DATA 572: Supervised Learning

2023W2

Shan Du

Generative Models for Classification

- Logistic regression involves directly modeling $\Pr(Y = k|X = x)$ using the logistic function, for the case of two response classes.
- In statistical jargon, we model the conditional distribution of the response Y, given the predictor(s) X.
- We now consider an alternative and less direct approach to estimating these probabilities.

Generative Models for Classification

Discriminant Analysis

- In this new approach, we model the distribution of the predictors X separately in each of the response classes (i.e., for each value of Y).
- We then use Bayes' theorem to flip these around into estimates for $\Pr(Y = k|X = x)$.
- When the distribution of X within each class is assumed to be normal, it turns out that the model is very similar in form to logistic regression.

Generative Models for Classification

Generative model vs discriminative model

- Why do we need another method, when we have logistic regression? There are several reasons:
  - When there is substantial separation between the two classes, the parameter estimates for the logistic regression model are surprisingly unstable. The methods that we consider in this section do not suffer from this problem.
  - If the distribution of the predictors X is approximately normal in each of the classes and the sample size is small, then the approaches in this section may be more accurate than logistic regression.
  - The methods in this section can be naturally extended to the case of more than two response classes.

Generative Models for Classification

Discriminant Analysis

- Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$.
- Let π_k represent the overall or *prior probability* that a randomly chosen observation comes from the kth class.
- Let $f_k(X) = \Pr(X|Y = k)$ denote the density function of X for an observation that comes from the kth class.

Generative Models for Classification

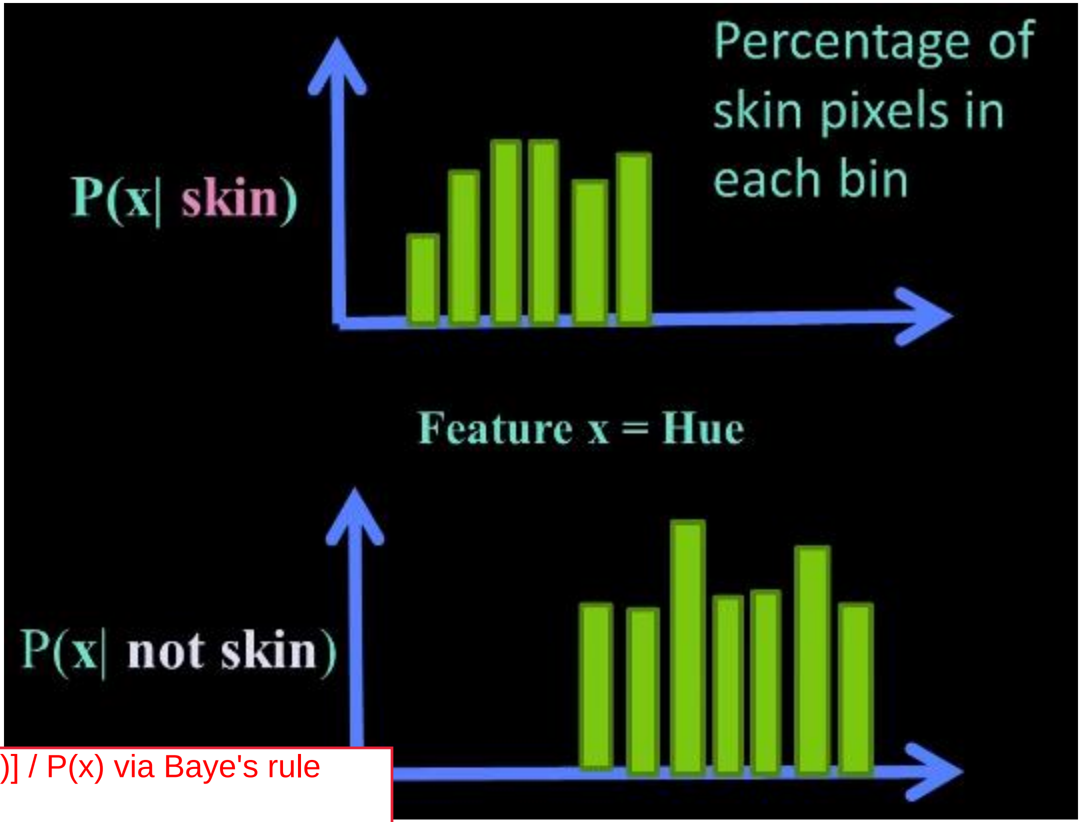
- Then *Bayes' theorem* states that

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

This is the *posterior probability* that an observation $X = x$ belongs to the kth class.

Bayes Thm: P(A|B) = [P(B|A) * P(A)] / P(B)

Example: Learning Skin Colors



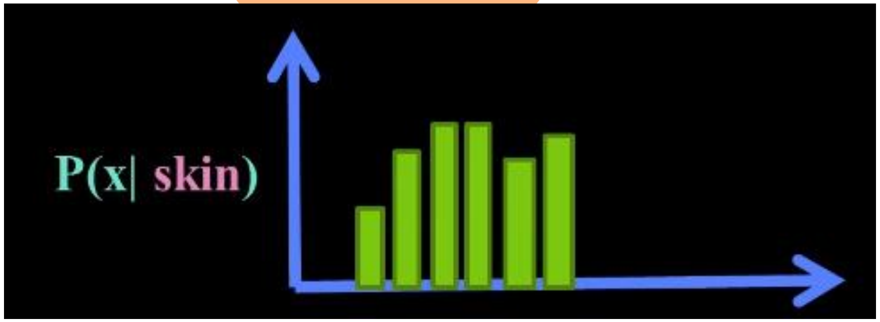
P(skin|x) is posterior probability => [P(skin) * P(x|skin)] / P(x) via Baye's rule

P(x|skin) is likelyhood
P(skin) is prior probability

Now we get a new image, and want to label each pixel as skin or not skin.

Example: Learning Skin Colors

$$\underbrace{P(skin|x)}_{\text{posterior}} = \frac{\underbrace{P(x|skin)}_{\text{likelihood}}\underbrace{P(skin)}_{\text{prior}}}{P(x)}$$



$$P(skin|x) \propto P(x|skin)P(skin)$$

Example: Classifying Skin Pixels

Now for every pixel in a new image, we can estimate probability that it is generated by skin:

If $P(skin|x) > \theta$ classify as skin; otherwise not

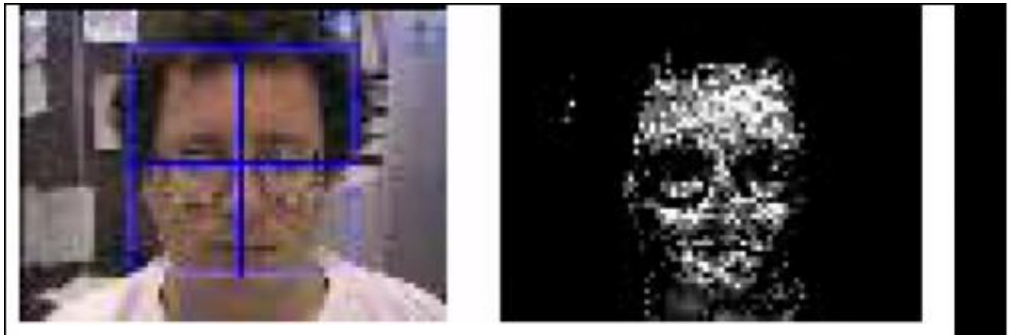


Figure 6: A video image and its flesh probability image

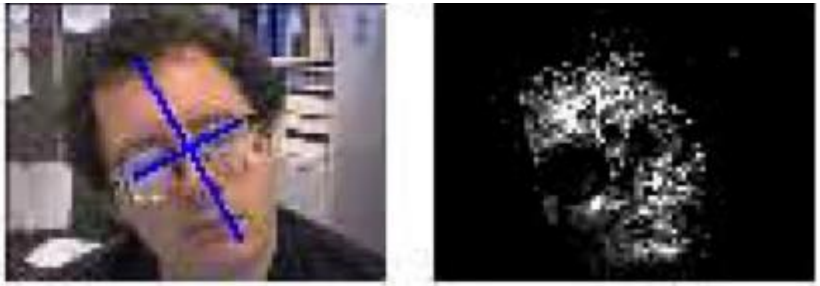


Figure 7: Orientation of the flesh probability distribution marked on the source video image

Generative Models for Classification

- In general, estimating π_k is easy if we have a random sample from the population: we simply compute the fraction of the training observations that belong to the kth class.
- However, estimating the density function $f_k(x)$ is much more challenging.
- To estimate $f_k(x)$, we will typically have to make some simplifying assumptions.

Generative Models for Classification

- By using different estimates of $f_k(x)$, we have three classifiers to approximate the Bayes classifier: *linear discriminant analysis*, *quadratic discriminant analysis*, and *naive Bayes*.

Linear Discriminant Analysis for $p = 1$

- Assume $p = 1$ — that is, we have only one predictor.
- Assume $f_k(x)$ is *normal* or *Gaussian*. In the one-dimensional setting, the normal density takes the form

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

where μ_k and σ_k^2 are the mean and variance parameters for the kth class.

Linear Discriminant Analysis for $p = 1$

- let us further assume that $\sigma_1^2 = \cdots = \sigma_k^2$: that is, there is a shared variance term across all K classes, which for simplicity we can denote by σ^2.
- Then $p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$
- Taking the log on each side

Linear Discriminant Analysis for $p = 1$

Probability of input being class k -> choose class with highest probability

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

- If $K = 2$ and $\pi_1 = \pi_2$, then 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.
- The Bayes decision boundary is the point for which $\delta_1(x) = \delta_2(x)$, then

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

Linear Discriminant Analysis for $p = 1$

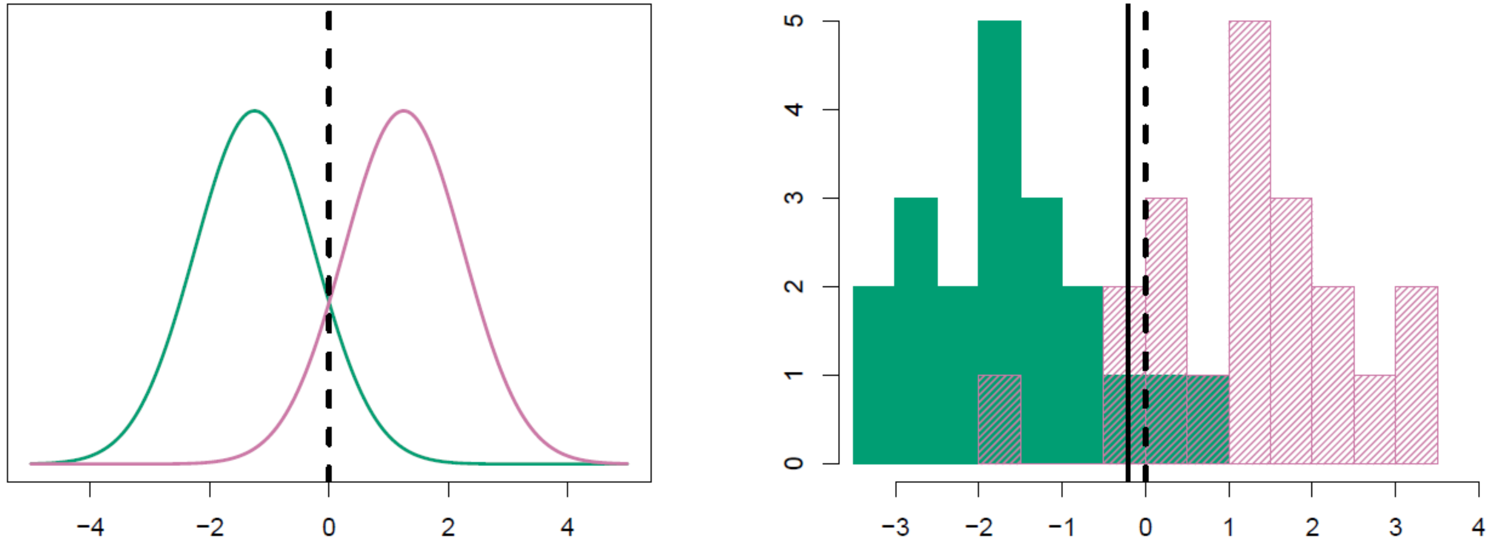


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line.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

Linear Discriminant Analysis for $p = 1$

- In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, to apply the Bayes classifier we still have to estimate the parameters $\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_k$, and σ^2.
- The *linear discriminant analysis* (LDA) method approximates the Bayes classifier by using the estimates:

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

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

where n is the total number of training observations, and n_k is the number of training observations in the kth class.

Linear Discriminant Analysis for $p = 1$

- $\hat{\mu}_k$ is simply the average of all the training observations from the kth class, while $\hat{\sigma}^2$ can be seen as a weighted average of the sample variances for each of the K classes.
- LDA estimates $\boxed{\pi_k}$ using the proportion of the training observations that belong to the kth class. In other words,

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

Linear Discriminant Analysis for $p = 1$

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

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

is largest.

Linear Discriminant Analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors. To do this, we will assume that $X = (X_1, X_2, \dots, X_p)$ is drawn from a *multivariate Gaussian* (or multivariate normal) distribution, with a class-specific multivariate mean vector and a common covariance matrix.
- The multivariate Gaussian distribution assumes that each individual predictor follows a one-dimensional normal distribution, with some correlation between each pair of predictors.

Linear Discriminant Analysis for $p > 1$

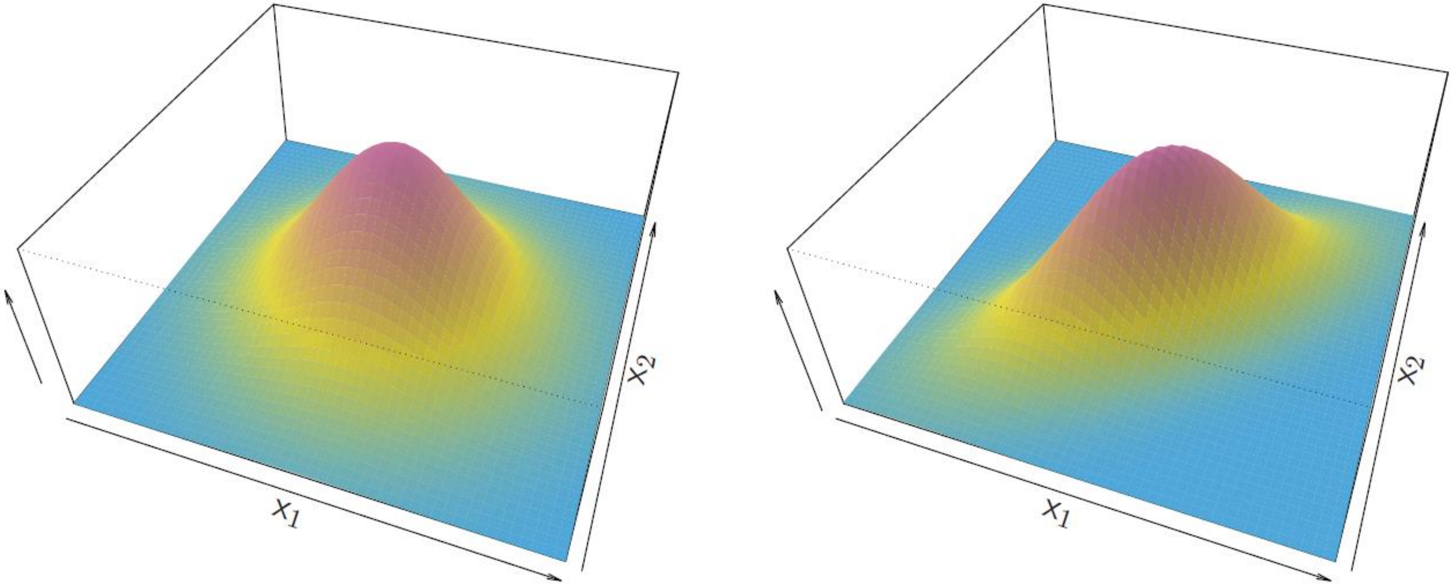


FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

Linear Discriminant Analysis for $p > 1$

- To indicate that a p-dimensional random variable X has a multivariate Gaussian distribution, we write $X \sim N(\mu, \Sigma)$. Here $\mathrm{E}(X) = \mu$ is the mean of X (a vector with p components), and $\mathrm{Cov}(X) = \Sigma$ is the $p \times p$ covariance matrix of X.
- Formally, the multivariate Gaussian density is defined as

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

Linear Discriminant Analysis for $p > 1$

- In the case of $p > 1$ predictors, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \Sigma)$, where μ_k is a class-specific mean vector, and Σ is a covariance matrix that is common to all K classes.

Linear Discriminant Analysis for $p > 1$

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

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

is largest.

Linear Discriminant Analysis for $p > 1$

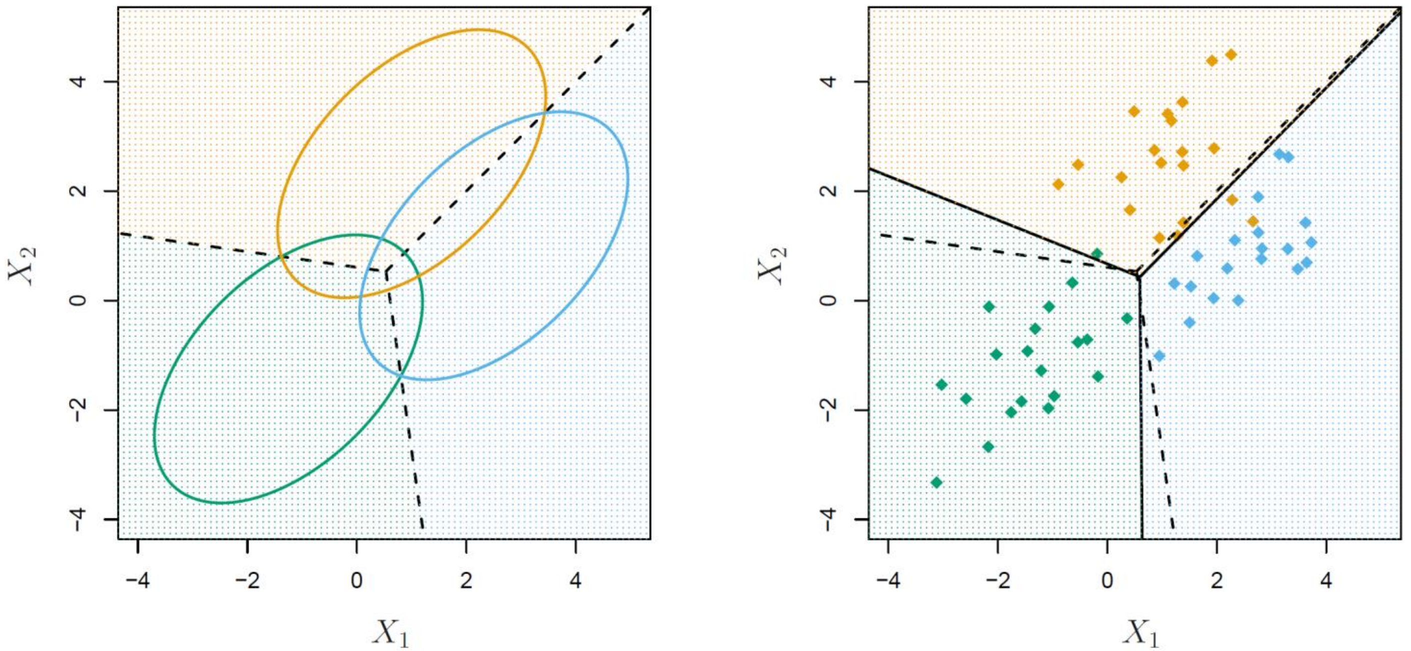


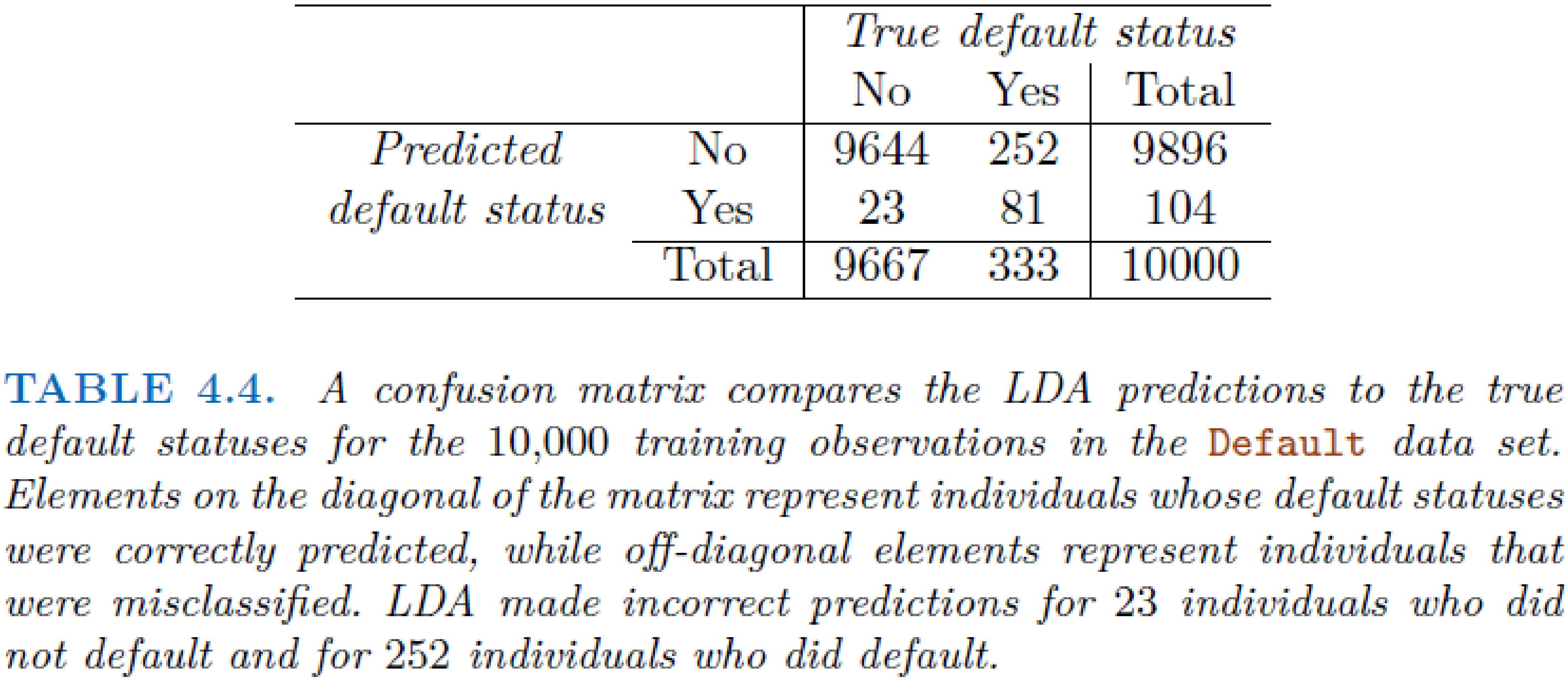
FIGURE 4.6. *An example with three classes. The observations from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix.* Left: *Ellipses that contain 95 % of the probability for each of the three classes are shown.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

Linear Discriminant Analysis for $p > 1$

- In practice, a binary classifier can make two types of errors: it can incorrectly assign an individual who defaults to the *no default* category, or it can incorrectly assign an individual who does not default to the *default* category.
- It is often of interest to determine which of these two types of errors are being made.

Linear Discriminant Analysis for $p > 1$

- A *confusion matrix* is a convenient way to display this information.



		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

TABLE 4.4. *A confusion matrix compares 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.*

Linear Discriminant Analysis for $p > 1$

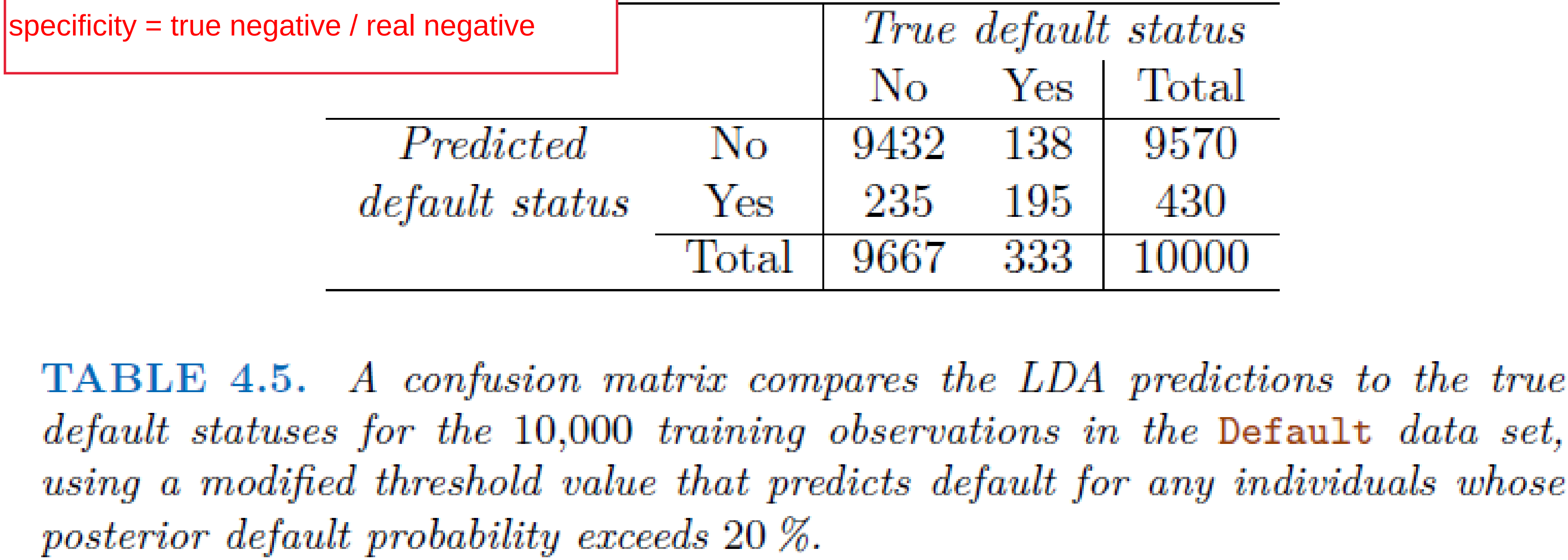
- LDA is trying to approximate the Bayes classifier, which has the lowest total error rate out of all classifiers. That is, the Bayes classifier will yield the smallest possible total number of misclassified observations, regardless of the class from which the errors stem.
- In contrast, a credit card company might particularly wish to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default, though still to be avoided, is less problematic.

Linear Discriminant Analysis for $p > 1$

Precision = true positive / detected positive

recall = true (or correct) positive / real positive

specificity = true negative / real negative



		True default status		
		No	Yes	Total
Predicted	No	9432	138	9570
default status	Yes	235	195	430
	Total	9667	333	10000

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 %.*

Linear Discriminant Analysis for $p > 1$

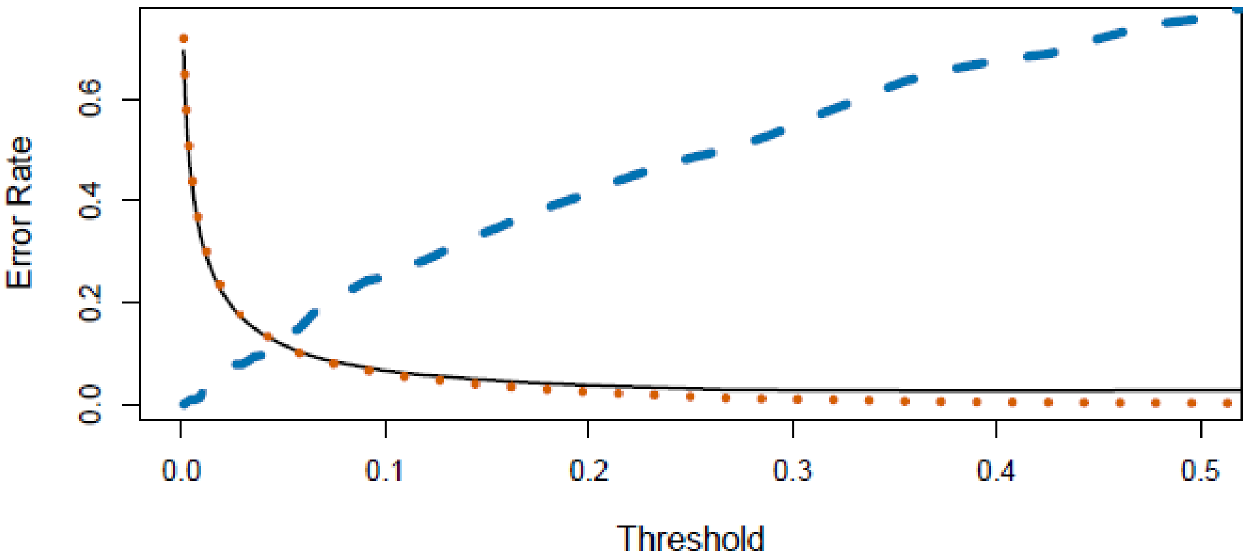


FIGURE 4.7. *For the* Default *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are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

Linear Discriminant Analysis for $p > 1$

- The *ROC curve* is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds. The name “ROC” is historic, and comes from communications theory. It is an acronym for *receiver operating characteristics*.
- The overall performance of a classifier, summarized over all possible thresholds, is given by the *area under* the (ROC) *curve* (AUC). An ideal ROC curve will hug the top left corner, so the larger area the AUC the better the classifier.

Linear Discriminant Analysis for $p > 1$

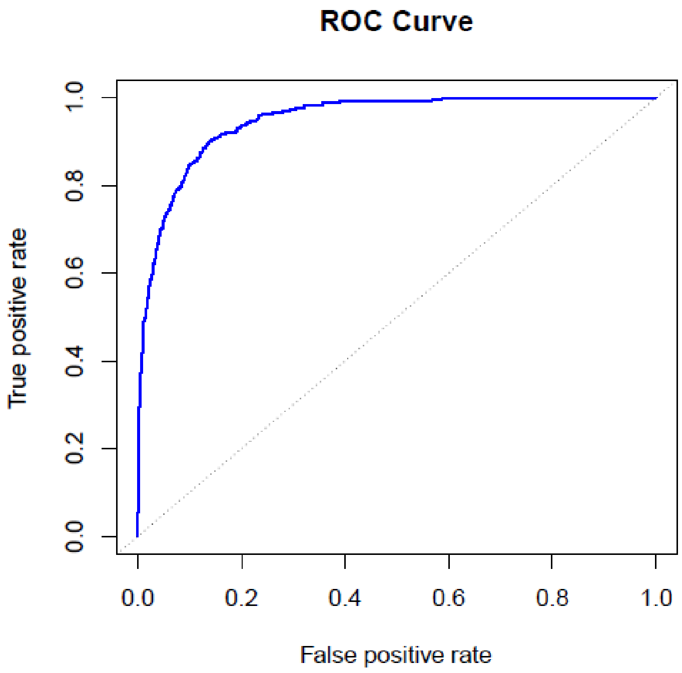


FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data. It traces out two types of error as we vary the threshold value for the posterior probability of default. The actual thresholds are not shown. The true positive rate is the sensitivity: the fraction of defaulters that are correctly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