4.4 Linear Discriminant Analysis

Zongyi Liu

2023-05-19

4.4 Linear Discriminant Analysis

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

There some reasons that we will not use the 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.
- It would be better to use when there are more than two response classes.

4.4.1 Using Bayes' Theorem for Classification

Suppose we wish to classify an observation into one of K classes, where K
$geq2$. We get the π_k to represent the overall **prior** probability that a randomly chosen observation comes from the kth class.

Let $f_k(x) = Pr(X = x|Y = k)$ denote the density function of X for an observation that comes from the kth class. The **Bayes's theorem** has that

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

Here we refer to $p_k(x)$ as the **posterior probability** that an observation X=x belongs to the k-th class.

4.4.2 Linear Discriminant Analysis for p=1

Here we assume that p=1 (meaning we only have one predictor). We would like to get an estimate for $f_k(x)$ in order to estimate $p_k(x)$. We will first assume that $f_k(x)$ is normal or Gaussian, then the normal density would take this form:

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

Then we can further assume that all σ^2 are equal, then plugging back, we can have

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

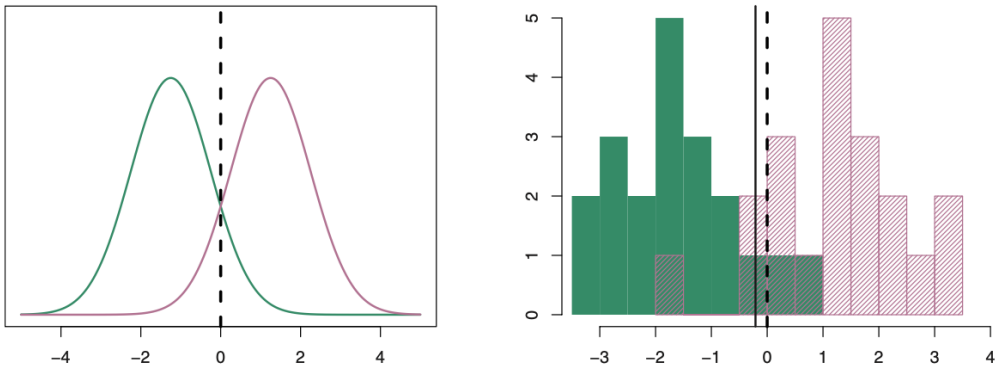
Taking the log of this equation and rearranging the terms, it is not hard to show that it is equivalent to assigning the observation to the class for which

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

is the largest. In this case, the Bayes decision boundary corresponds to the point where

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

Here is an example, in which two normal density functions are displayed. Here $\mu_1 = -1.25$, $mu_2 = 1.25$, and $\sigma_1^2 = \sigma_2^2 = 1$. If we assume that an observation is equally likely to come from either class, then by inspection of equation above, we see that the Bayes classifier assigns the observation to class 1 if x < 0 and class 2 otherwise.



LDA analysis

In practice, even if we know the that X is drawn from a Gaussian distribution within each class, we still need to estimate the parameters
$mu_1, ...$
mu_K,
$pi_1, ...$,
pi_K, and σ^2. The linear discriminant analysis (LDA) method approximates the Bayes classifier by plugging estimates for π_k, mu_k, and $sigma^2$ into equations above.

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

LDA estimates pi_K using the proportion of the training observations that belong to the kth class:

$$\hat{\pi}_k = n_k/n.$$

and get

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

the "linear" in LDA refers to the linear relationship in equation above.

In the right hand panel of the picture above, we can see that the LDA decision boundary is slightly to the left of the optimal Bayes decision boundary.

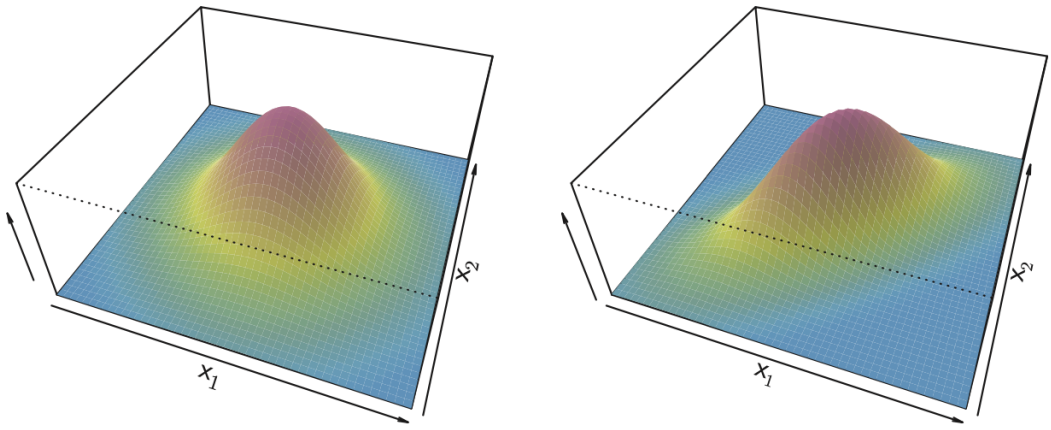
Summary:

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 σ^2, and plugging estimates for these parameters into the Bayes classifier.

4.4.3 Linear Discriminant Analysis for p >1

Then we can 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, which assumes that each individual predictor has a one-dimensional normal distribution. Examples can be seen as below:

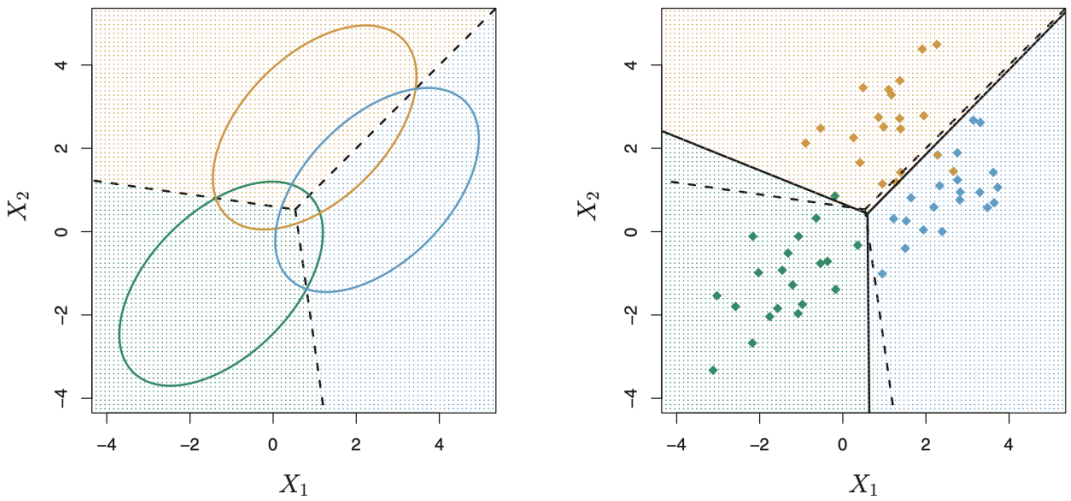


The height of the surface at any particular point represents the probability that both X_1 and X_2 fall in a small region around that point.

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.



To indicate that a p-dimensional RV X has a multivariate Gaussian distribution, we can write $X \sim N(\mu, \sum)$, where $E(X) = \mu$ is the mean of X and $Cov(X) = \sum$ is the $p * p$ covariance matrix of X. The multivariate Gaussian density is defines as

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

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

Thus, we can realize that the Bayes classifier assigns an observation $X = x$ to the class for which

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

is the largest

In the left-hand panel of the picture above, we can see three ellipses, which represent regions that contain 95% of the probability for each of the three classes. They represent the set of values x for which $\sigma_k(x) = \sigma_l(x)$, ie:

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

for $k \neq l$. Also, we should notice that there are three lines above, separating there pairs of classes one by one.

Example: Perform LDA on the `Defualt` data.

After performing it on 10,000 training samples, we can get a training error rate of 2.75%, this seems to be a small rate, but there are still two disadvantages:

- The training error rates will usually be lower than test error rates. The higher the ratio of parameters p to number of samples n, the more we expect an **overfitting** to exist in the model (here we have p=2 and n=10,000).

- The trivial null classifier will achieve an error rate that is only a bit higher than the LDA training set error rate.

In practice, the binary classifier like this might assign the observation to the category which it should not belong to. For this error, which can have a **confusion matrix** as below:

		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

From the table we can come up with two terms:

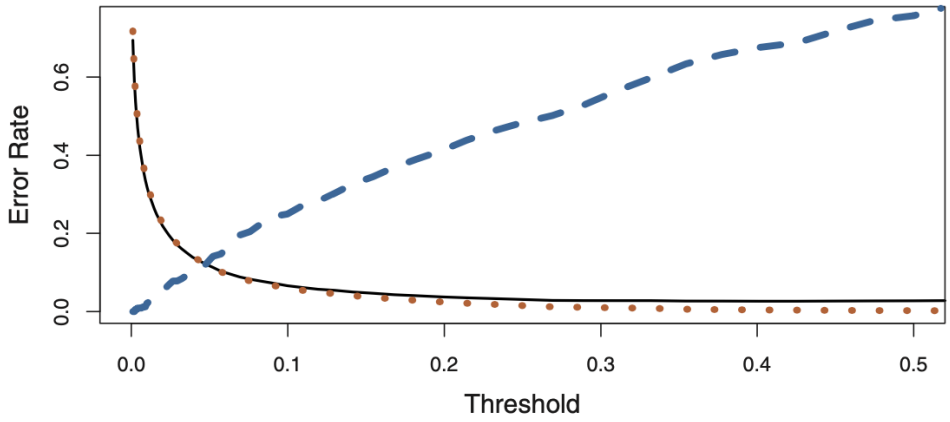
- **Sensitivity**: the percentage of true defaulters that are identified. In this case it's 24.3%
- **Specificity**: the percentage of non-defaulters that are correctly identified. In this case it's 99.8%.

The reason for such low sensitivity: because LDA is trying to approximate the Bayes classifier, which has the lowest total error rate out of all classifiers. Which means the Bayes classifier will yield the smallest possible total number of misclassified observations, regardless of which class the error comes.

The Bayes classifier works by assigning an observation to the class which maximizes the posterior probability $p_k(X)$, in the two-class example, we have:

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5.$$

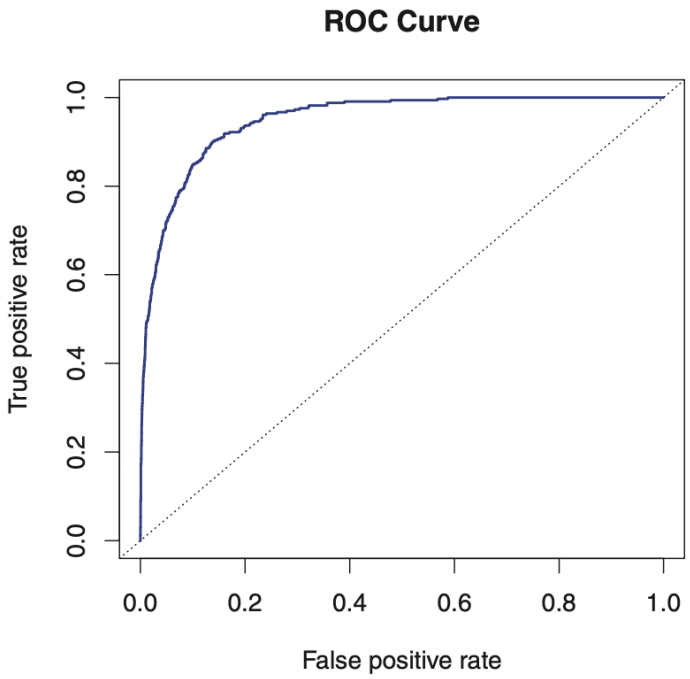
If we hope to solve the low sensitivity issue, we can change the threshold of the posterior probability into a small number, like 0.2, or 0.3. There relationships can be found as below:



The way to determine the best threshold value is based on domain knowledge, such as detailed information about the costs associated with default.

ROC Curve

ROC (Receiver Operating Characteristics) curve is a popular graphic for simultaneously displaying the two types of errors for all possible thresholds.



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.

4.4.4 Quadratic Discriminant Analysis

In LDA, we assume that the observations are drawn from multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes.

Here, the Quadratic Discriminant Analysis (QDA) assumes that observations from each class are drawn from a Gaussian distribution, but each class has its own covariance matrix. In other words, it assumes that an observation from the kth class is of the form $XN($
mu_k,
sum_k), where $\sum_k$ is a covariance matrix for the kth class. Under this assumption, we can get the Bayes classifier assigns an observation $X = x$ to class for which:

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

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

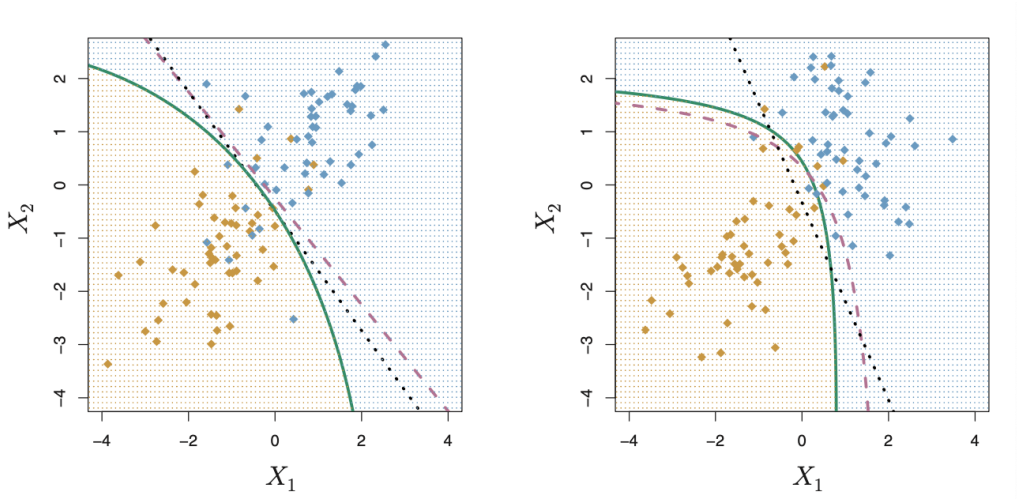
Trade-off between LDA and QDA

The rationale of this is based on the trade-off between bias and variance. When there are p predictors, estimating a covariance matrix requires estimating p(p+1)/2 parameters.

QDA estimates a separate covariance matrix for each class, for a total of $Kp(p+1)/2$ parameters.

For LDA, it is assumed that the K classes share a common covariance matrix, which means there are Kp linear coefficients to estimate. Thus LDA is a much less flexible classifier than QDA, and thus has a lower variance.

However, the assumption LDA holds may not be correct, or more badly than that of QDA, which would lead to higher bias.



- Left: LDA is more accurate, because QDA suffers from higher variance without a corresponding decrease in bias.
- Right: QDA is more accurate, because the Bayes decision boundary is quadratic.