11
Linear and Quadratic Discriminant Analysis, Logistic Regression, and Partial Least Squares Regression

In this chapter, we review, for the most part, linear methods for classification. The only exception is quadratic discriminant analysis, a straightforward generalization of a linear technique. These methods are best known for their simplicity. A linear decision boundary is easy to understand and visualize, even in many dimensions. An example of such a boundary is shown in Figure 11.1 for Fisher iris data.

Because of their high interpretability, linear methods are often the first choice for data analysis. They can be the only choice if the analyst seeks to discover linear relationships between variables and classes. If the analysis goal is maximization of the predictive power and the data do not have a linear structure, nonparametric nonlinear methods should be favored over simple interpretable techniques.

Linear discriminant analysis (LDA), also known as Fisher discriminant, has been a very popular technique in particle and astrophysics. Quadratic discriminant analysis (QDA) is its closest cousin.

11.1
Discriminant Analysis

Suppose we observe a sample drawn from a multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The data are D-dimensional, and vectors, unless otherwise noted, are column-oriented. The multivariate density is then

$$P(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right], \tag{11.1}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

Suppose we observe a sample of data drawn from two classes, each described by a multivariate normal density

$$P(\boldsymbol{x}|k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)\right] \tag{11.2}$$

Statistical Analysis Techniques in Particle Physics, First Edition. Ilya Narsky and Frank C. Porter.

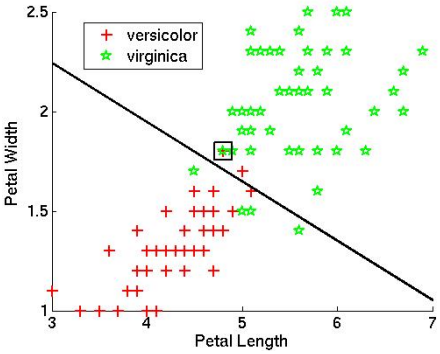

Figure 11.1 Class boundary obtained by linear discriminant analysis for Fisher iris data. The square covers one observation of class versicolor and two observations of class virginica.

for classes $k = 1, 2$. Recall that Bayes rule gives

$$P(k|\boldsymbol{x}) = \frac{\pi_k P(\boldsymbol{x}|k)}{P(\boldsymbol{x})} \tag{11.3}$$

for the posterior probability $P(k|\boldsymbol{x})$ of observing an instance of class k at point $\boldsymbol{x}$. The unconditional probability $P(\boldsymbol{x})$ in the denominator does not depend on k. The prior class probability π_k was introduced in Chapter 9; we discuss its role in the discriminant analysis below.

Let us take a natural logarithm of the posterior odds:

$$\begin{aligned}\log \frac{P(k=1|\boldsymbol{x})}{P(k=2|\boldsymbol{x})} = \log \frac{\pi_1}{\pi_2} &- \frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \\ &+ \boldsymbol{x}^\mathsf{T} \left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) \\ &- \frac{1}{2}\boldsymbol{x}^\mathsf{T} \left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right) \boldsymbol{x} \\ &- \frac{1}{2} \left(\boldsymbol{\mu}_1^\mathsf{T}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\mathsf{T}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right) .\end{aligned} \tag{11.4}$$

The hyperplane separating the two classes is obtained by equating this log-ratio to zero. This is a quadratic function of $\boldsymbol{x}$, hence *quadratic discriminant analysis*. If the two classes have the same covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, the quadratic term disappears and we obtain *linear discriminant analysis*.

Fisher (1936) originally derived discriminant analysis in a different fashion. He searched for a direction $\boldsymbol{q}$ maximizing separation between two classes,

$$S(\boldsymbol{q}) = \frac{\left[\boldsymbol{q}^\mathsf{T}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\right]^2}{\boldsymbol{q}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{q}} . \tag{11.5}$$

This separation is maximized at $\boldsymbol{q} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The terms not depending on $\boldsymbol{x}$ in (11.4) do not change the orientation of the hyperplane separating the two distributions – they only shift the boundary closer to one class and further away from the other. The formulation by Fisher is therefore equivalent to (11.4) for LDA.

To use this formalism for $K > 2$ classes, choose one class, for example the last one, for normalization. Compute posterior odds $\log[P(k|\boldsymbol{x})/P(K|\boldsymbol{x})]$ for $k = 1, \ldots, K-1$. The logarithm of the posterior odds is additive, that is,

$$\log[P(i|\boldsymbol{x})/P(j|\boldsymbol{x})] = \log[P(i|\boldsymbol{x})/P(K|\boldsymbol{x})] - \log[P(j|\boldsymbol{x})/P(K|\boldsymbol{x})]$$

for classes i and j. The computed $K-1$ log-ratios give complete information about the hyperplanes of separation. If we need to compute the posterior probabilities, we require that $\sum_{k=1}^{K} P(k|\boldsymbol{x}) = 1$ and obtain estimates of $P(k|\boldsymbol{x})$ from the log-ratios. The same trick is used in other multiclass models such as multinomial logistic regression. For prediction on new data, the class label is assigned by choosing the class with the largest posterior probability.

In this formulation, the class prior probabilities merely shift the boundaries between the classes without changing their orientations (for LDA) or their shapes (for QDA). They are not used to estimate the class means or covariance matrices; hence, they can be applied after training. Alternatively, we could ignore the prior probabilities and classify observations by imposing thresholds on the computed log-ratios. These thresholds would be optimized using some physics-driven criteria. Physicists often follow the second approach.

As discussed in Chapter 9, classifying into the class with the largest posterior probability $P(\gamma|\boldsymbol{x})$ minimizes the classification error. If the posterior probabilities are accurately modeled, this classifier is optimal. If classes indeed have multivariate normal densities, QDA is the optimal classifier. If classes indeed have multivariate normal densities with equal covariance matrices, LDA is the optimal classifier. Most usually, we need to estimate the covariance matrices empirically.

LDA is seemingly simple, but this simplicity may be deceiving. Subtleties in LDA implementation can change its result dramatically. Let us review them now.

11.1.1 Estimating the Covariance Matrix

Under the LDA assumptions, classes have equal covariance matrices and different means. Take the training data with known class labels. Let $\mathbf{M}$ be an $N \times K$ class membership matrix for N observations and K classes: $m_{nk} = 1$ if observation n is from class k and 0 otherwise. First, estimate the mean for each class in turn,

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{n=1}^{N} m_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} m_{nk}} . \tag{11.6}$$

Then compute the pooled-in covariance matrix. For example, use a maximum likelihood estimate,

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} m_{nk}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)^{\mathsf{T}} . \tag{11.7}$$

Vectors $\boldsymbol{x}_n$ and $\hat{\boldsymbol{\mu}}_k$ are $D \times 1$ (column-oriented), and $(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)^{\mathsf{T}}$ is therefore a symmetric $D \times D$ matrix. This maximal likelihood estimator is biased. To remove

the bias, apply a small correction:

$$\hat{\boldsymbol{\Sigma}} = \frac{N}{N-K}\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}\,. \tag{11.8}$$

Elementary statistics textbooks derive a similar correction for a univariate normal distribution and include a formula for an unbiased estimate, $S^2 = \sum_{n=1}^{N}(x_n - \bar{x})^2/(N-1)$, of the variance, σ^2. The statistic $(N-1)S^2/\sigma^2$ is distributed as χ^2 with $N-1$ degrees of freedom. We use $N-K$ instead of $N-1$ because we have K classes. Think of it as losing one degree of freedom per linear constraint. In this case, there are K linear constraints for K class means.

In physics analysis, datasets are usually large, $N \gg K$. For unweighted data, this correction can be safely neglected. For weighted data, the problem is a bit more involved. The weighted class means are given by

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{n=1}^{N} m_{nk} w_n \boldsymbol{x}_n}{\sum_{n=1}^{N} m_{nk} w_n}\,. \tag{11.9}$$

The maximum likelihood estimate (11.7) generalizes to

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \sum_{k=1}^{K}\sum_{n=1}^{N} m_{nk} w_n (\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_k)^{\mathsf{T}}\,. \tag{11.10}$$

Above, we assume that the weights are normalized to sum to one: $\sum_{n=1}^{N} w_n = 1$. The unbiased estimate is then

$$\hat{\boldsymbol{\Sigma}} = \frac{\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}}{1 - \sum_{k=1}^{K} \frac{W_k^{(2)}}{W_k}}\,, \tag{11.11}$$

where $W_k = \sum_{n=1}^{N} m_{nk} w_n$ is the sum of weights in class k and $W_k^{(2)} = \sum_{n=1}^{N} m_{nk} w_n^2$ is the sum of squared weights in class k. For class-free data $K = 1$, this simplifies to $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}/(1 - \sum_{n=1}^{N} w_n^2)$. If all weights are set to $1/N$, (11.11) simplifies to (11.8). In this case, the corrective term $\sum_{n=1}^{N} w_n^2$ attains minimum, and the denominator in (11.11) is close to 1. If the weights are highly nonuniform, the denominator in (11.11) can get close to zero.

For LDA with two classes, this bias correction is, for the most part, irrelevant. Multiplying all elements of the covariance matrix by factor a is equivalent to multiplying $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ by $1/a$. This multiplication does not change the orientation of the hyperplane separating the two classes, but it does change the posterior class probabilities at point $\boldsymbol{x}$. Instead of using the predicted posterior probabilities directly, physicists often inspect the ROC curve and select a threshold on classification scores (in this case, posterior probabilities) by optimizing some function of true positive and false positive rates. If we fix the false positive rate, multiply the log-ratio at any point $\boldsymbol{x}$ by the same factor and measure the true positive rate, we will obtain the same value as we would without multiplication.

Unfortunately, this safety mechanism fails for QDA, multiclass LDA, and even LDA with two classes if the covariance matrix is estimated as a weighted combination of the individual covariance matrices, as described in Section 11.1.3. We refrain from recommending the unbiased estimate over the maximum likelihood estimate or the other way around. We merely point out this issue. If you work with highly nonuniform weights, you should investigate the stability of your analysis procedure with respect to weighting.

11.1.2 Verifying Discriminant Analysis Assumptions

The key assumptions for discriminant analysis are multivariate normality (for QDA and LDA) and equality of the class covariance matrices (for LDA). You can verify these assumptions numerically.

Two popular tests of normality proposed in Mardia (1970) are based on multivariate skewness and kurtosis. The sample kurtosis is readily expressed in matrix notation

$$\hat{\kappa} = \frac{1}{N}\sum_{n=1}^{N}\left[(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}})\right]^2 , \tag{11.12}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the usual estimates of the mean and covariance matrix. Asymptotically, $\hat{\kappa}$ has a normal distribution with mean $D(D+2)$ and variance $8D(D+2)/N$ for the sample size N and dimensionality D. A large observed value indicates a distribution with tails heavier than normal, and a small value points to a distribution with tails lighter than normal.

A less rigorous but more instructive procedure is to inspect a *quantile-quantile (QQ) plot* of the squared Mahalanobis distance (Healy, 1968). If $\boldsymbol{X}$ is a random vector drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, its squared Mahalanobis distance $(\boldsymbol{X}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$ has a χ^2 distribution with D degrees of freedom. We can plot quantiles of the observed Mahalanobis distance versus quantiles of the χ^2_D distribution. A departure from a straight line would indicate the lack of normality, and extreme points would be considered as candidates for outliers. In practice we know neither $\boldsymbol{\mu}$ nor $\boldsymbol{\Sigma}$ and must substitute their estimates. As soon as we do, we, strictly speaking, can no longer use the χ^2_D distribution, although it remains a reasonable approximation for large N. No statistical test is associated with this approach, but visual inspection often proves fruitful.

Equality of the class covariance matrices can be verified by a Bartlett multivariate test described in popular textbooks such as Andersen (2003). Formally, we test hypothesis $H_0\colon \boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_K$ against $H_1\colon$ at least two $\boldsymbol{\Sigma}'$s are different. The test statistic,

$$-2\log V = (N-K)\log|\hat{\boldsymbol{\Sigma}}| - \sum_{k=1}^{K}(n_k - 1)\log|\hat{\boldsymbol{\Sigma}}_k| , \tag{11.13}$$

resembles a log-likelihood ratio for the pooled-in unbiased estimate $\hat{\boldsymbol{\Sigma}}$ and unbiased class estimates $\hat{\boldsymbol{\Sigma}}_k$. Here, n_k is the number of observations in class k. This formula would need to be modified for weighted data. When the sizes of all classes are comparable and large, $-2\log V$ can be approximated by a χ^2 distribution with $(K-1)D(D+1)/2$ degrees of freedom (see, for example, Box, 1949). For small samples, the exact distribution can be found in Gupta and Tang (1984). H_0 should be rejected if the observed value of $-2\log V$ is large. Intuitively, $-2\log V$ measures the lack of uniformity of the covariance matrices across the classes. The pooled-in estimate is simply a sum over class estimates $(N-K)\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K}(n_k-1)\hat{\boldsymbol{\Sigma}}_k$. If the sum of several positive numbers is fixed, their product (equivalently, the sum of their logs) is maximal when the numbers are equal. This test is based on the same idea. The Bartlett test is sensitive to outliers and should not be used in their presence.

The tests described here are mostly of theoretical value. Practitioners often apply discriminant analysis when its assumptions do not hold. The ultimate test of any classification model is its performance. If discriminant analysis gives a satisfactory predictive power for nonnormal samples, don't let the rigor of theory stand in your way. Likewise, you can verify that QDA improves over LDA by comparing the accuracies of the two models using one of the techniques reviewed in Chapter 10.

11.1.3 Applying LDA When LDA Assumptions Are Invalid

Under the LDA assumptions, all classes have multivariate normal distributions with different means and the same covariance matrix. The maximum likelihood estimate (11.10) is equal to the weighted average of the covariance matrix estimates per class:

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \sum_{k=1}^{K} W_k \hat{\boldsymbol{\Sigma}}_k \,. \tag{11.14}$$

Above, $W_k = \sum_{n=1}^{N} m_{nk} w_n$ is the sum of weights in class k. As usual, we take $\sum_{n=1}^{N} w_n = 1$.

In practice, physicists apply LDA when none of the LDA conditions holds. The class densities are not normal and the covariance matrices are not equal. In these circumstances, you can still apply LDA and obtain some separation between the classes. But there is no theoretical justification for the pooled-in covariance matrix estimate (11.14). It is tempting to see how far we can get by experimenting with the covariance matrix estimate.

Let us illustrate this problem on a hypothetical example. Suppose we have two classes with means

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix} \quad \boldsymbol{\mu}_2 = \begin{pmatrix} -1/2 \\ 0 \end{pmatrix} \tag{11.15}$$

and covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{11.16}$$

The two covariance matrices are inverse to each other, up to some constant. If we set the covariance matrix for LDA to $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$, the predicted line of optimal separation is orthogonal to the first coordinate axis and the optimal classification is given by x_1. If we set the covariance matrix for LDA to $\boldsymbol{\Sigma}_1$, the optimal classification is $2x_1 - x_2$. If we set the covariance matrix for LDA to $\boldsymbol{\Sigma}_2$, the optimal classification is $2x_1 + x_2$. If we had to estimate the pooled-in covariance matrix on a training set, we would face the same problem. If classes 1 and 2 were represented equally in the training data, we would obtain $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$. If the training set were composed mostly of observations of class 1, we would obtain $\boldsymbol{\Sigma}_1$. Similarly for $\boldsymbol{\Sigma}_2$.

Let us plot ROC curves for the three pooled-in matrix estimates. As explained in the previous chapter, a ROC curve is a plot of true positive rate (TPR) versus false positive rate (FPR), or accepted signal versus accepted background. Take class 1 to be signal and class 2 to be background. The three curves are shown in Figure 11.2. Your choice of the optimal curve (and therefore the optimal covariance matrix) would be defined by the specifics of your analysis. If you were mostly concerned with background suppression, you would be interested in the lower left corner of the plot and choose $\boldsymbol{\Sigma}_2$ as your estimate. If you goal were to retain as much signal as possible at a modest background rejection rate, you would focus on the upper right corner of the plot and choose $\boldsymbol{\Sigma}_1$. If you wanted the best overall quality of separation measured by the area under the ROC curve, you would choose $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)/2$. It is your analysis, so take your pick!

If we took this logic to the extreme, we could search for the best a in the linear combination $\hat{\boldsymbol{\Sigma}} = a\hat{\boldsymbol{\Sigma}}_1 + (1-a)\hat{\boldsymbol{\Sigma}}_2$ by minimizing some criterion, perhaps FPR at fixed TPR. If you engage in such optimization, you should ask yourself if LDA is the right tool. At this point, you might want to give up the beloved linearity and switch to a more flexible technique such as QDA. In this example, QDA beats LDA at any FPR, no matter what covariance matrix estimate you choose.

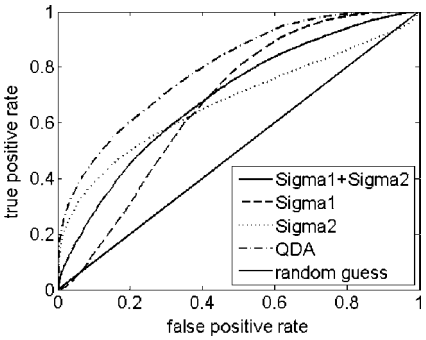

Figure 11.2 ROC curves for LDA with three estimates of the covariance matrix and QDA.