STATISTICAL LEARNING

CHAPTER 4: CLASSIFICATION

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

- In classification, the response variable Y is instead **qualitative** (or **categorical**)
- 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, since it involves assigning the observation to a category, or class
  - 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
- Many possible **classifiers** (classification techniques) exist
  - Most widely-used classifiers: **logistic regression**, **linear discriminant analysis**, and **K-nearest neighbors** (in this chapter)
  - More computer-intensive methods: **generalized additive models** (Chapter 7), **trees**, **random forests**, and **boosting** (Chapter 8), and **support vector machine** (Chapter 9)

An Overview of Classification

- Classification problems occur often, perhaps even more so than regression problems
  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

- Objective of classification
  - Training data: $(x_1, y_1), \ldots, (x_n, y_n)$
  - We use the training data 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

An Overview of Classification

- Default data set
    - We are interested in predicting whether an individual will default on his or her credit are payment, on the basis of annual income and monthly credit card balance
    - Response: default (Y)
    - Predictors: balance (X_1), income (X_2)

Estimating the Coefficients

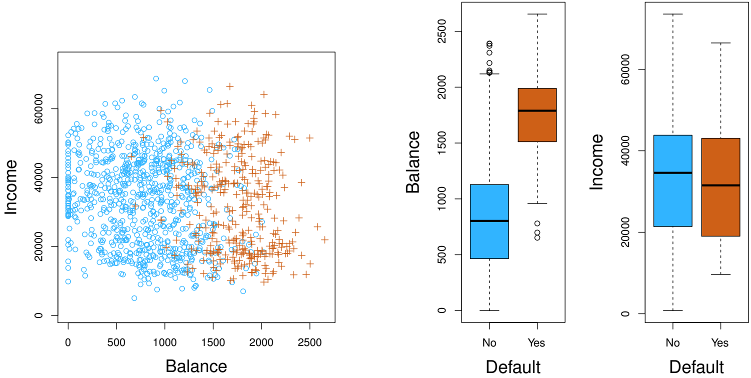

Figure 4.1: The Default data set. Left: The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit are payments are shown in orange, and those who did not are shown in blue. Center: Boxplots of balance as a function of default status. Right: Boxplots of income as a function of default status.

Why Not Linear Regression?

- Linear regression is not appropriate for a qualitative response. Why not?
- Emergency room example
  - Three possible diagnoses: stroke, drug overdose, and epileptic seizure
  - Response variable is encoded as

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

    - This response coding insists that the difference between stroke and drug overdose is the same as the difference between drug overs ode and epileptic seizure. In practice there is no particular reason that this needs to be the case
    - Other combinations are not free from such criticism
- In general there is no natural way to convert a qualitative response with more than two levels into a quantitative response
- If the response has a natural ordering, such as mild, moderate, and severe, then a 1, 2, 3 coding would be reasonable

Why Not Linear Regression?

- For a **binary** (two level) qualitative response, the situation is better
- Suppose there are only tow possibilities for the patient's medical condition: stroke and drug overuse

$$Y = \begin{cases} 0 & \text{if stroke} \\ 1 & \text{if drug overuse} \end{cases}$$

- We could fit a linear regression to this binary response, and predict drug overuse if $\hat{Y} > 0.5$ and stroke otherwise
  - Linear regression will produce the same final prediction even if we flip the above coding
  - It can be shown that $X\hat{\beta}$ is an estimate of $\Pr(\text{drug overuse}|X)$
  - Some of our estimates might be outside the [0,1] interval
  - The dummy variable approach cannot be easily extended to accommodate qualitative responses with more than two levels

Why Not Linear Regression?

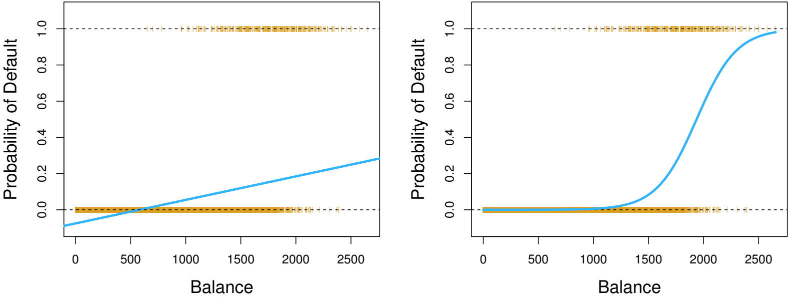

Figure 4.2: Classification using the Default data. Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for default (No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Logistic Regression

- Rather than modeling the response Y directly, logistic regression models the *probability* that Y belongs to a particular category
- For Default data set, for example,
  - The probability of default given balance can be written as

$$\Pr(\text{default} = \text{Yes}|\text{balance}) := p(\text{balance})$$

  - $p(\text{balance})$ will range between 0 and 1
  - One might predict default=Yes for any individual for whom $p(\text{balance}) > 0.5$
  - If a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\text{balance}) > 0.1$

The Logistic Model

- How to model the relationship between $p(X) = \Pr(Y = 1|X)$ and X?
- Linear regression

$$p(X) = \beta_0 + \beta_1 X \tag{4.1}$$

  - This model is shown in the left-hand panel of Figure 4.2
  - In principle, we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others (unless the range of X is limited)

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

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \tag{4.2}$$

  - It ranges between 0 and 1 regardless of a value of X
  - To fit the model (4.2), we use a method called **maximum likelihood**

The Logistic Model

- **Odds**

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{4.3}$$

  - The odds can take on any value between 0 and ∞

- **Log-odds** or **logit**

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{4.4}$$

  - A logit can take on any value between $-\infty$ and ∞
  - A logit is linear in X
  - Increasing X by one unit changes the log odds by β_1 - (4.4)
    $\Leftrightarrow$ Increasing X by one unit multiplies the odds by e^{β_1} - (4.3)

Estimating the Regression Coefficients

- **Maximum likelihood** method is preferred to estimate the parameters
  - **Likelihood function**:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{4.5}$$

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

	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

Table 4.1: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance. A one-unit increase in balance is associated with an increase in the log odds of default by 0.0055 units.

Making Predictions

- From Table 4.1 for Default data set,
  - We predict that the default probability for an individual with a balance of \$1,000 is
    $$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576$$
    which is blow 1%
  - The predicted probability of default for an individual with a balance of \$2,000 is much higher, and equals 0.586 or 58.6%

- From Table 4.2 with a qualitative variable student for Default data set
  - Predicted probabilities
    $$\widehat{\Pr}(\text{default} = \text{Yes}|\text{student} = \text{Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0439,$$
    $$\widehat{\Pr}(\text{default} = \text{Yes}|\text{student} = \text{No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Making Predictions

	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

Table 4.2: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable student[Yes] in the table.

Multiple Logistic Regression

- We now consider the problem of predicting a binary response using multiple predictors
  - We can generalize (4.4) as follows:
  $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{4.6}$$
  - Probability is expressed as
  $$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{4.7}$$
  - We use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

Multiple Logistic Regression

	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

Table 4.3: For the Default data, estimated coefficients of the logistic regression model that predicts the probability of default using balance, income, and student status. Student status is encoded as a dummy variable student[Yes], with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, income was measured in thousands of dollars.

- For Table 4.3
  - A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058 \qquad (4.8)$$

  - A non-student with the same balance and income has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105 \qquad (4.9)$$

Multiple Logistic Regression

- For Table 4.3
  - Confounding effect in the variable student can be found
  - **Confounding**: The results obtained using one predictor may be different from those obtained using multiple predictors especially when there is correlation among the predictors

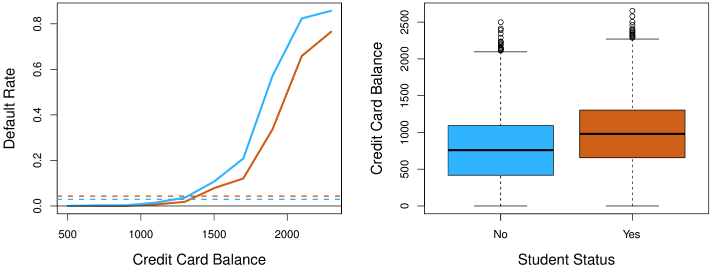

Figure 4.3: Confounding in the Default data. Left: Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of balance, while the horizontal broken lines display the overall default rates. Right: Boxplots of balance for students (orange) and non-students (blue) are shown.

Logistic Regression for >2 Response Classes

- Consider medical example where we had three categories of medical condition in the emergency room: stroke, drug overdose, epileptic seizure
- We wish to model both $\Pr(Y = \text{stroke}|X)$ and $\Pr(Y = \text{drug overuse}|X)$, with the remaining $\Pr(Y = \text{epileptic seizure}|X) = 1 - \Pr(Y = \text{stroke}|X) - \Pr(Y = \text{drug overuse}|X)$
- Two-class logistic regression models have multiple-class extension, which can be done with software available in R
- **Discriminant analysis** is popular for multiple-class classification

Linear Discriminant Analysis

- Logistic regression involves directly modeling $\Pr(Y = k|X = x)$ using the logistic function
- Alternative (and less direct) approach
  1. We model the distribution of the predictors X separately in each of the response classes (i.e. given Y) - these distributions are assumed to be normal
  2. Then use Bayes's theorem to flip these around into estimate for $\Pr(Y = k|X = x)$
- Why do we need another method, when we have logistic regression?
  - 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 model is popular when we have more than two response classes

Using Bayes' Theorem for Classification

- Suppose that the qualitative response variable Y can take on K possible distinct and unordered values
- π_k is the **prior** probability that a randomly chosen observation comes from the kth class
- $f_k(X) \equiv \Pr(X = x|Y = k)$: is the **density function** of X for an observation that comes from the kth class
- **Bayes's theorem** states that

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{4.10}$$

  - We use the abbreviation $p_k(X) = \Pr(Y = k|X)$
  - This suggests that instead of directly computing $p_k(X)$, we can simply plug in estimates of π_k and $f_k(X)$ into (4.10)
  - We will classify an observation to the class for which $p_k(x)$ is greatest
  - We refer to $p_k(x)$ as the **posterior** probability that an observation $X = x$ belongs to the kth class

Linear Discriminant Analysis for $p = 1$

- Assume $p = 1$, that is, we have only one predictor
- Suppose we assume that $f_k(x)$ is **normal** or **Gaussian**

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right) \tag{4.11}$$

- With additional assumption $\sigma_1^2 = \cdots = \sigma_K^2 := \sigma^2$, plug (4.11) into (4.10), we find that

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \tag{4.12}$$

- The Bayes classifier involves assigning an observation $X = x$ to the class for which (4.12) is largest

Linear Discriminant Analysis for $p = 1$

- Taking the log of (4.12) and rearranging the terms. Bayes classifier is equivalent to assigning the observation to the class for which the below is largest:

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \tag{4.13}$$

- If $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier is

$$\hat{y}_k = \begin{cases} 1 & \text{if } 2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2 \\ 2 & \text{if } 2x(\mu_1 - \mu_2) < \mu_1^2 - \mu_2^2 \end{cases}$$

  - In this case, the Bayes decision boundary corresponds to the point where

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \tag{4.14}$$

- Test error rate from Figure 4.4
  - Bayes test error rate : 10.6%
  - LDA test error rate : 11.1%

Linear Discriminant Analysis for $p = 1$

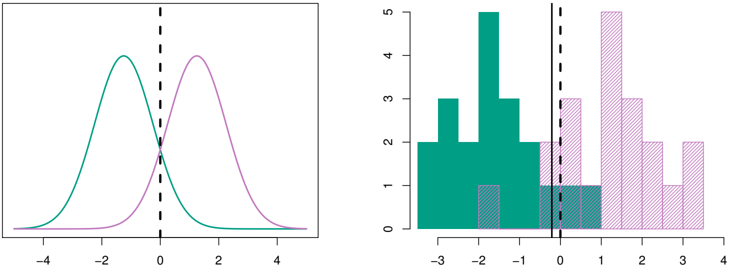

Figure 4.4: Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn fro each of the two classes, and are shown as histograms. The Bayes decision boundary is again as a dashed vertical line. The solid vertical line represent the LDA decision boundary estimated from the training data.

Linear Discriminant Analysis for $p = 1$

- **Linear discriminant analysis** (LDA) method approximates the Bayes classifier by plugging estimates for π_k, μ_k, and σ^2 into (4.13)

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \end{aligned} \tag{4.15}$$

  - The above estimates are calculated using training data set
  - Sometimes we have knowledge of the class membership probabilities $\pi_1, \ldots, \pi_K$, which can be used directly
  - In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class

$$\hat{\pi}_k = n_k / n \tag{4.16}$$

  - LDA classifier

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k \tag{4.17}$$

  - The word **linear** in the classifier's name stems from the fact that the **discriminant functions** $\hat{\delta}_k(x)$ in (4.17) are linear functions of x

Linear Discriminant Analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors
  - We will assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a **multivariate Gaussian** distribution, say $X \sim N(\mu, \mathbf{\Sigma})$
  - The multivariate Gaussian density is defined as

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right) \qquad (4.18)$$

  - Assume that the observation in the kth class are drawn from a multivariate Gaussian distribution $f_k(x) = N(\mu_k, \mathbf{\Sigma})$
  - Bayes classifier assigns an observation $X = x$ to the class for the below is largest

$$\delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + \log\pi_k \qquad (4.19)$$

  - Bayes decision boundary for $k \neq l$ when $\pi_k = \pi_l$,

$$x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k = x^T\mathbf{\Sigma}^{-1}\mu_l - \frac{1}{2}\mu_l^T\mathbf{\Sigma}^{-1}\mu_l \qquad (4.20)$$

- Test error rate from Figure 4.6
  - Bayes test error rate : 0.0746
  - LDA test error rate : 0.0770

Linear Discriminant Analysis for $p > 1$

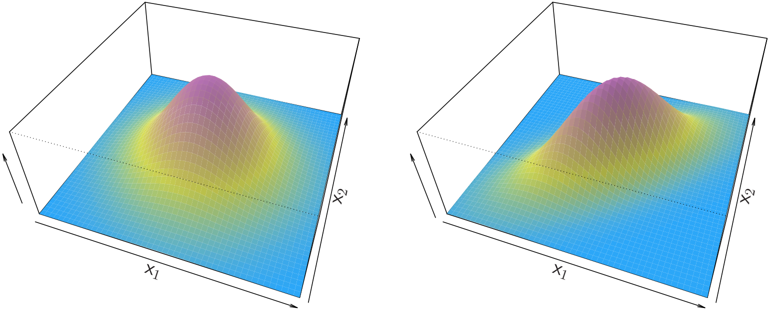

Figure 4.5: Two multivariate Gaussian density functions are shown, with $p = 2$. Left: The two predictors are uncorrelated. Right: The two variables have a correlation of 0.7.

Linear Discriminant Analysis for $p > 1$

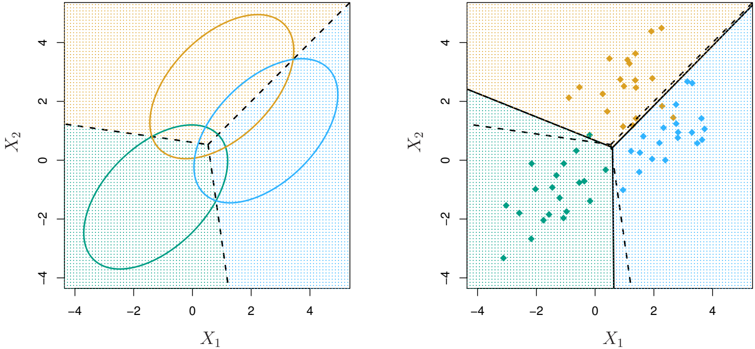

Figure 4.6: An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix. Left: Ellipse that contain 95% of the probability for each of the three classes are shown. The 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.

Linear Discriminant Analysis for $p > 1$

- For Default data set
    - The LDA fit to the 10,000 training samples results in a **training** error rate of 2.75%

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

Table 4.4: A confusion matrix compare 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.

- Two caveats must be noted
    1. Training error rates will be lower than test error rates, which are the real quantity of interest. The higher the ratio of p to n, the more we expect the **overfitting** to play a role (In this example, $p = 3$ and $n = 10000$ so it does not really matter)
    2. Since only 3.33% of the individuals in the training sample defaulted, a simple but useless classifier that always predicts that each individuals will not default, regardless of his or her credit card balance and student status, will result in an training error rate of 3.33%

Linear Discriminant Analysis for $p > 1$

- Two types of errors in a binary classifier
  - It can incorrectly assign an individual who defaults to the *no default* category
    - **Sensitivity** is the percentage of true defaulters that are identified : $1 - 252/333 = 1 - 0.757 = 24.3\%$
  - It can incorrectly assign an individual who does not default to the *default* category
    - **Specificity** is the percentage of non-defaulters that are correctly identified : $1 - 23/9,667 = 99.8\%$
- Why does LDA have such a low sensitivity?
  - Credit company might particularly wish to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default is less problematic
  - However, the Bayes classifier will yield the smallest possible total number of misclassified observations, irrespectively of which class the errors come from
  - The Bayes classifier assigns an observation to the *default* class if

$$\Pr(\text{default} = \text{Yes}|X = x) > 0.5 \tag{4.21}$$

Linear Discriminant Analysis for $p > 1$

- Why does LDA have such a low sensitivity?
  - Instead of (4.21), we can assign an observation to this class if

$$\Pr(\text{default} = \text{Yes} | X = x) > 0.2 \tag{4.22}$$

  - Confusion matrix with (4.22)

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

Table 4.5: A confusion matrix compare 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 is $195/333 = 58.5\%$ and specificity is $9,432/9,667 = 97.6\%$
- Overall error rate has slightly increased to $(138 + 235)/10,000 = 3.73\%$

Linear Discriminant Analysis for $p > 1$

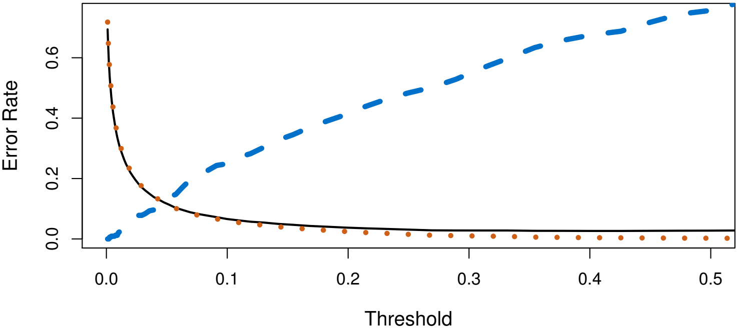

Figure 4.7: For the Default data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.

Linear Discriminant Analysis for $p > 1$

- Overall error rates, sensitivity, specificity vary as a function of the threshold value
    - Decision of the threshold value must be based on *domain knowledge*, such as detailed information about the cost associated with default
- **Receiver operating characteristics** (**ROC**) curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds
    - It is displayed in the axes, one of which is the **true positive rate** (same as sensitivity) and the other of which is the **false positive rate** (same as 1-specificity)
    - The overall performance of a classifier, summarized over all possible thresholds, is given by the **area under the curve** (AUC)
    - An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier
    - For the Default data set, AUC is 0.95, which is close to the maximum of 1 so would be considered very good

Linear Discriminant Analysis for $p > 1$

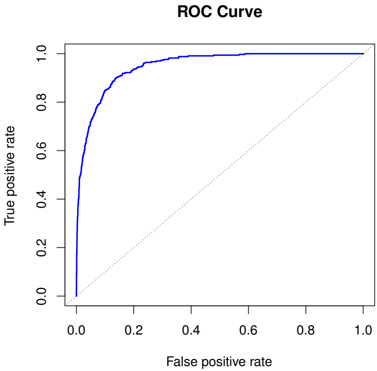

Figure 4.8: A ROC curve for the LDA classifier on the Default data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly identified, 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 the same threshold value. The ideal ROC curve hugs the top left corner, indicating a high true positive classifier; this is what we would expect if student status and credit card balance are not associated with probability of default.

Linear Discriminant Analysis for $p > 1$

		Predicted class		
		- or Null	+ or Non-null	Total
True	- or Null	True Neg. (TN)	False Pos. (FP)	N
class	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

Table 4.6: Possible results when applying a classifier or diagnostic test to a population.

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^*	

Table 4.7: Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.

Quadratic Discriminant Analysis

- **Quadratic discriminant analysis** (QDA) assumes that each class has its own covariance matrix, relaxing the common covariance assumption in LDA
  - QDA assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \mathbf{\Sigma}_k)$
  - The Bayes classifier assigns an observation $X = x$ to the class for which the below is largest

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log \pi_k - \frac{1}{2}|\mathbf{\Sigma}_k| \\ &= -\frac{1}{2}x^T \mathbf{\Sigma}_k^{-1} x + x^T \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}_k^{-1}\mu_k + \log \pi_k - \frac{1}{2}|\mathbf{\Sigma}_k| \quad (4.23)\end{aligned}$$

  - QDA classifier involves plugging estimates for $\mathbf{\Sigma}_k$, μ_k, and π_k into (4.23), and then assigning an observation $X = x$ to the class for which this quantity is largest
  - Unlike in (4.19), the quantity x appears as a **quadratic** function in (4.23)

Quadratic Discriminant Analysis

- Why would one prefer LDA to QDA, or vice-versa? The answer lies in the bias-variance trade-off
  - QDA estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.
    - With $p = 50$ predictors this is some multiple of 1,225, which is a lot of parameters
  - LDA is a much less flexible classifier than QDA, and so has substantially lower variance. This can potentially lead to improved prediction performance
  - But there is a trade-off: if LDA's assumption that the K classes share a common covariance matrix is badly off, then LDA can suffer from high bias
  - Roughly speaking, LDA tends to be a better bet than QDA if there are relatively few training observations and so reducing variance is crucial. In contrast, QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern, or if the assumption of a common covariance matrix for the K classes is clearly untenable

Quadratic Discriminant Analysis

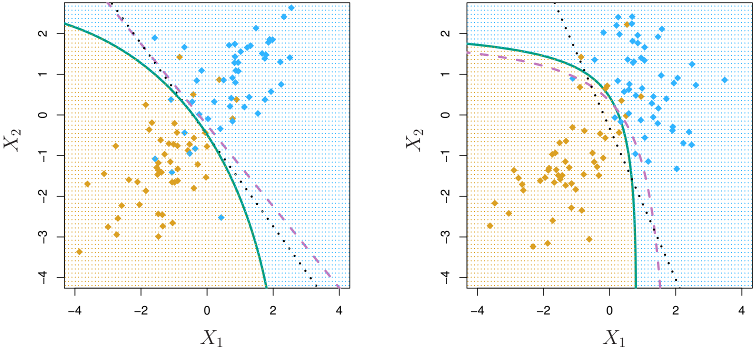

Figure 4.9: Left: The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA. Right: Details are as given in the left-hand panel, except that $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.

A Comparison of Classification Methods

- Connection between the logistic regression and LDA methods for two classes with $p = 1$
  - Let $p_1(x)$ and $p_2(x)$ be the probabilities that the observation $X = x$ belongs to the class 1 and 2, respectively
  - In the LDA framework, the log odds is given by

$$\log\left(\frac{p_1(x)}{1 - p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x \qquad (4.24)$$

    where c_0 and c_1 are functions of μ_1, μ_2 and σ^2
  - In the logistic regression framework, from (4.4),

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x \qquad (4.25)$$

  - Both (4.24) and (4.25) are linear functions in x. Hence, both logistic regression and LDA produces linear decision boundaries
  - The only difference between 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
  - LDA can provide some improvements over logistic regression when the Gaussian assumption approximately holds

A Comparison of Classification Methods

- KNN is a completely non-parametric approach
    - No assumptions are made about the shape of the decision boundary
    - We can expect KNN to dominate LDA and logistic regression when the decision boundary is highly non-linear
    - KNN does not tell us which predictors are important; we don't get a table of coefficients as in Table 4.3
- QDA serves as a compromise between the non-parametric KNN method and the linear LDA and logistic regression approaches
    - Since QDA assumes a quadratic decision boundary, it can accurately model a wider range of problems than can the linear methods
    - Though not as flexible as KNN, QDA can perform better in the presence of a limited number of training observations because it does make some assumption about the form of the decision boundary

A Comparison of Classification Methods

- 6 scenarios for Figures 4.10 and 4.11
  1. Scenario 1: 20 training observations for each class, two predictors, Gaussian distributions with different means and the same independent covariance
  2. Scenario 2: same as scenario 1, except that the two predictors had a correlation of -0.5
  3. Scenario 3: 50 training observation for each class, X_1 and X_2 from the t-distribution
  4. Scenario 4: Gaussian distribution, correlation of 0.5 for the first class and -0.5 for the second class
  5. Scenario 5: Gaussian distribution with uncorrelated predictors, the responses were sampled from the logistic function using X_1^2, X_2^2, and $X_1 \times X_2$ as predictors
  6. Scenario 6: Details are as in the scenario 5, but the responses were sampled from a more complicated non-linear function

A Comparison of Classification Methods

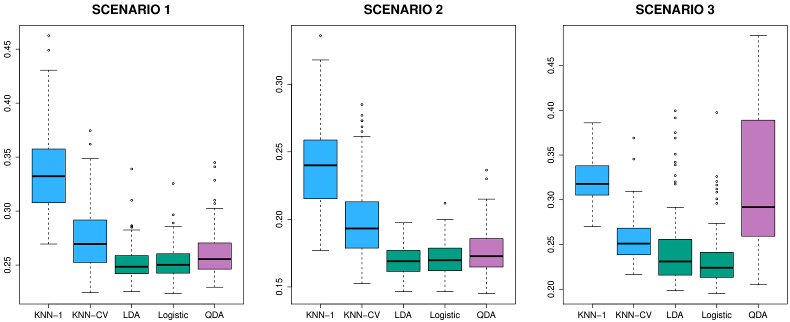

Figure 4.10: Boxplots of the test error rates for each of the linear scenarios described in the main text.

A Comparison of Classification Methods

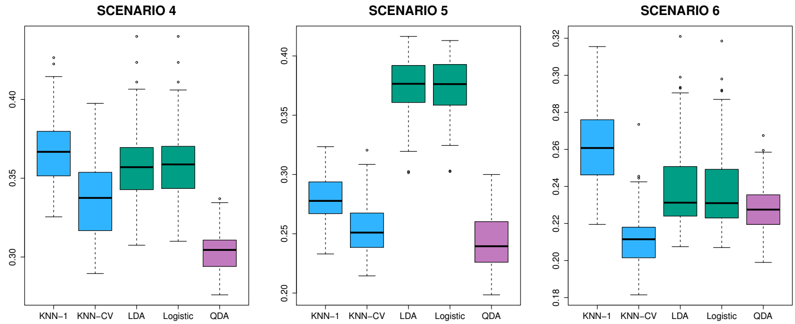

Figure 4.11: Boxplots of the test error rates for each of the non-linear scenarios described in the main text.

Lab: Logistic Regression, LDA, QDA, and KNN