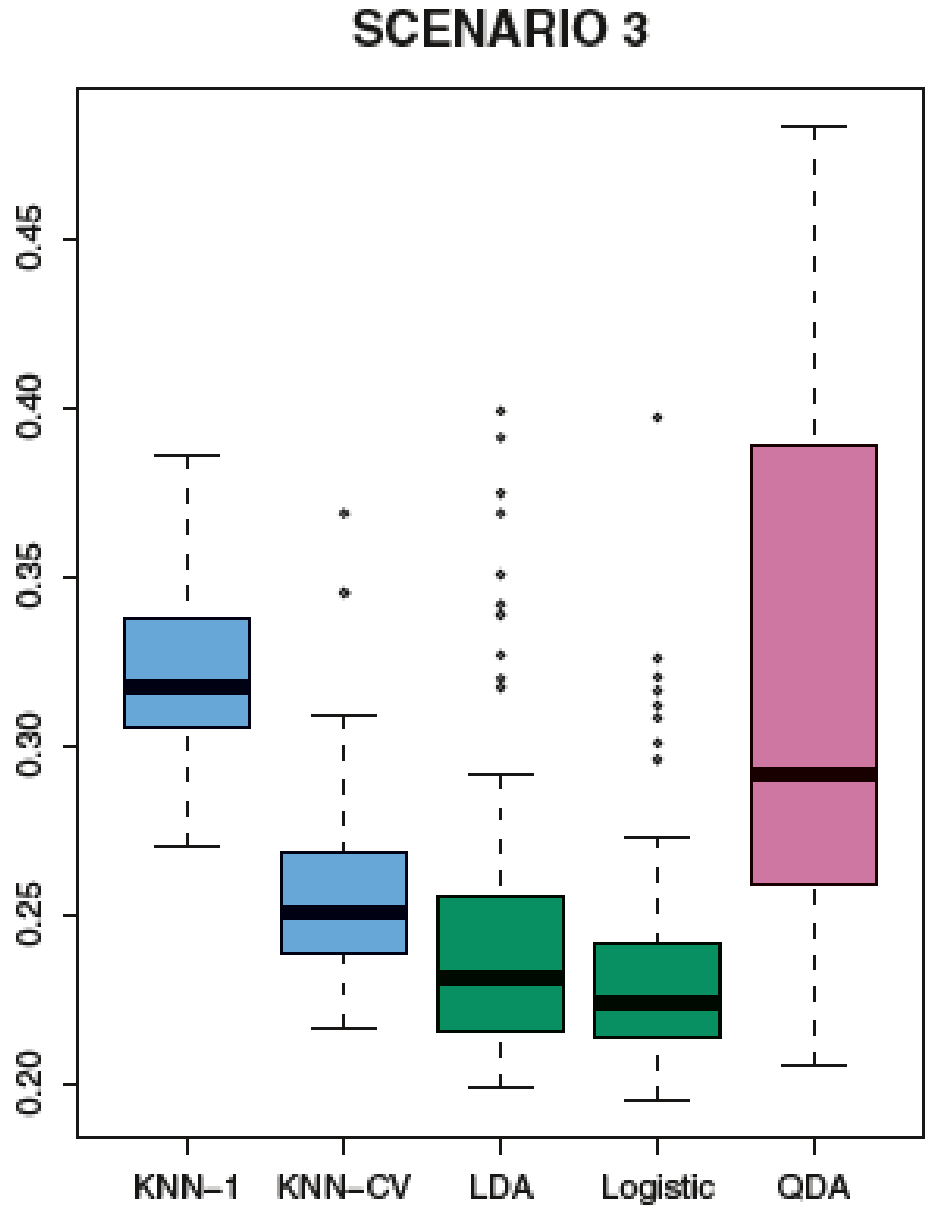
- The Figure indicates little change in the relative performances of the methods as compared to the previous scenario.



Scenario 3:

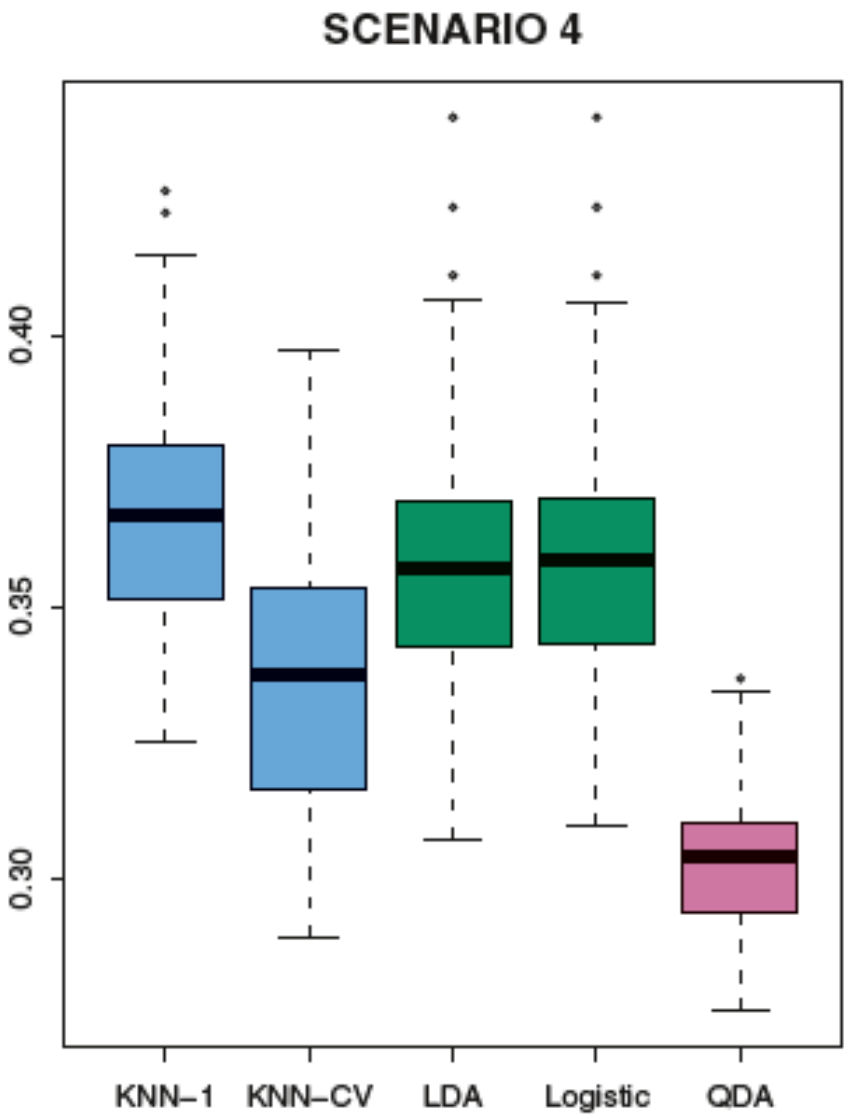
Scenario 3: We generated X_1 and X_2 from the *t-distribution*, with 50 observations per class.

- The *t*-distribution has a similar shape to the normal distribution, but it has a tendency to yield more extreme points—that is, more points that are far from the mean.
- In this setting, the decision boundary was still linear, and so fit into the logistic regression framework. The set-up violated the assumptions of LDA, since the observations were not drawn from a normal distribution.
- The Figure shows that logistic regression outperformed LDA, though both methods were superior to the other approaches. In particular, the QDA results deteriorated considerably as a consequence of non-normality.



Scenario 4:

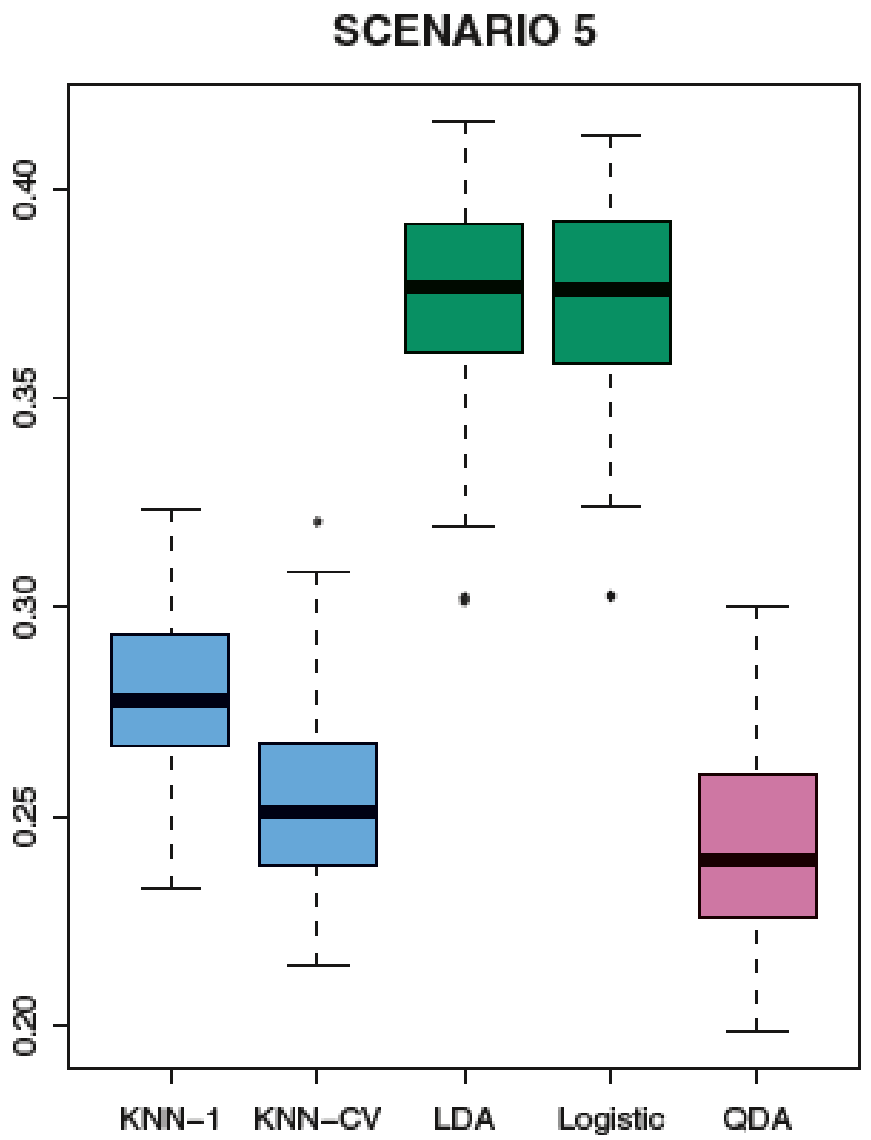
- The data were generated from a normal distribution, with a correlation of 0.5 between the predictors in the first class, and correlation of −0.5 between the predictors in the second class.
- This setup corresponded to the QDA assumption, and resulted in quadratic decision boundaries.
- The Figure shows that QDA outperformed all of the other approaches.



Scenario 5

Scenario 5: Within each class, the observations were generated from a normal distribution with uncorrelated predictors. However, the responses were sampled from the logistic function using X_{21}, X_{22}, and $X_1 \times X_2$ as predictors.

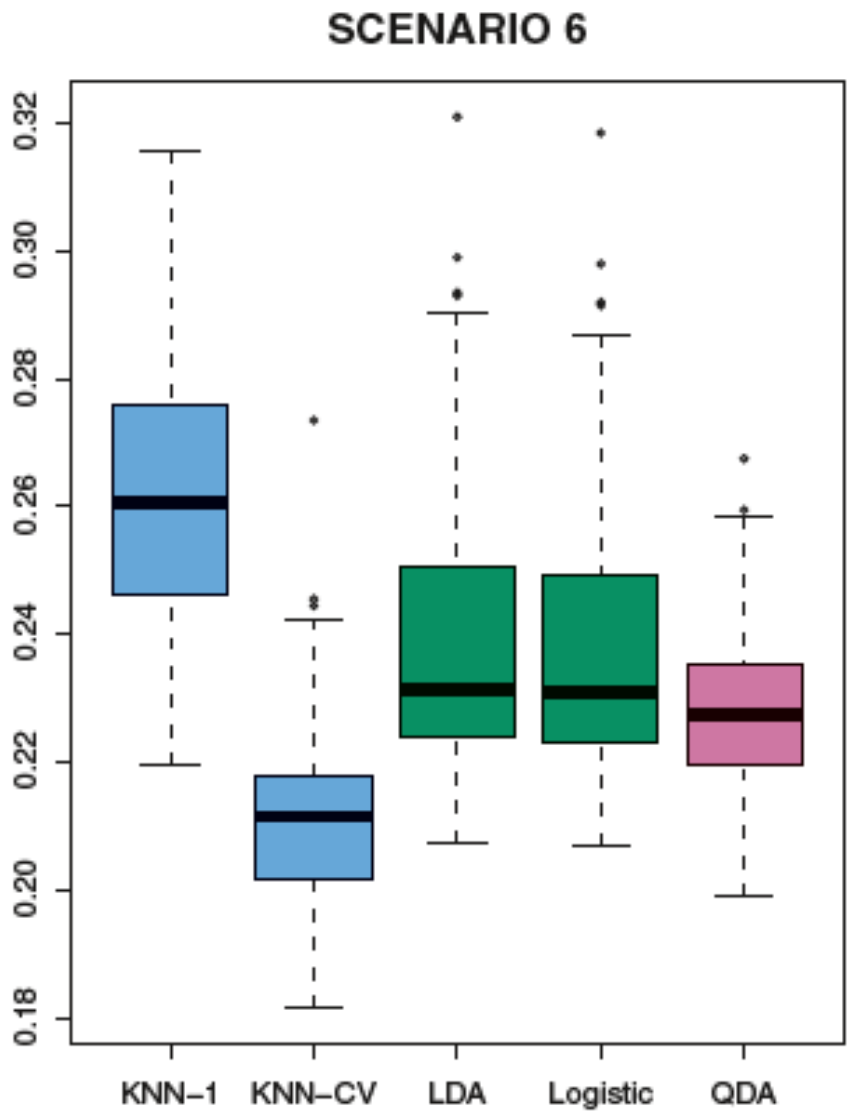
- Consequently, there is a quadratic decision boundary.
- The Figure indicates that QDA once again performed best, followed closely by KNN-CV.
- The linear methods had poor performance.



Scenario 6:

Scenario 6: Details are as in the previous scenario, but the responses were sampled from a more complicated non-linear function.

- As a result, even the quadratic decision boundaries of QDA could not adequately model the data.
- The Figure shows that QDA gave slightly better results than the linear methods, while the much more flexible KNN-CV method gave the best results.
- But KNN with $K = 1$ gave the worst results out of all methods.
- This highlights the fact that even when the data exhibits a complex nonlinear relationship, a non-parametric method such as KNN can still give poor results if the level of smoothness is not chosen correctly.



Conclusion

- These six examples illustrate that no one method will dominate the others in every situation.
- When the true decision boundaries are linear, then the LDA and logistic regression approaches will tend to perform well.
- When the boundaries are moderately non-linear, QDA may give better results.
- Finally, for much more complicated decision boundaries, a non-parametric approach such as KNN can be superior. But the level of smoothness for anon-parametric approach must be chosen carefully.