Outline: **Supervised Classification**
(F. Chiaromonte)

Introduction to Statistical Learning
Chapter 4

Units\Features	X_1	X_2	…	X_p	Y
Unit 1	x_{11}	x_{12}		x_{1p}	y_1
Unit 2	x_{21}	x_{22}		x_{2p}	y_2
. . .					
Unit n	x_{n1}	x_{n2}		x_{np}	y_n

p features and one **categorical response** measured on n units:

- An n x p data matrix X, plus a columns of labels
- A data cloud of n labeled points in R^p .

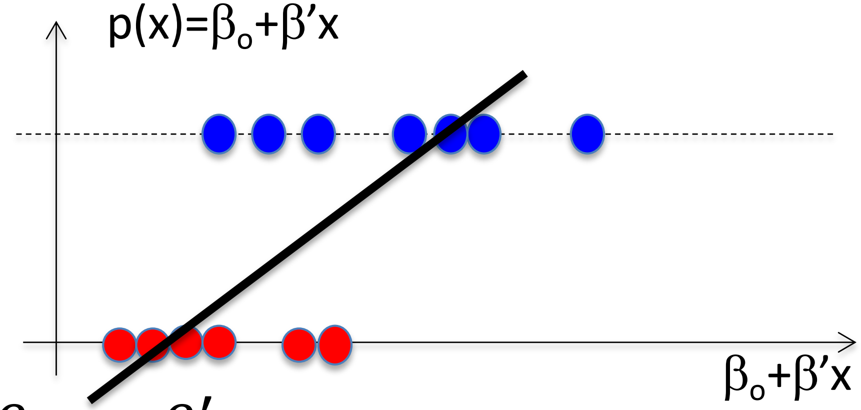
General idea: using the data available, train (estimate the parameter of) an algorithm (model) that will allow one to predict the label of a new unit based on the values in its feature vector $x_{(new)} = (x_{1(new)} \dots x_{p(new)})^T$

Simple methods (Chapter 4):

- **Logistic Regression**
- **Linear and Quadratic Discriminant Analysis** (LDA, QDA)
- **K-Nearest Neighbors** (see also Chapter 2)

Many more sophisticated algorithms exist, e.g., Tree-Based methods (Chapter 8), Support Vector Machines (SVM, Chapter 9)

LOGISTIC REGRESSION



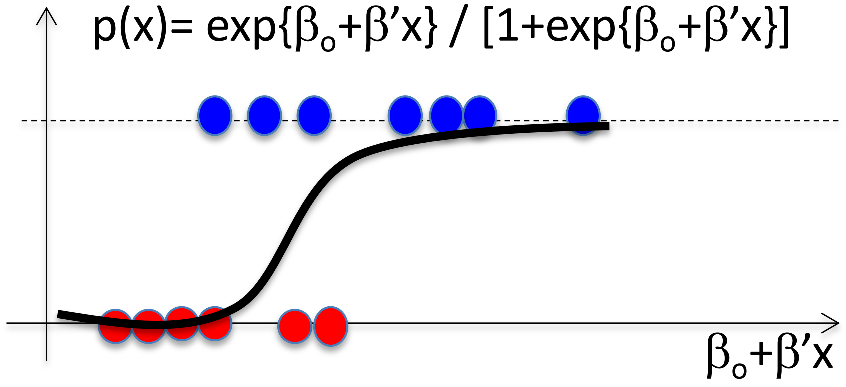
Binary Y, encoded (arbitrarily) as {0,1}

The standard Linear Model would be:

$$E(Y|x) = \Pr(Y = 1|x) = p(x) = \beta_o \ + \beta' \ x$$

$$Y|x \sim p(x) + \varepsilon, \varepsilon \sim N(0, \beta\sigma^2)$$

… can produce p(x) estimates smaller than 0 or larger than 1, and adding Gaussian error makes no sense. Instead, use a *link function* (Generalized Linear Model); the LOGIT:



$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_o + \beta' x$$

Log of the *odds ratio* of a "1" modeled as a linear function of x.

Stochastic component in the mechanism producing the data is represented through a Bernoulli scheme, not through a Gaussian error about the expected value.

Parameters are estimated by *maximum likelihood*; p(x) can be predicted for any level of x:

$$\hat{p}(x) = \frac{\exp\{\hat{\beta}_o + \hat{\beta}' x\}}{1 + \exp\{\hat{\beta}_o + \hat{\beta}' x\}}$$

and the label can be predicted (classification) based on whether this is > 0.5 (a threshold).

Remark: likewise a standard linear model, some of the features included as predictors could be *binary or categorical* (through dummy variables).

Remark: In the credit default example used in the book, student status (=0,1) has a positive effect on the log-odds ratio of default when considered by itself (being a student, marginally, increases the odds of defaulting) but a negative value when considered together with credit card balance (given any given level of outstanding balance, being a student decreases the odds of defaulting). An example of *confounding*. Due to associations among predictors; the effects of a feature considered by itself or in the context of other features are different!

Remark: MLEs in Logistic regression are highly *unstable* when the two classes are well separated.

LINEAR DISCRIMINANT ANALYSIS

Very close relative of Logistic regression that:

- Works also when Y has K > 2 than two classes
- Overcomes instability presented by the MLEs in Logistic regression when the two classes are well separated.

For k=1,2… K, instead of modeling $p_k(x)$ = Pr{Y=k|x}, model $f_k(x)$ = density of X given Y = k .

Let π_k = *prior* Pr{Y=k}, by Bayes Theorem we have:

$$p_k(x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

The *Bayes classifier rule* (optimal performance under certain assumptions) will attribute a new unit to the class $k^*(x_{(new)})$ for which $p_k(x_{(new)})$ is highest.

LDA assumes a model for $f_k(x)$ and implements estimation of $p_k(x)$ accordingly.

Assume $X \mid Y = k \sim N_p(\mu_k, \Sigma)$ p-variate Gaussians with different mean vectors but the same variance-covariance matrix

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

Bayes classifier *discriminant function* (classify to highest)

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

Note: linear function of x

Estimate $\hat{\pi}_k = \frac{n_k}{n}$, $k = 1, 2 \ldots K$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \quad , \quad k = 1, 2 \ldots K$$

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

pooled estimate of the (common) “within class” Σ

LDA discriminant function (classify to highest) approximates Bayes classifier

$$\hat{\delta}_k(x) = x'\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k'\hat{\Sigma}^{-1}\hat{\mu}_k + \log \hat{\pi}_k$$

Example p=1, K=2. *Boundary*: x s.t. $\delta_1(x) = \delta_2(x)$ (here $x = (\mu_1+\mu_2)/2$ since $\pi_1=\pi_2=1/2$)

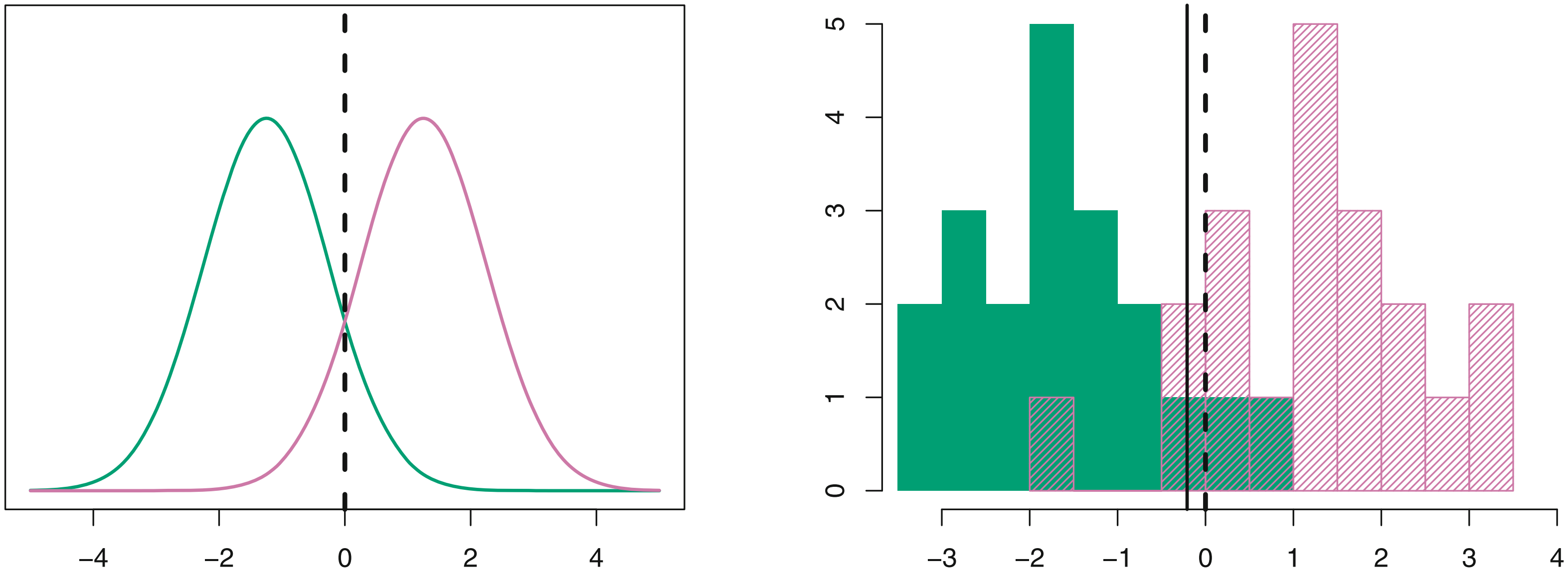


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line.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

Example p=2, K=3. *Boundaries*: x s.t. $\delta_k(x) = \delta_j(x)$, for each k, j pair

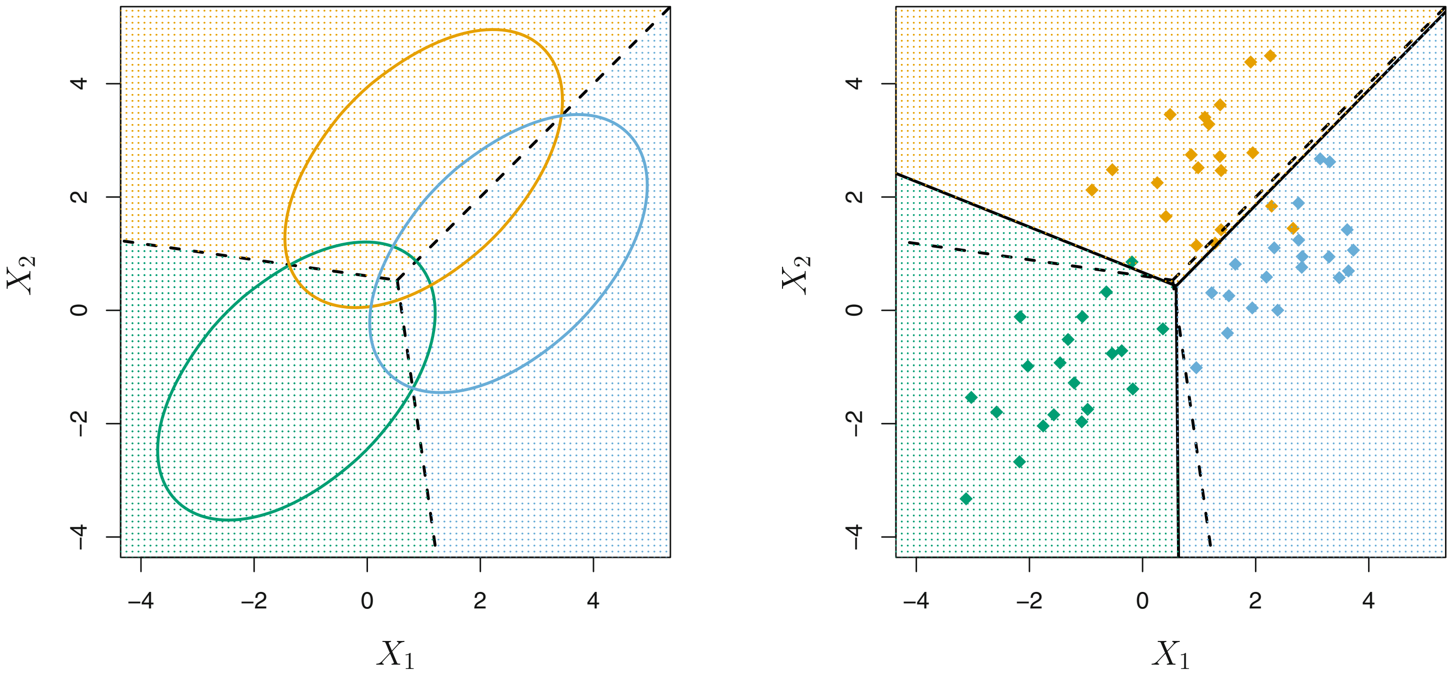


FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown.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

QUADRATIC DISCRIMINANT ANALYSIS

Works exactly the same way, but variance-covariance matrices are allowed to differ among the classes

- We need to estimate a much larger number of parameters!
- The boundaries are allowed to curve; more flexible classifier
- Less *bias* but more *variance*
- More prone to overfitting

In a given application, do you have a large enough sample size to use QDA?

Is it really a stretch to assume common variance-covariance across classes and use LDA?

LDA classifier discriminant function (classify to highest) approximates Bayes classifier

$$\hat{\delta}_k(x) = x'\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\boxed{x_k{}'\hat{\Sigma}_k^{-1}x_k} - \frac{1}{2}\hat{\mu}_k{}'\hat{\Sigma}_k^{-1}\hat{\mu}_k - \frac{1}{2}\log\det(\hat{\Sigma}_k) + \log\hat{\pi}_k$$

Note: quadratic term in x

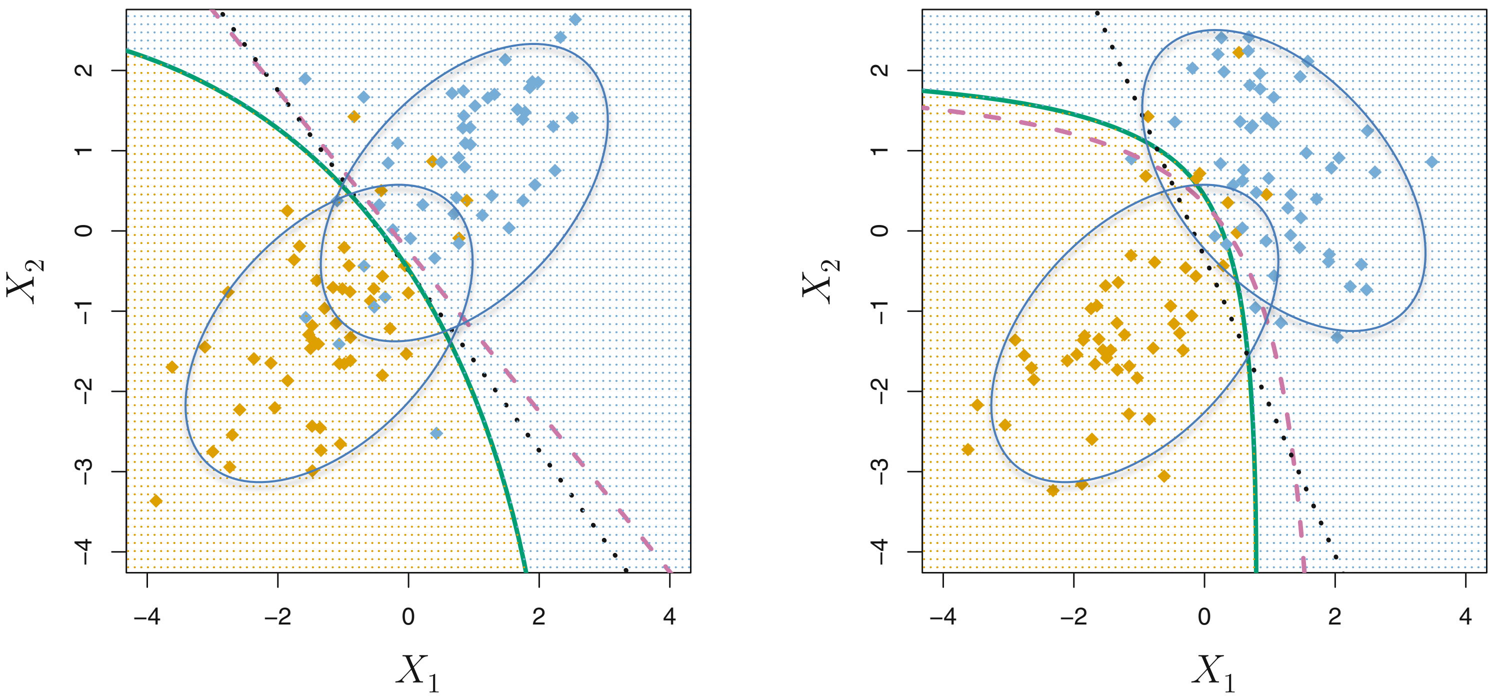
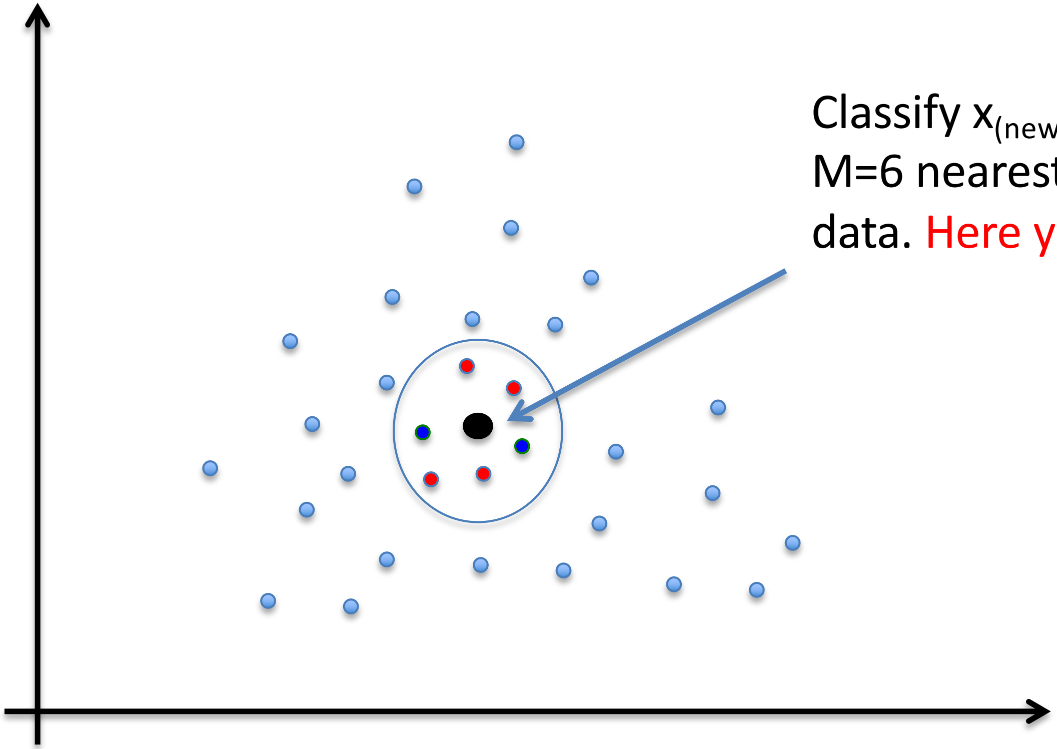


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$*. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated 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.*

K-NEAREST NEIGHBORS CLASSIFICATION

Say M nearest, K used for number of classes (Chapter 2)

- Completely *non-parametric*!
- The boundaries can take any shape; no modeling
- Without modeling one may have less power and/or accuracy
- Need to *select M* appropriately (*tuning parameter* of the algorithm; a “degree of smoothing”)



Classify $x_{(new)}$ with a majority vote by the M=6 nearest neighbors in the training data. Here $y_{(new)}$ = red.

<u>CLASSIFICATION ACCURACY</u>: Training set and Testing set error (misclassification) rates

$$Err_{(train)} = \frac{1}{n_{(train)}} \sum_{i \in Train} Ind(\hat{y}_i \neq y_i)$$

$$Err_{(test)} = \frac{1}{n_{(test)}} \sum_{i \in Test} Ind(\hat{y}_i \neq y_i)$$

Better estimate of the performance of the classifier (on training set, may overfit)

Bayes classifier (and L(Q)DA approximations of it) have optimality properties in terms of overall misclassification rates.

But what if the training set is very unbalanced, misclassification rates for different classes may differ; even with a good overall classification performance, some classes (the least sampled) may be predicted very poorly!

e.g., in the credit default example used in the book, overall misclassification rate is rather low with both Logistic regression and LDA, but misclassification rate for people who do default (a very small fraction of the total) is as bad as ~75%! (and this is likely what the researchers care about the most).

Can move boundaries (thresholds) as to improve classification for a class of interest.

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

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

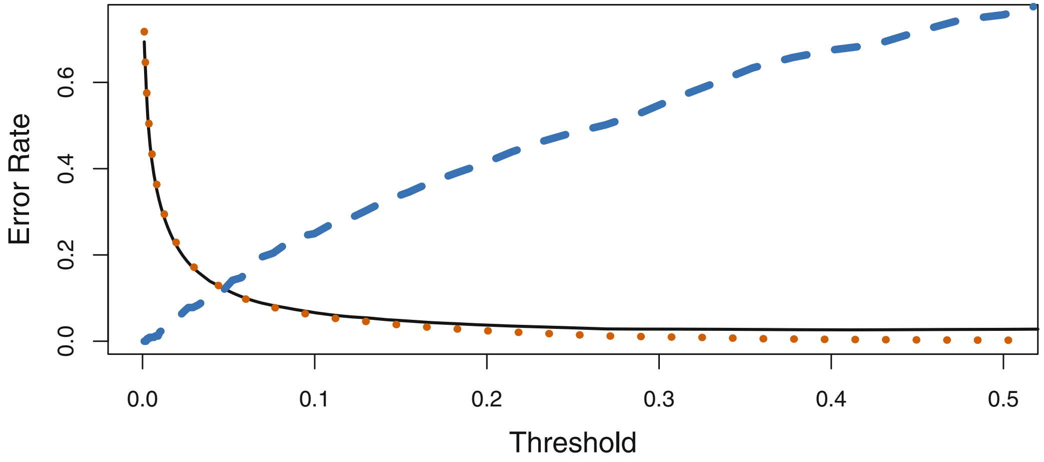
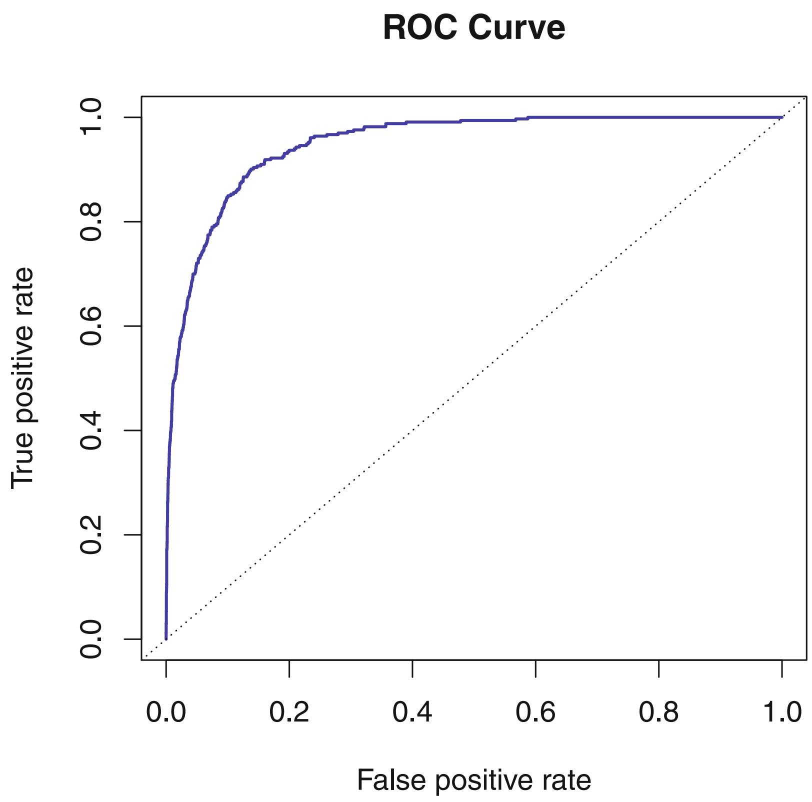


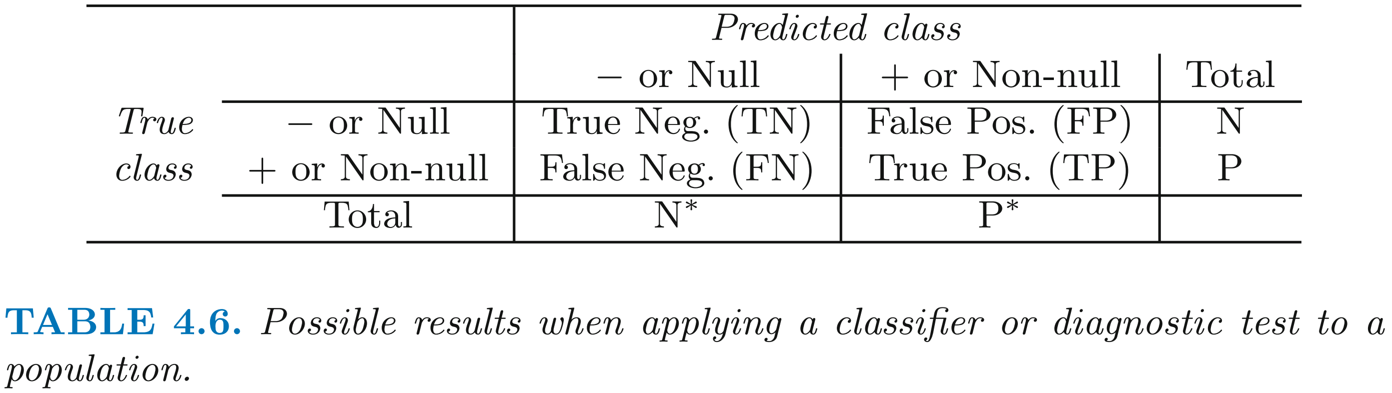
FIGURE 4.7. *For the* Default *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are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



Area Under the Curve (AUD) is a measure of the quality of the classifier – across choices of boundaries/ thresholds

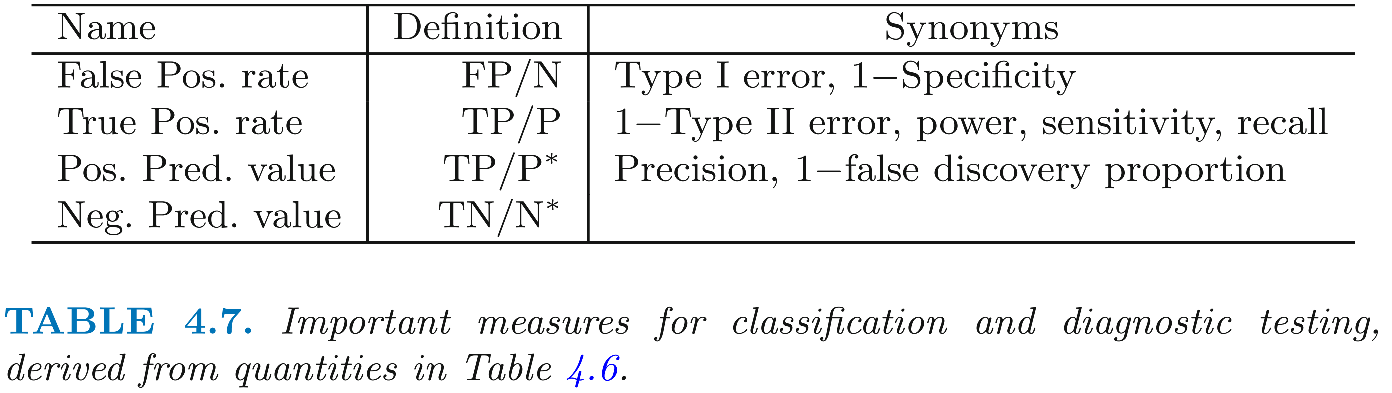
FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly 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 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.*

Nomenclature on classifiers’ performance; K=2



		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.*