Classification

Some of the figures are taken from “An Introduction to Statistical Learning with Applications in R” with permission from the authors G. James, D. Witten, T. Hastie, and R. Tibshirani

Introduction

- The linear regression model assumes that the response variable Y is quantitative.
- In many situations, the response variable is qualitative (*categorical*).
  - For example, eye color is qualitative, taking on values blue, brown, or green.
- Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

Classification Techniques

- There are many possible classification techniques, or classifiers.
- In this lecture we will discuss three of the most widely-used classifiers:
  - Logistic regression
  - Linear discriminant analysis (LDA)
  - K-nearest neighbors.
- We will discuss other classifiers such as generalized additive models, trees, random forests, and support vector machines later in this course.

Data Description

- We will use Default dataset, which is available in ISLR library, for the illustration purposes in this lecture.

```
names(Default)
```

```
'default'  'student'  'balance'  'income'
```

```
dim(Default)
```

```
10000  4
```

```
head(Default)
```

default	student	balance	income
No	No	729.5265	44361.625
No	Yes	817.1804	12106.135
No	No	1073.5492	31767.139
No	No	529.2506	35704.494
No	No	785.6559	38463.496
No	Yes	919.5885	7491.559

```
summary(Default)
```

```
 default    student       balance           income
 No :9667   No :7056   Min.   :   0.0   Min.   :  772
 Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
                       Median : 823.6   Median :34553
                       Mean   : 835.4   Mean   :33517
                       3rd Qu.:1166.3   3rd Qu.:43808
                       Max.   :2654.3   Max.   :73554
```

Overview of Classification

- Qualitative variables take values in an unordered set C, such as:

  $eye\ color \in \{brown, blue, green\}$

  $email \in \{spam, not\ spam\}$

- Given a vector X and a qualitative response Y taking values in the set C, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y ; i.e. $C(X) \in C$.

- Often we are more interested in estimating the probabilities that X belongs to each category in C.

  - For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

An Example

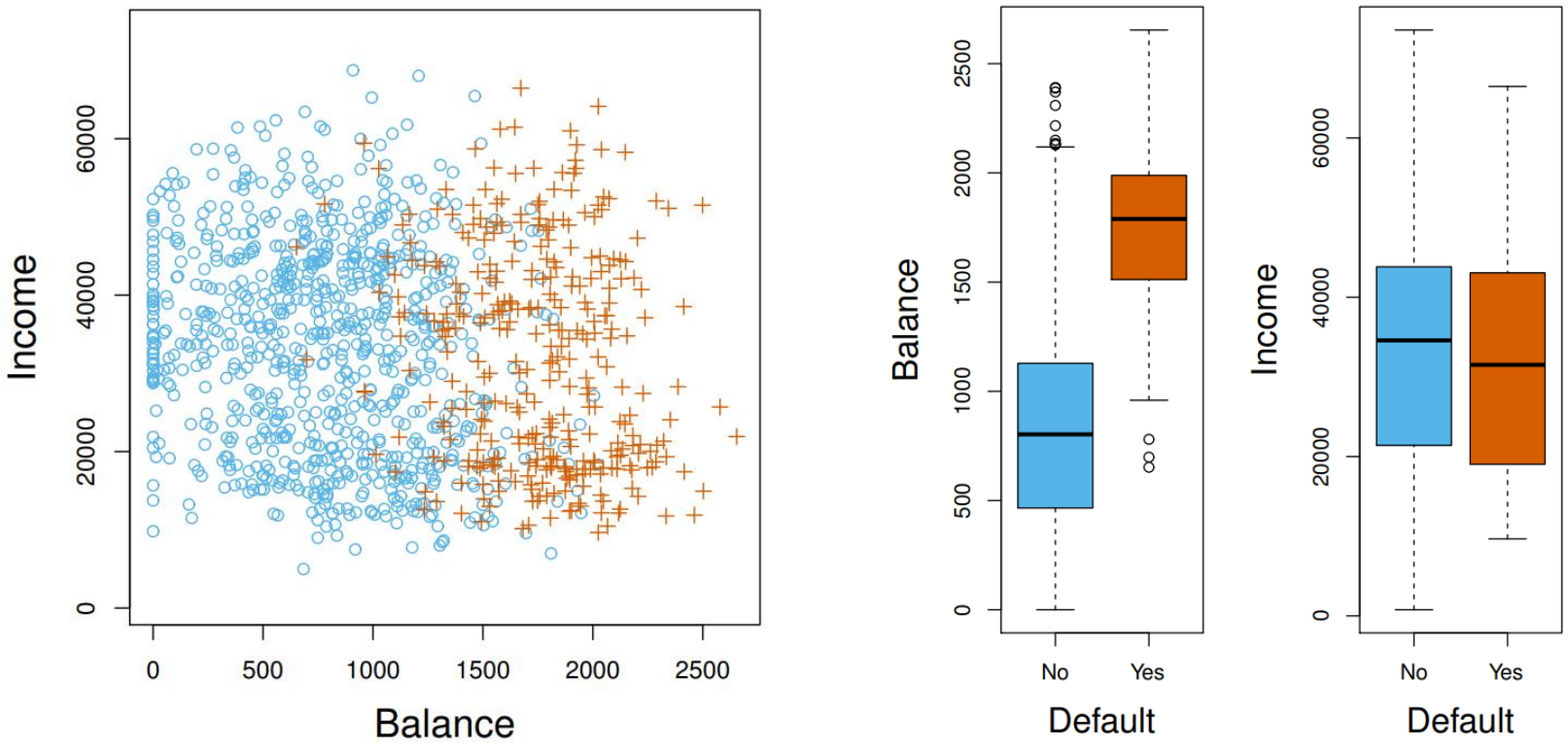


Figure. The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

Why not linear regression?

- Suppose for the Default classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classier, and is equivalent to linear discriminant analysis which we discuss later.
- Since in the population $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.

Why not linear regression?

Linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

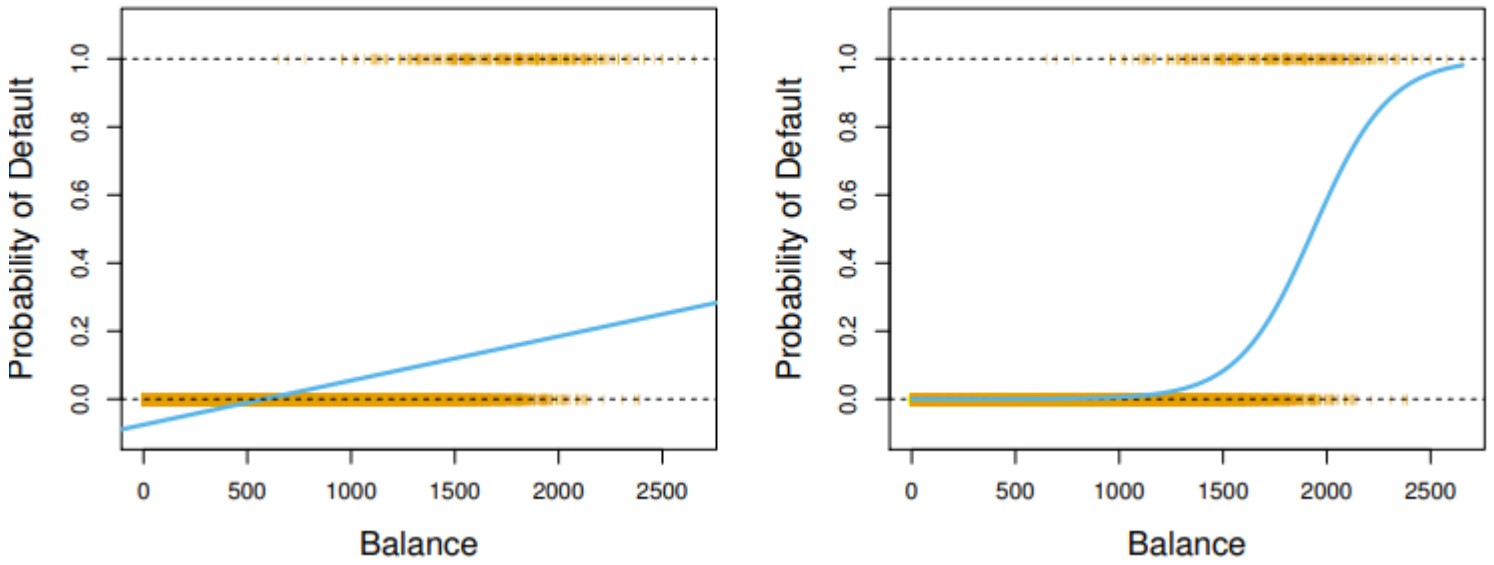


Figure. Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default(No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Logistic Regression

An Example

- Consider the Default data set, where the response default falls into one of two categories, Yes or No.
  - Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category.
  - The probability of default given balance is

$$\Pr(\text{default} = \text{Yes} \mid \text{balance})$$

  - Then for any given value of balance, a prediction can be made for default.
  - For example, one might predict default = Yes for any individual for whom $\Pr(\text{default} = \text{Yes} \mid \text{balance})$, $p(\text{balance})$ for short, is greater than 0.5.

How should we model the relationship?

- For convenience, we are using the generic 0/1 coding for the response (0: No, 1:Yes).
- Let's write $p(X) = \Pr(Y = 1 \mid X)$ for short and consider using balance to predict default. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])

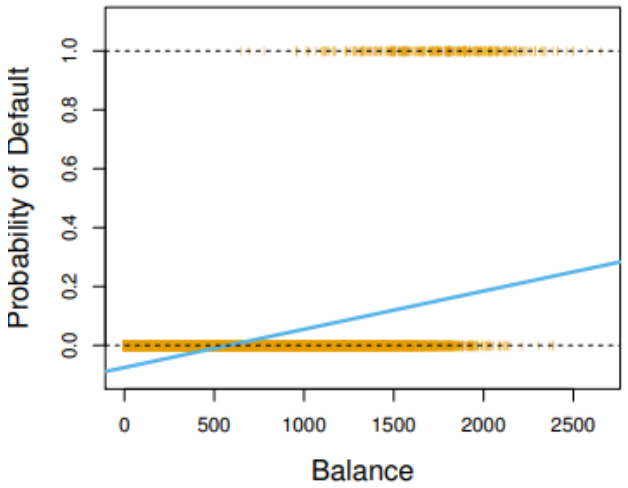
- It is easy to see that no matter what values β_0, β_1 or X take, p(X) will have values between 0 and 1.

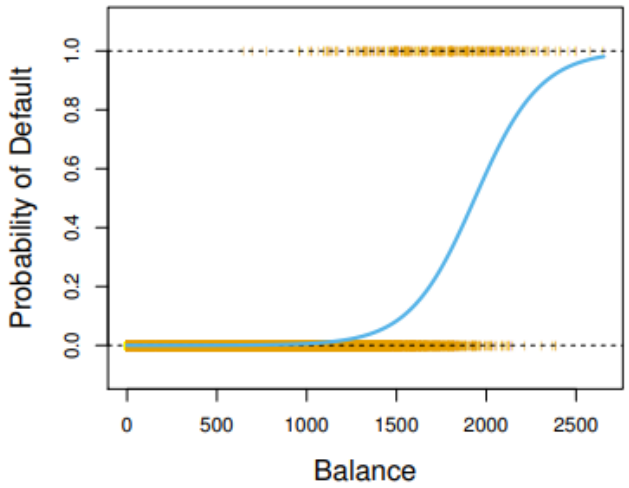
The Logistic Model

- A bit of rearrangement gives

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

- This monotone transformation is called the log odds or logit transformation of $p(X)$.





Logistic regression ensures that our estimate for $p(X)$ lies between 0 and 1.

Interpretations of coefficients

- Because the relationship between $p(X)$ and X is not a straight line, β_1 does not correspond to the change in $p(X)$ associated with a one-unit increase in X.
- The rate of change in $p(X)$ per unit change in X depends on the current value of X.
- But regardless of the value of X, if β_1 is positive then increasing X will be associated with increasing $p(X)$, and if β_1 is negative then increasing X will be associated with decreasing $p(X)$.

Estimating the coefficients

- We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

- Most statistical packages can fit logistic regression models by maximum likelihood. In R we use the glm function.

An Example

	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

- $\beta_1 = 0.0055$; this indicates that an increase in balance is associated with an increase in the probability of default.
  - A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.
- The estimated intercept is typically not of interest; its main purpose is to adjust the average fitted probabilities to the proportion of ones in the data

An Example

	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

- Many aspects of the logistic regression output are similar to the linear regression output.
- The *z*-statistic plays the same role as the *t*-statistic in the linear regression output
- The *z*-statistic associated with β_1 is equal to $\widehat{\beta_1}/SE(\widehat{\beta_1})$, and so a large (absolute) value of the z-statistic indicates evidence against the null hypothesis $H_0: \beta_1 = 0$ (implies that the probability of default does not depend on balance).

An Example

	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

- Since the p-value associated with balance is tiny, we can reject

$H_0: \beta_1 = 0$

Making predictions

- What is our estimated probability of default for someone with a balance of $1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

- With a balance of $2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Categorical predictors

- One can use categorical predictors with logistic regression model using the dummy variable approach.
- Example for Default data set: we have a qualitative variable student.
  - To fit the model we simply create a dummy variable that takes on a value of 1 for the students and 0 for non-students.

Categorical predictors

	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

- The coefficient associated with the dummy variable is positive, and the associated p-value is statistically significant.
- This indicates that students tend to have higher default probabilities than non-students:

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

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

Logistic regression with several variables

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

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

	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

- Why is coefficient for student negative, while it was positive before?

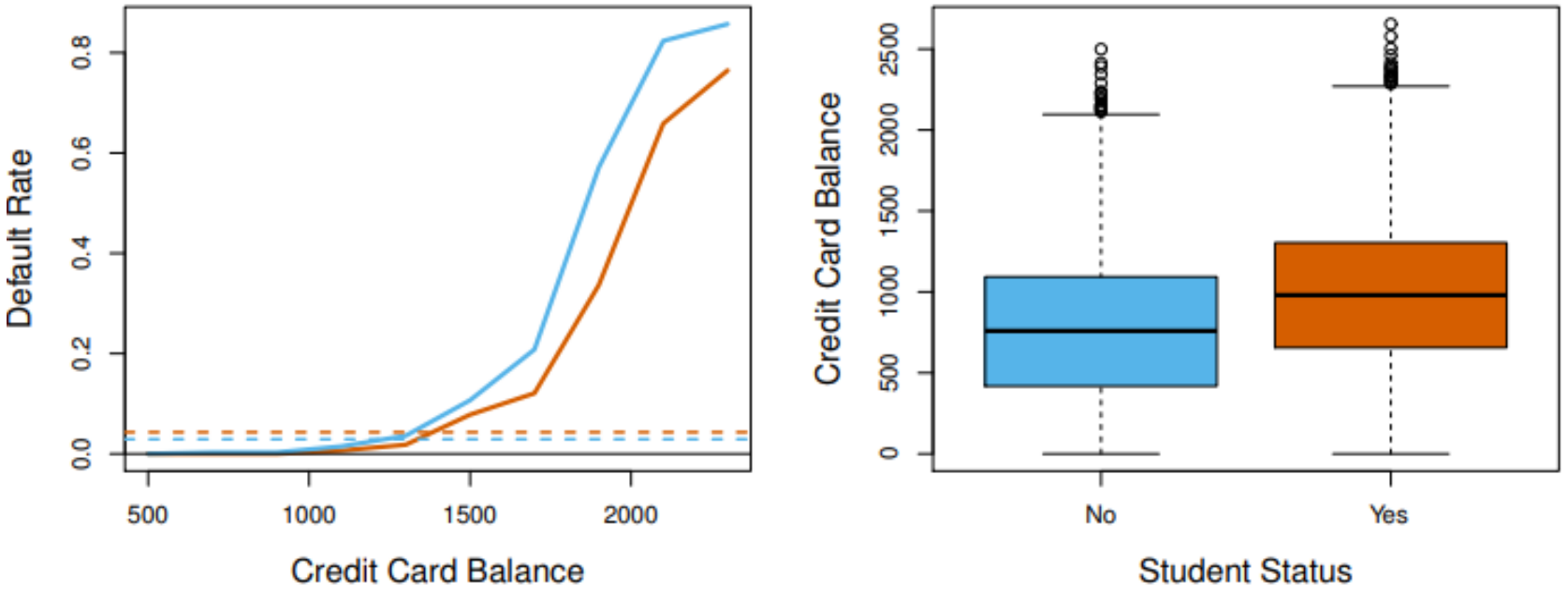
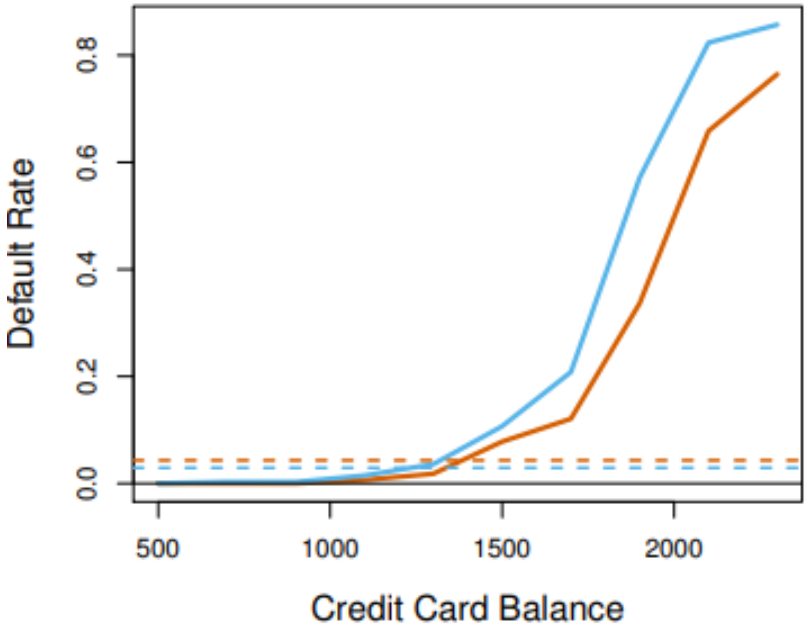
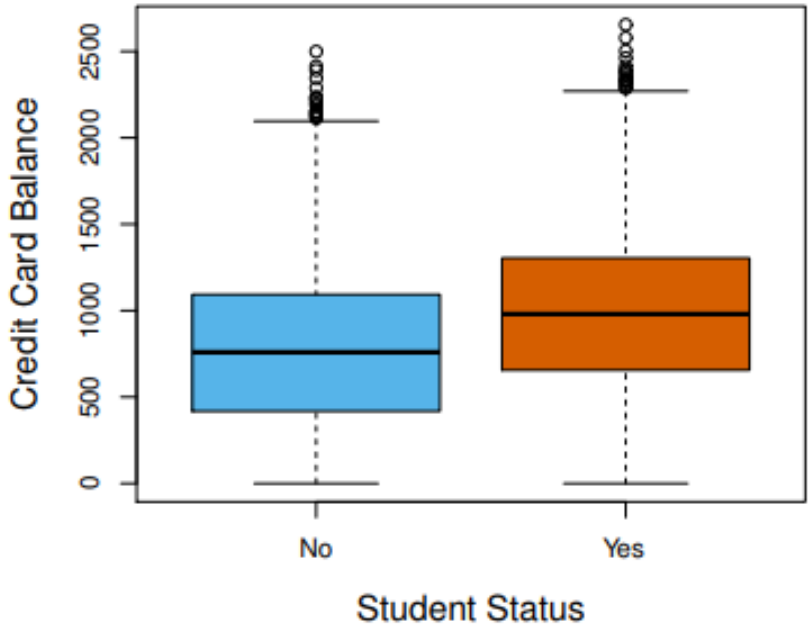


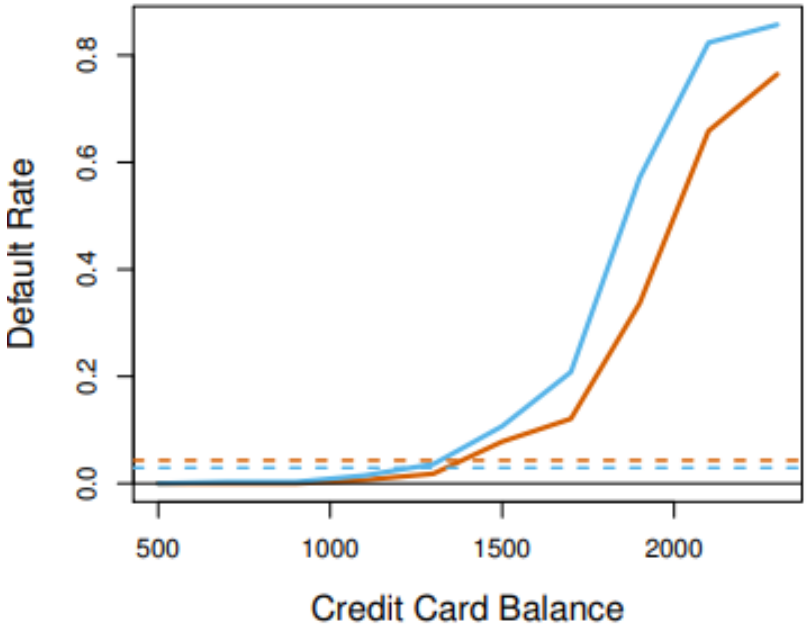
Figure. Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

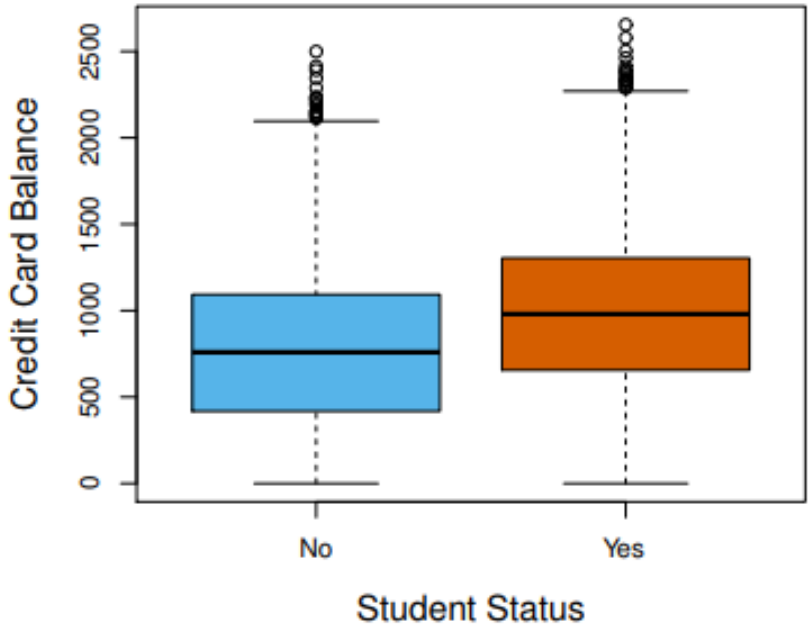
- Students tend to hold higher levels of debt, which is in turn associated with higher probability of default.
  - In other words, students are more likely to have large credit card balances, which, as we know from the left-hand panel of Figure, tend to be associated with high default rates.





- Even though an individual student with a given credit card balance will tend to have a lower probability of default than a non-student with the same credit card balance, the fact that students on the whole tend to have higher credit card balances means that overall, students tend to default at a higher rate than non-students.





- This example illustrates the dangers and subtleties associated with performing regressions involving only a single predictor when other predictors may also be relevant.
- This phenomenon is known as confounding.

Logistic regression with more than two classes

- So far we have discussed logistic regression with two classes.
- It is easily generalized to more than two classes. One version (used in the R package glmnet) has the symmetric form

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

- Here there is a linear function for each class. (The *mathier* students will recognize that some cancellation is possible, and only $K - 1$ linear functions are needed as in 2-class logistic regression.)
- Multiclass logistic regression is also referred to as multinomial regression.

Linear Discriminant Analysis

Discriminant Analysis

- Here 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 $\Pr(Y \mid X)$.
- When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.
- However, this approach is quite general, and 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 probabilistic modeling. Here we focus on a simple result, known as Bayes theorem:

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

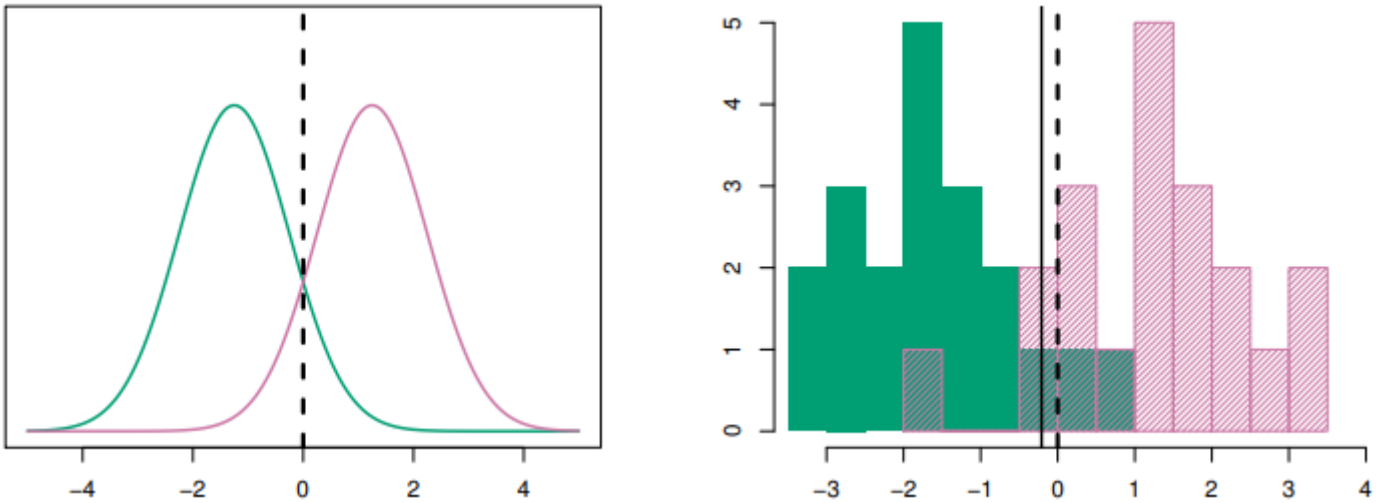
Bayes theorem for classification

- For discriminant analysis

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

  - $p_k(X) = \Pr(Y = k|X = x)$ as the posterior probability that an observation $X = x$ belongs to the kth class. That is, it is the probability that the observation belongs to the kth class, given the predictor value for that observation.
  - $f_k(x) = \Pr(X = x \mid Y = k)$ denote the density function of X for an observation that comes from the kth class.. 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.

Overall idea: 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.

Why discriminant analysis?

- 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.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

Linear discriminant analysis ($p = 1$)

- Suppose we assume that $f_k(x)$ is normal or Gaussian, which takes 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 is 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 $p_k(x) = \Pr(Y = k \mid X = x)$:

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

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

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

An Example

- Example with $\mu_1 = -1.25, \mu_2 = 1.25, \sigma_1^2 = \sigma_2^2 = 1$, and $\pi_1 = \pi_2 = 0.5$.

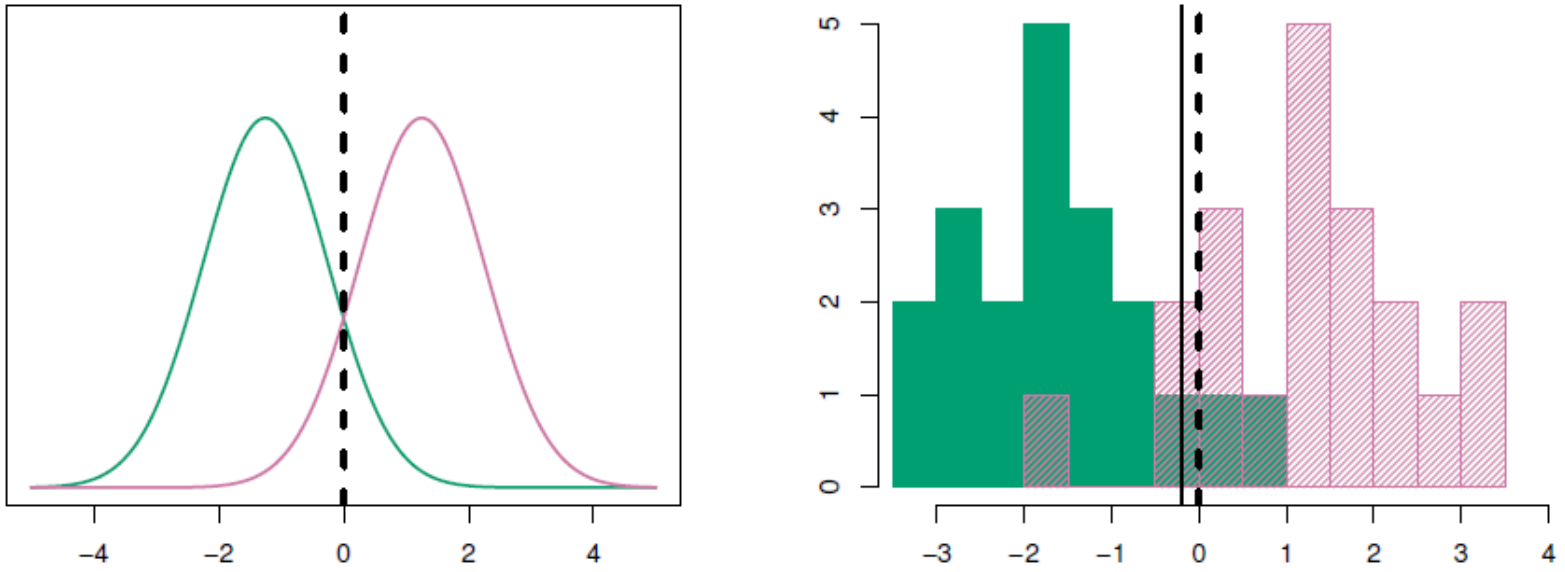


Figure. Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The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

- Note that in this case, we can compute the Bayes classifier because we know that X is drawn from a Gaussian distribution within each class, and we know all of the parameters involved.
- In a real-life situation, we are not able to calculate the Bayes classifier. Typically, we don't know the parameters; we just have the training data.
- In that case we simply estimate the parameters and plug them into the rule.

Estimating the parameters

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

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

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

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

LDA classifier

- The LDA classifier plugs the estimates into discriminant function and assigns an observation $X = x$ to the class for which

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

is largest.

- The word linear in the classifier's name stems from the fact that the discriminant functions $\hat{\delta}_k(x)$ are linear functions of x.

- The LDA classifier results from assuming that the observations within each class come from a normal distribution with a class-specific mean vector and a common variance σ2, and plugging estimates for these parameters into the Bayes classifier.

An Example

- Example with $\mu_1 = -1.25, \mu_2 = 1.25, \sigma_1^2 = \sigma_2^2 = 1$, and $\pi_1 = \pi_2 = 0.5$.

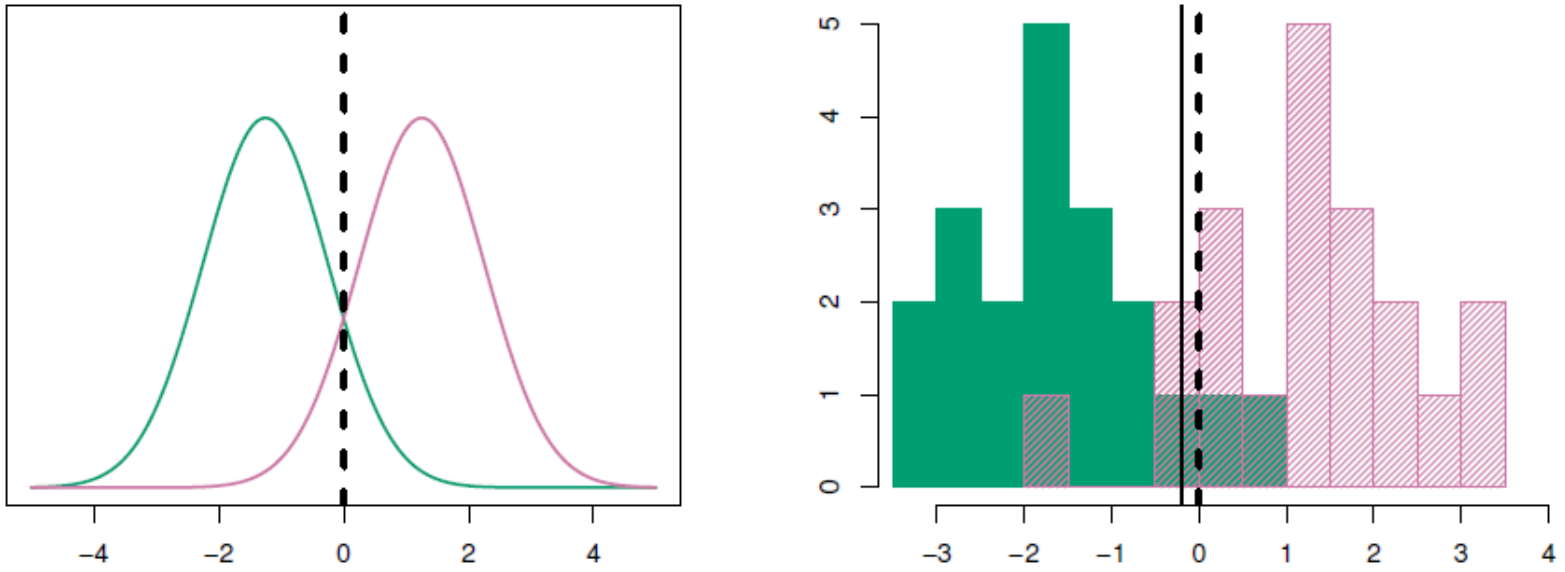


Figure. Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The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

LDA for the case of multiple predictors

- We will assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a multivariate Gaussian (or multivariate normal) distribution, with a class-specific multivariate mean vector and a common covariance matrix.
  - Density: $f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$
  - Discriminant function: $\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$

An Example

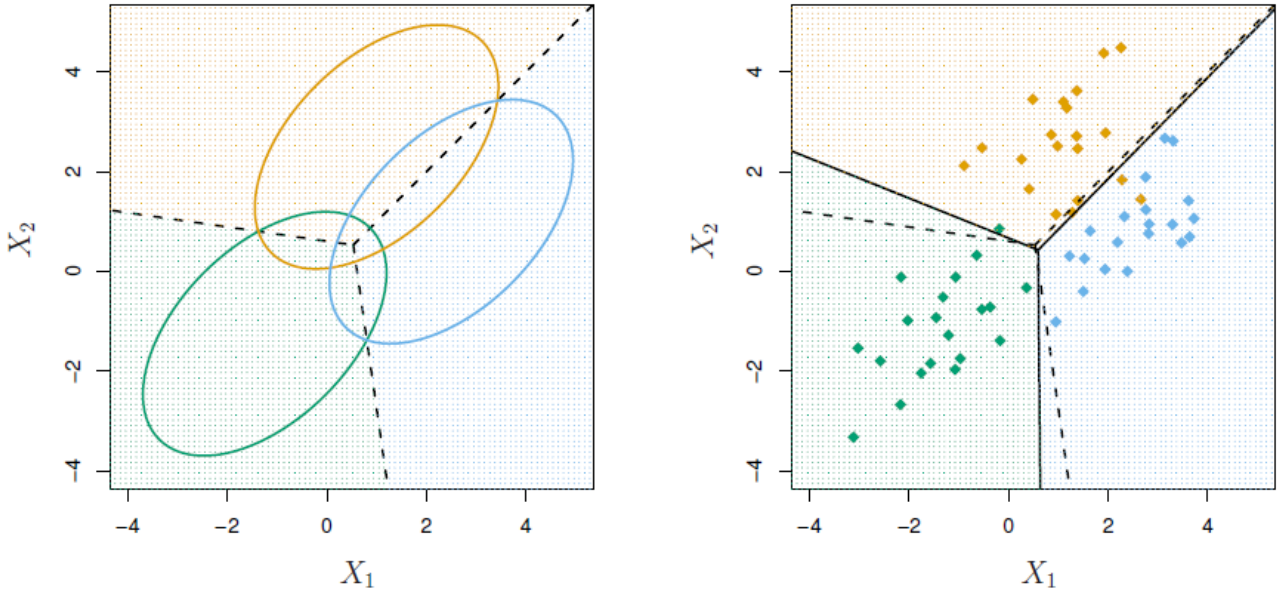


Figure. An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with p = 2, with a class-specific mean vector and a common covariance matrix. Left: Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries. Right: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.

- Here $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$
- The dashed lines are known as the Bayes decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

How to calculate the 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 $\widehat{\Pr}(Y = k \mid X = x)$ is largest.
- When $K = 2$, we classify to class 2 if $\widehat{\Pr}(Y = 2 \mid X = x) \geq 0.5$, else to class 1.

LDA on Default data set

		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!
- **However,**
  - This is training error, and we may be overfitting. Training error rates will usually be lower than test error rates, which are the real quantity of interest. The reason is that we specifically adjust the parameters of our model to do well on the training data.

LDA on Default data set

		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!
- **However,**
  - Since only 3.33% of the individuals in the training sample defaulted, a simple but useless classifier that always predicts that each individual will not default, regardless of his or her credit card balance and student status, will result in an error rate of 3.33%.

LDA on Default data set

		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!
- **However,**
  - Of the true No's, we make $23/9667 = 0.2\%$ errors; of the true Yes's, we make $252/333 = 75.7\%$ errors!
  - So while the overall error rate is low, the error rate among individuals who defaulted is very high. From the perspective of a credit card company that is trying to identify high-risk individuals, an error rate of 252/333 = 75.7% among individuals who default may well be unacceptable.

Types of errors

		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*	

- False positive rate: The fraction of negative examples that are classified as positive – 0.2% in the Default example.
- False negative rate: The fraction of positive examples that are classified as negative – 75.7% in the Default example.
- We produced this table by classifying to class Yes if

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

Varying the threshold

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

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

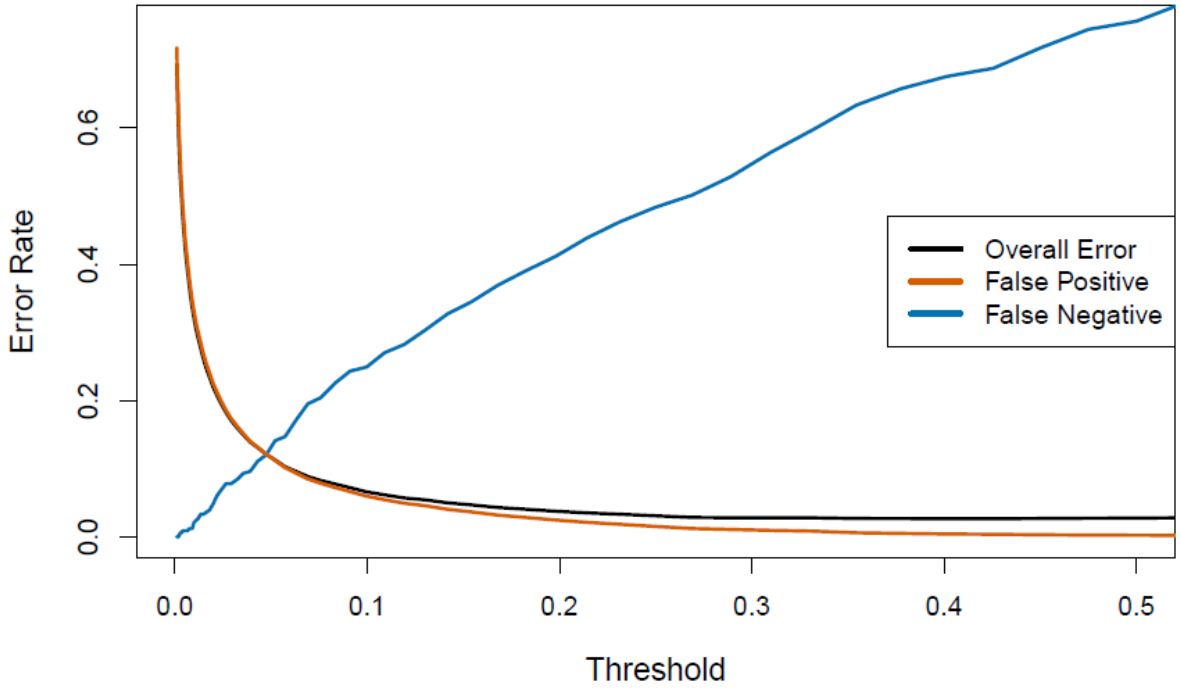


Figure. For the Default data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue line represents the fraction of defaulting customers that are incorrectly classified, and the orange line indicates the fraction of errors among the non-defaulting customers.

ROC curve

- The ROC curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds.
- Sometimes we use the AUC or area under the curve to summarize the overall performance. Higher AUC is good.

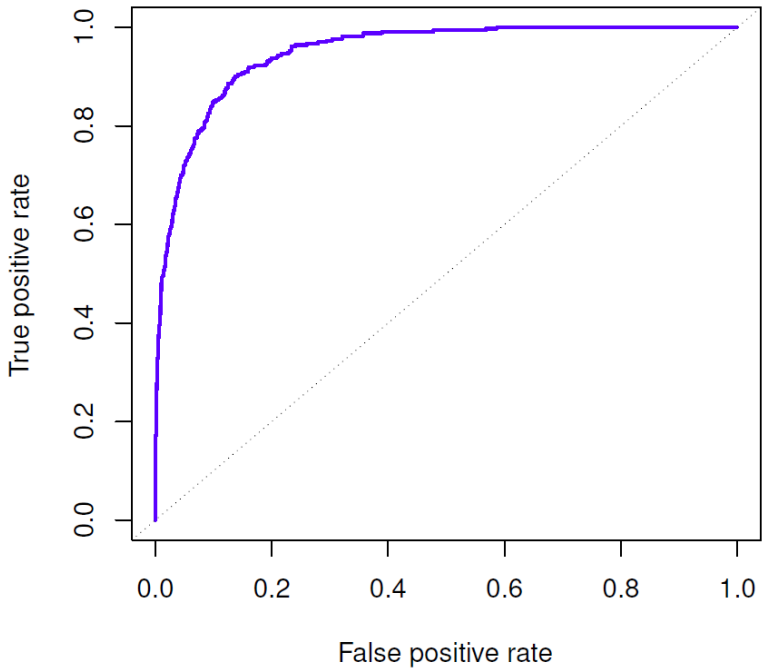


Figure. A ROC curve for the LDA classifier on the Default data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is 1-specificity: the fraction of non-defaulters that we classify incorrectly as defaulters, using that same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. The dotted line represents the "no information" classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.

Quadratic Discriminant Analysis (QDA)

- 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.
- However, unlike LDA, QDA assumes that each class has its own covariance matrix. That is, it assumes that an observation from the *k*th class is of the form $X \sim N(\mu_k, \mathbf{\Sigma}_k)$, where $\mathbf{\Sigma}_k$ is a covariance matrix for the *k*th class.

- The Bayes classifier assigns an observation $X = x$ to the class for which

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

- The QDA classifier involves plugging estimates for Σ_k, μ_k, and π_k into above discriminant function, and then assigning an observation $X = x$ to the class for which this quantity is largest.

- The quantity $\delta_k(x)$ appears as a quadratic function in above discriminant function. This is where QDA gets its name.

An Example

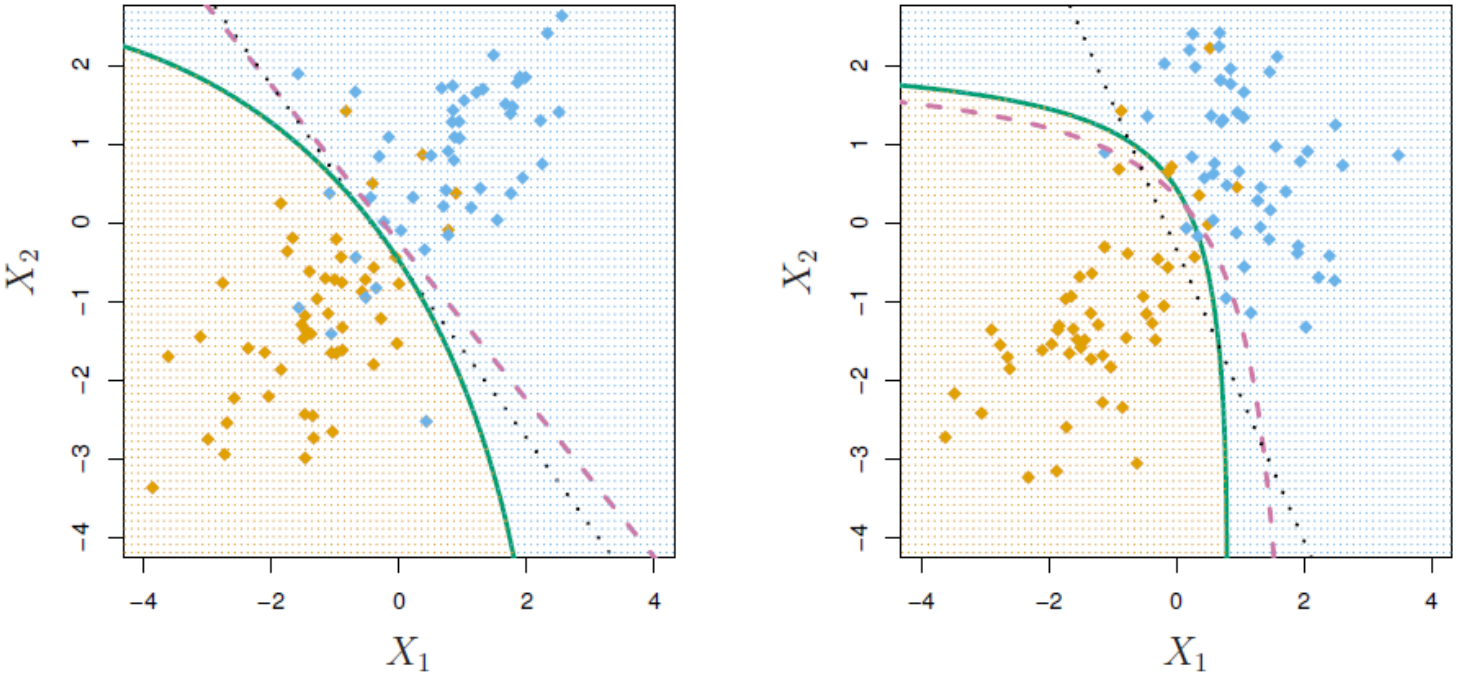


Figure. Left: The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated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

K-nearest neighbors (KNN)

Introduction

- In theory we would always like to predict qualitative responses using the Bayes classifier.
- But for real data, we do not know the conditional distribution of Y given X, and so computing the Bayes classifier is impossible.
- Therefore, the Bayes classifier serves as an unattainable gold standard against which to compare other methods.
- Many approaches attempt to estimate the conditional probability of Y given X, and then classify a given observation to the class with highest estimated probability.
  - One such method is K-nearest neighbors (KNN)

KNN

- Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_o, represented by $\mathcal{N}_0$.
- It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

An example

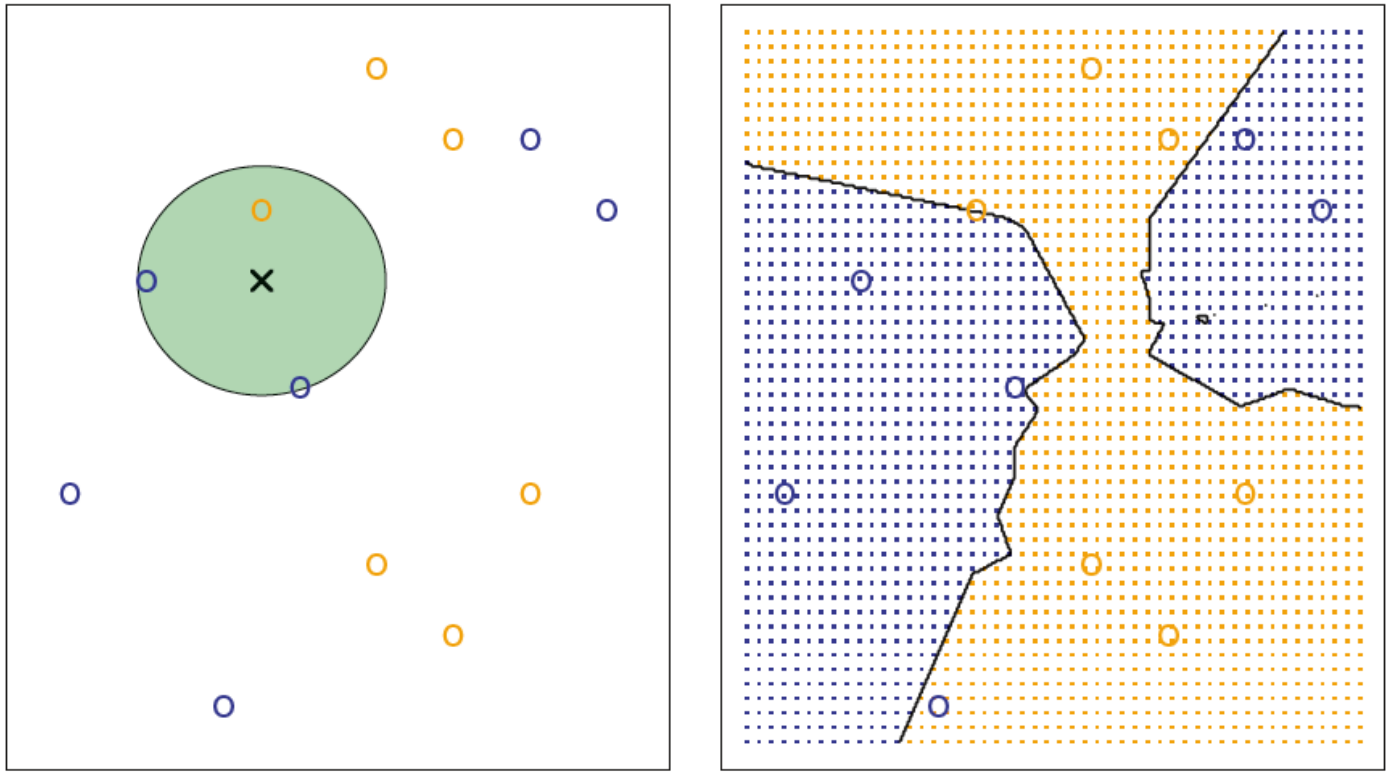


Figure. The KNN approach, using K = 3, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: The KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.