CSCI 6521: Advanced Machine Learning I

Chapter #2: LDA & QDA

(Reference: Chapter 4 of [1])

Objective: Here, we will extend our previous models for classification problems. Approaches are similar to the regression problem; however, the outputs are broken into discrete ranges and labeled categorically.

Here, we discuss the classification problem and focus on linear methods for classification. Since our predictor $G(x)$ takes values in a discrete set G, we can always divide the input space into a collection of regions labeled according to the classification. The boundaries of these regions can be rough or smooth, depending on the prediction function. For an important class of procedures, these *decision boundaries* are linear; this is what we will mean by linear methods for classification.

There are several different ways in which linear decision boundaries can be found. One example can be to fit linear regression models to the class indicator variables and classify the sample input based on the largest fit. Suppose there are K classes for convenience labeled 1, 2, …, K, and the fitted linear model for the k^{th} indicator response variable is $\hat{f}_k(x) = \hat{\beta}_{k0} + \hat{\beta}_k^T x$. The decision boundary between class k and ℓ is that set of points for which $\hat{f}_k(x) = \hat{f}_l(x)$, this is, $\{x : (\hat{\beta}_{k0} - \hat{\beta}_{l0}) + (\hat{\beta}_k - \hat{\beta}_l)^T x = 0\}$, an affine set or hyperplane[1]. Since the same is true for any pair of classes, the input space is divided into regions of constant classification, with piecewise hyperplanar decision boundaries. This regression approach is a member of a class of methods that model *discriminant functions* $\delta_k(x)$ for each class, and then classify x to the class with the largest value for its discriminant function. Methods that model the posterior probabilities $\Pr(G = k \mid X = x)$ are also in this class. Clearly, if either the $\delta_k(x)$ or $\Pr(G = k|X = x)$ are linear in x, then the decision boundaries will be linear.

Actually, all we require is that some monotone transformation of δ_k or $\Pr(G = k \mid X = x)$ be linear for the decision boundaries to be linear. For

[1] Strictly speaking, a hyperplane passes through the origin, while an affine set need not. We sometimes ignore the distinction and refer in general to hyperplanes.

example, if there are two classes, a popular model for the posterior probabilities is

$$\begin{aligned} \Pr(G = 1|X = x) &= \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}, \\ \Pr(G = 2|X = x) &= \frac{1}{1 + \exp(\beta_0 + \beta^T x)}. \end{aligned} \tag{4.1}$$

Here the monotone transformation is the ***logit*** transformation: log $[p/(1-p)]$, and in fact, we see that

$$\log \frac{\Pr(G = 1 \mid X = x)}{\Pr(G = 2 \mid X = x)} = \beta_0 + \beta^T x \tag{4.2}$$

Note [Starts]:

The ***logit*** of a probability p is given by $logit(p) = \log\left(\frac{p}{1-p}\right)$. The term $\left(\frac{p}{1-p}\right)$ is called the odds of probability p; thus ***logit*** and ***log-odds*** are synonyms. So, we can write, $logit(p) = log\ (odds(p)) = \log\left(\frac{p}{1-p}\right)$.

Note [Ends]

The decision boundary is the set of points for which the log-odds are zero, and this is a hyperplane defined by $\{ x \mid \beta_0 + \beta^T x = 0\}$. We discuss two very popular but different methods that result in linear *log-odds* or *logits*: linear discriminant analysis (LDA) and linear logistic regression. Although they differ in their derivation, the essential difference between them is in the way the linear function is fit to the training data.

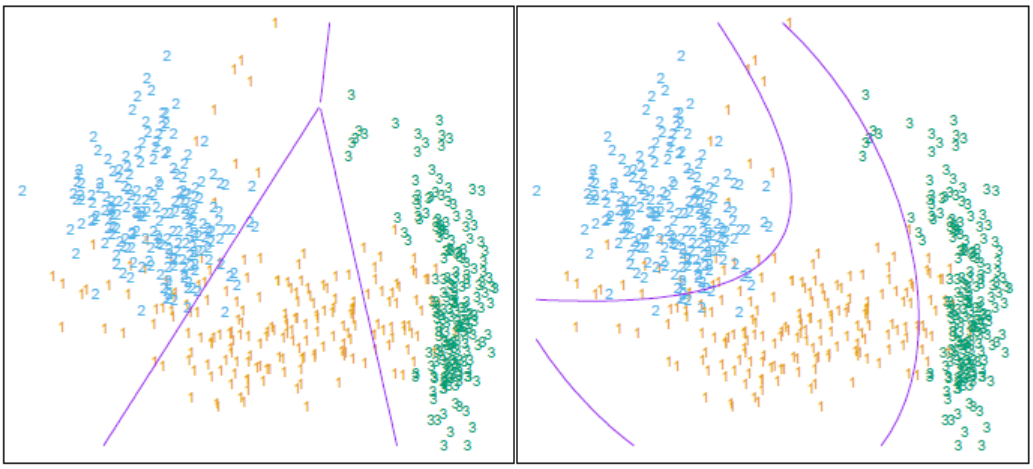

Figure 4.1 [1]: *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space* X_1, X_2, X_1X_2, X_1^2, X_2^2. *Linear inequalities in this space are quadratic inequalities in the original space.*

While this entire chapter is devoted to linear decision boundaries, there is considerable scope for generalization. For example, we can expand our variable set $X_1, \ldots, X_p$ by including their squares and cross-products X_1^2, X_2^2, …, X_1X_2, …, thereby adding $p(p+1)/2$ additional variables (of power 2, for example, for existing 3 variables, $\{x_1, x_2, x_3\}$, we can have 6 additional variables: $\{x_1^2, x_2^2, x_3^2, x_1x_2, x_2x_3, x_2x_3\}$). Linear functions in the augmented space map down to quadratic functions in the original space—hence linear decision boundaries to quadratic decision boundaries. Figure 4.1 illustrates the idea. The data are the same: the left plot uses linear decision boundaries in the two-dimensional space shown, while the right plot uses linear decision boundaries in the augmented five-dimensional space described above. This approach can be used with any basis transformation $h(X)$ where $h: \Re^p \rightarrow \Re^q$ with q > p.

4.2 Linear Regression of an Indicator Matrix

Here each of the response categories is coded via an indicator variable. Thus if G has K classes, there will be K such indicators Y_k, $k = 1, \ldots, K$, with $Y_k = 1$ if $G = k$, else 0. These are collected together in a vector $Y = (Y_1, \ldots, Y_K)$, and the N training instances of these form an $N \times K$ *indicator response matrix* Y. Y is a matrix of 0's and 1's, with each row having a single 1.

For example, we can think of representing our cancer classification problem as following in this context:

$$\begin{array}{cccc} \mathbf{X}_0 & \mathbf{X}_1 & \mathbf{X}_2 & \\ \begin{pmatrix} 1 & 0.9 & 1.9 \\ 1 & 3.4 & 3.0 \\ 1 & \vdots & \vdots \\ 1 & 2.45 & 3.67 \end{pmatrix}_{N\times(p+1)} & \begin{matrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 \end{matrix} & \begin{matrix} \mathbf{Y}_1=\mathbf{B} & \mathbf{Y}_2=\mathbf{M} \end{matrix} \\ & \begin{pmatrix} \beta_{1,0} & \beta_{2,0} \\ \beta_{1,1} & \beta_{2,1} \\ \vdots & \vdots \\ \beta_{k=1,p} & \beta_{k=2,p} \end{pmatrix}_{(p+1)\times K} & = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}_{N\times K} \\ \mathbf{X}_{N\times(p+1)} & \boldsymbol{\beta}_{(p+1)\times K} & = \mathbf{Y}_{N\times K} \end{array}$$

More precisely, for the two classes we can write:

$$Y_1 = X_0\beta_{1,0} + \sum_{j=1}^{p} X_j\,\beta_{1,j} + \varepsilon_1 \text{ and } Y_2 = X_0\beta_{2,0} + \sum_{j=1}^{p} X_j\,\beta_{2,j} + \varepsilon_2$$

And in general:

$$Y_k = X_0\beta_{k,0} + \sum_{j=1}^{p} X_j\,\beta_{k,j} + \varepsilon_k$$

Here, card (K) = 2 {*Benign* (**B**), *Malignant* (**M**)} and p = 2 {*Size* (X_1), *Age* (X_2)} are assumed. X_0 is the intercept column as we have experienced before.

We fit a linear regression model to each of the columns of **Y** simultaneously, and the fit is given by

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \tag{4.3}$$

Note that we have a coefficient vector for each response column $\boldsymbol{y}_k$, and hence a $(p+1)\times K$ coefficient matrix $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$. Here X is the model matrix with $p+1$ columns corresponding to the p inputs and a leading column of 1's for the intercept.

A new observation with input x is classified as follows:

- compute the fitted output $\hat{f}(x)^T = (1, x^T)\hat{\boldsymbol{\beta}}$, a K vector;
- identify the largest component and classify accordingly:

$$\hat{G}(x) = \text{argmax}_{k\in\mathcal{G}}\hat{f}_k(x) \tag{4.4}$$

We can view the regression here as an estimate of conditional expectations. For the random variable Y_k, $E(Y_k|X = x) = \Pr(G = k|X = x)$, so conditional expectation of each of the Y_k seems a sensible goal.

It is quite straightforward to verify that $\sum_{k\in\mathcal{G}} \hat{f}_k(x) = 1$ for any x, as long as there is an intercept in the model (column of 1's in X). However, the $\hat{f}_k(x)$ can be negative or greater than 1, and typically some are. This is a consequence of the rigid nature of linear regression, especially if we make predictions outside the hull of the training data. These violations in themselves do not guarantee that this approach will not work, and in fact, on many problems, it gives similar results to more standard linear methods for classification. If we allow linear regression onto basis expansions $h(X)$ of the inputs, this approach can lead to consistent estimates of the probabilities. As the size of the training set N grows bigger, we adaptively include more basis elements so that linear regression onto these basis functions approaches conditional expectation (more in Chapter 5 of [1]).

A more simplistic viewpoint is to construct *targets* t_k for each class, where t_k is the k^{th} column of the $K \times K$ identity matrix. Our prediction problem is to try and reproduce the appropriate target for observation. With the same coding as before, the response vector y_i (i^{th} row of $\mathbf{Y}$) for observation i has the value $y_i = t_k$ if $g_i = k$. We might then fit the linear model by least squares:

$$\min_{\beta} \sum_{i=1}^{N} \| y_i - [(1, x_i^T)\beta]^T \|^2 \qquad (4.5)$$

The criterion is a sum-of-squared Euclidean distances of the fitted vectors from their targets. A new observation is classified by computing its fitted vector $\hat{f}(x)$ and classifying to the closest target:

$$\hat{G}(x) = \arg\min_k \| \hat{f}(x) - t_k \|^2 \qquad (4.6)$$

This is the same as the previous approach:

- The sum-of-squared-norm criterion is exactly the criterion for multiple response linear regression, just viewed slightly differently. Since a squared norm is itself a sum of squares, the components decouple and can be rearranged as a separate linear model for each element. Note that this is only possible because there is nothing in the model that binds the different responses together.

- The closest target classification rule (4.6) is easily seen to be exactly the same as the maximum fitted component criterion (4.4) but does require that the sum of the fitted value to 1.

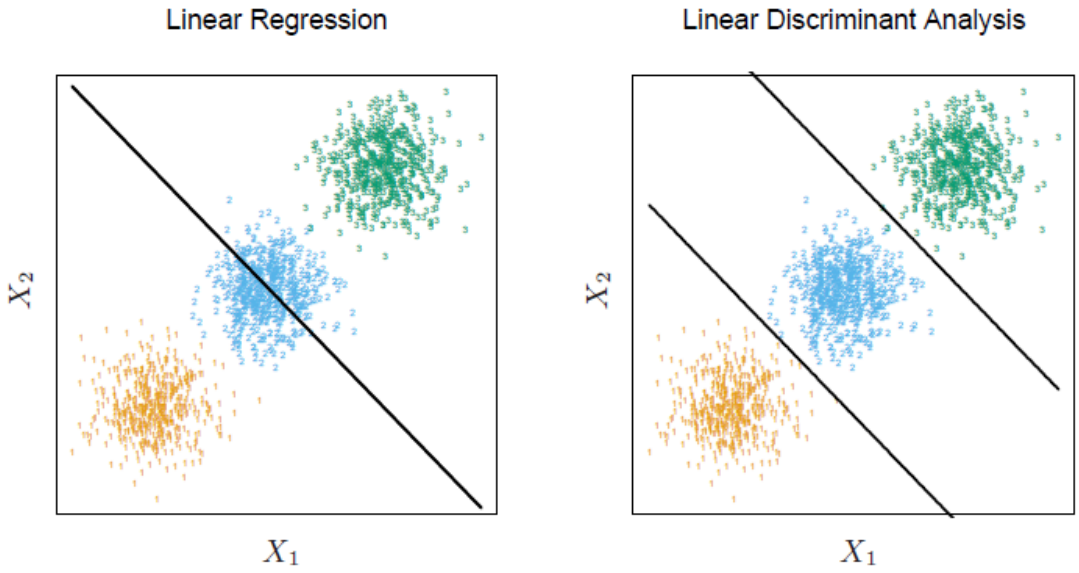

Figure 4.2 [1]: The data come from three classes in $\Re^2$ and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis (*LDA*). The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates).

There is a serious problem with the regression approach when the number of classes $K \geq 3$, especially prevalent when K is large. Because of the rigid nature of the regression model, classes can be masked by others. Figure 4.2 illustrates an extreme situation when $K = 3$. The three classes are perfectly separated by linear decision boundaries, yet linear regression misses the middle class completely.

For this simple example, a quadratic rather than linear fit (for the middle class at least) would solve the problem. However, it can be seen that if there were four rather than three classes lined up like this, a quadratic would not come down fast enough, and a cubic would be needed as well. A loose but general rule is that if $K \geq 3$ classes are lined up, polynomial terms up to degree $(K - 1)$ might be needed to resolve them.

However, for large K and small p, such maskings naturally occur. As a more realistic illustration, Figure 4.4 is a projection of the training data for a vowel recognition problem onto an informative two-dimensional subspace. There are $K = 11$ classes in $p = 10$ dimensions. This is a difficult classification problem, and the best methods achieve around 40% errors on the test data. The main point here is summarized in Table 4.1; linear regression has an error

rate of 67%, while a close relative, linear discriminant analysis, has an error rate of 56%. It seems that masking has hurt in this case. While all the other methods in this chapter are based on linear functions of x as well, they use them in such a way that avoids this masking problem.

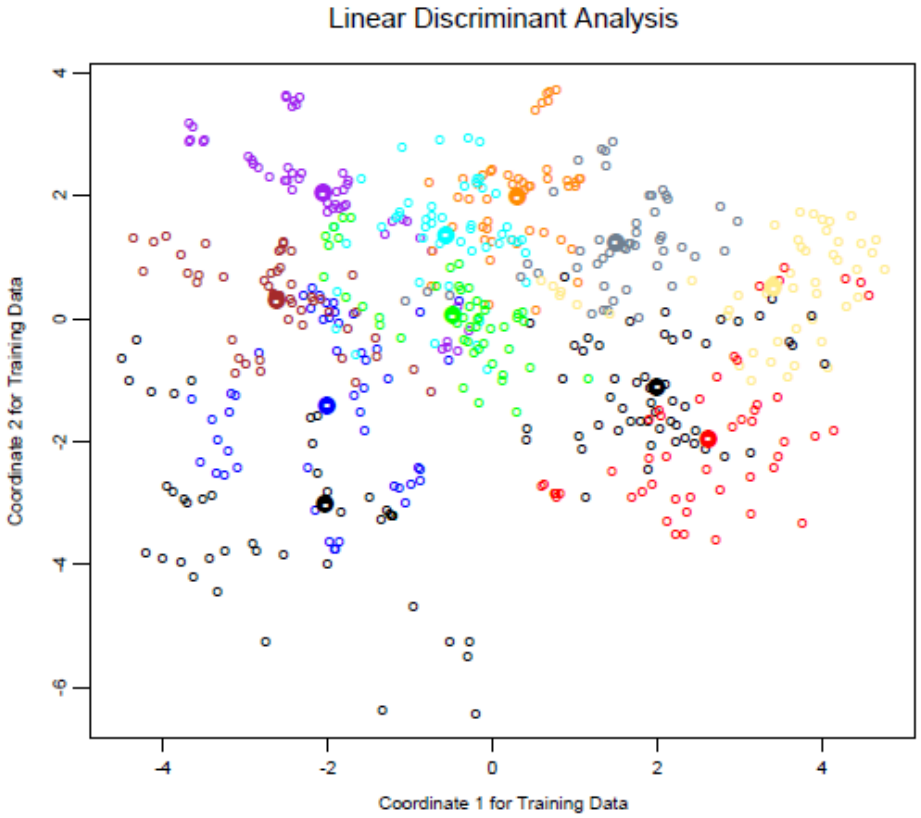

Figure 4.4 [1]: A two-dimensional plot of the vowel training data. There are eleven classes with $X \in \Re^{10}$, and this is the best view in terms of an LDA model (Section 4.3.3 in [1]). The heavy circles are the projected mean vectors for each class. The class overlap is considerable.

Table 4.1: Training and test error rates using a variety of linear techniques on the vowel data. There are eleven classes in ten dimensions, of which three account for 90% of the variance. We see that linear regression is hurt by masking, increasing the test and training error by over 10%.

Technique	Error Rates	
	Training	Test
Linear regression	0.48	0.67
Linear discriminant analysis	0.32	0.56
Quadratic discriminant analysis	0.01	0.53
Logistic regression	0.22	0.51

4.3 Linear Discriminant Analysis

Decision theory for classification (Section 2.4 in [1]) tells us that we need to know the class posteriors Pr($G|X$) for optimal classification. Suppose $f_k(x)$ is the class-conditional density of X in class $G = k$, and let π_k be the prior probability of class k, with $\sum_{k=1}^{K} \pi_k = 1$. A simple application of Bayes theorem gives us:

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{4.7}$$

We see that in terms of ability to classify, having the $f_k(x)$ is almost equivalent to having the quantity Pr($G = k|X = x$). Many techniques are based on models for the class densities, such as linear and quadratic discriminant analysis use Gaussian densities, and more flexible mixtures of Gaussians allow for nonlinear decision boundaries.

Suppose that we model each class density as multivariate Gaussian

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\mathbf{\Sigma}_k\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \tag{4.8}$$

Linear discriminant analysis (LDA) arises in the special case when we assume that the classes have a common covariance matrix $\Sigma_k = \Sigma \,\forall k$. In comparing two classes k and ℓ, it is sufficient to look at the log-ratio, and we see that

$$\log \frac{\Pr(G = k \mid X = x)}{\Pr(G = l \mid X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \frac{\pi_k}{\pi_l} + \log f_k(x) - \log f_l(x)$$

$$= \log \frac{\pi_k}{\pi_l} + \left[-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right]$$
$$-\left[-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma_l|) - \frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)\right]$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left[(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - (x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l) + \log(|\Sigma_k|) - \log(|\Sigma_l|)\right] \quad \textbf{(A0)}$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}[(x^T \Sigma_k^{-1} x - \mu_k^T \Sigma_k^{-1} x - x^T \Sigma_k^{-1} \mu_k + \mu_k^T \Sigma_k^{-1} \mu_k)$$

$$-(x^T\Sigma_l^{-1}x-\mu_l^T\Sigma_l^{-1}x-x^T\Sigma_l^{-1}\mu_l+\mu_l^T\Sigma_l^{-1}\mu_l)+\log(|\Sigma_k|)-\log(|\Sigma_l|)]$$

$$= \log\frac{\pi_k}{\pi_l}-\frac{1}{2}[x^T(\Sigma_k^{-1}-\Sigma_l^{-1})x-(\mu_k^T\Sigma_k^{-1}x+x^T\Sigma_k^{-1}\mu_k)+(\mu_l^T\Sigma_l^{-1}x+x^T\Sigma_l^{-1}\mu_l)$$
$$+(\mu_k^T\Sigma_k^{-1}\mu_k-\mu_l^T\Sigma_l^{-1}\mu_l)+\log(|\Sigma_k|)-\log(|\Sigma_l|)]$$

$$= \log\frac{\pi_k}{\pi_l}-\frac{1}{2}[x^T(\Sigma_k^{-1}-\Sigma_l^{-1})x-2(\mu_k^T\Sigma_k^{-1}x)+2(\mu_l^T\Sigma_l^{-1}x)+(\mu_k^T\Sigma_k^{-1}\mu_k-\mu_l^T\Sigma_l^{-1}\mu_l)$$
$$+\log(|\Sigma_k|)-\log(|\Sigma_l|)]$$

$$\left[\because x^T\Sigma^{-1}\mu=(\Sigma^{-1}\mu)^T x=\mu^T(\Sigma^{-1})^T x=\mu^T\Sigma^{-1}x\right.$$
$$\left.\text{Note}:\because \Sigma \text{ is symmetric},\ \therefore \Sigma^T=\Sigma \text{ and } (\Sigma^{-1})^T=\Sigma^{-1}\right]$$

$$= -\frac{1}{2}[x^T(\Sigma_k^{-1}-\Sigma_l^{-1})x+2(\mu_l^T\Sigma_l^{-1}-\mu_k^T\Sigma_k^{-1})x+$$
$$+\left(\mu_k^T\Sigma_k^{-1}\mu_k-\mu_l^T\Sigma_l^{-1}\mu_l+log\frac{|\Sigma_k|}{|\Sigma_l|}-2\,log\frac{\pi_k}{\pi_l}\right)] \qquad \ldots \qquad \textbf{(A1)}$$

[It is important to note the format of the above Equation (A1). The RHS is a quadratic term for variable *x* => *Quadratic Discriminant Analysis (QDA)*]

$$=-\tfrac{1}{2}[2(\mu_l^T\Sigma^{-1}-\mu_k^T\Sigma^{-1})x+\left(\mu_k^T\Sigma^{-1}\mu_k-\mu_l^T\Sigma^{-1}\mu_l-2\,log\frac{\pi_k}{\pi_l}\right)]$$

$$\left[\text{Assuming},\Sigma_k=\Sigma_l=\Sigma\right.$$
$$\text{and the quadratic part disappears.}$$
$$\left.\text{Thus QDA gets converted into LDA}\right]$$

$$= \log\frac{\pi_k}{\pi_l}-(\mu_l^T\Sigma^{-1}-\mu_k^T\Sigma^{-1})x-\frac{1}{2}\left(\mu_k^T\Sigma^{-1}\mu_k-\mu_l^T\Sigma^{-1}\mu_l\right)$$

$$= \log\frac{\pi_k}{\pi_l}-x^T\Sigma^{-1}\mu_l+x^T\Sigma^{-1}\mu_k-\frac{1}{2}\left(\mu_k^T\Sigma^{-1}\mu_k-\mu_l^T\Sigma^{-1}\mu_l\right)$$
$$[\because x^T\Sigma^{-1}\mu=(\Sigma^{-1}\mu)^T x=\mu^T(\Sigma^{-1})^T x=\mu^T\Sigma^{-1}x]$$

$$= \log\frac{\pi_k}{\pi_l}-\frac{1}{2}\left(\mu_k^T\Sigma^{-1}\mu_k-\mu_l^T\Sigma^{-1}\mu_l\right)+x^T\Sigma^{-1}(\mu_k-\mu_l) \qquad \text{(A2)}$$

$$= \log\frac{\pi_k}{\pi_l}-\frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l)+x^T\Sigma^{-1}(\mu_k-\mu_l) \qquad \text{(4.9)}$$

[We can think of it, as: $a^2 - b^2 = a^2 - ba + ab - b^2 = (a+b)(a-b)$, the derivation is shown in detail in the note after the appearance of equation (4.10)]

$$= \left(x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \right) - \left(x^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l + \log \pi_l \right)$$ [see from (A2)]

$$= \delta_k(x) - \delta_l(x)$$

an equation linear in *x*.

Therefore, finally, we see:

$$\log \frac{\Pr(G = k \mid X = x)}{\Pr(G = l \mid X = x)} = \delta_k(x) - \delta_l(x) \qquad \textbf{(A3)}$$

At the decision boundary, we will have $\Pr(G = k \mid X = x) = \Pr(G = l \mid X = x)$, therefore, $\log \frac{\Pr(G = k \mid X = x)}{\Pr(G = l \mid X = x)} = \log(1) = 0$. From Equaton (A3), we can write:

$$\delta_k(x) - \delta_l(x) = 0$$
$$\text{Or, } \delta_k(x) = \delta_l(x)$$

This linear log-odds function implies that the decision boundary between classes k and ℓ—the set where $\Pr(G = k|X = x) = \Pr(G = \ell|X = x)$ – is linear in *x*; in *p* dimensions a hyperplane. This is, of course, true for any pair of classes, so all the decision boundaries are linear. If we divide $\Re^p$ into regions that are classified as class 1, class 2, etc., these regions will be separated by hyperplanes.

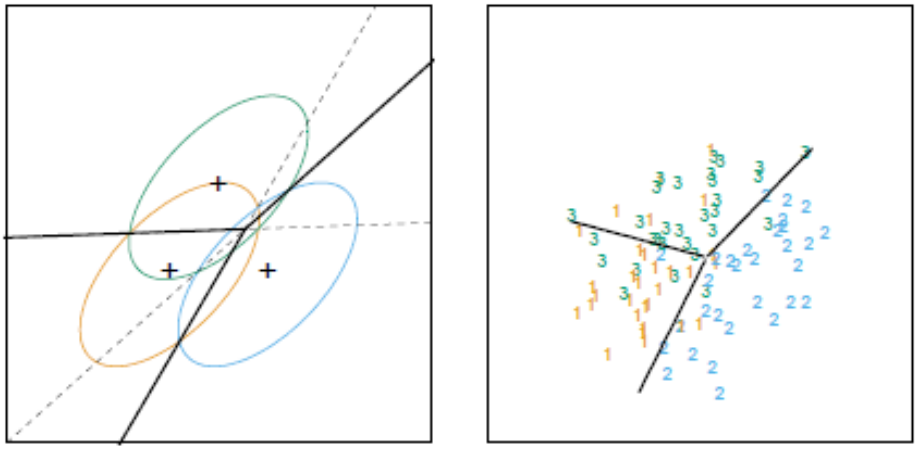

Figure 4.5 [1]: The left panel shows three Gaussian distributions, with the same covariance and different means. Included are the contours of constant density enclosing 95% of the probability in each case. The Bayes decision boundaries between each pair of classes are shown (broken straight lines), and the Bayes decision boundaries separating all three classes are the thicker solid lines (a subset of the former). On the right, we see a sample of 30 drawn from each Gaussian distribution, and the fitted LDA decision boundaries.

Figure 4.5 (left panel) shows an idealized example with three classes and p = 2. Here the data do arise from three Gaussian distributions with a common covariance matrix. We have included in the figure the contours corresponding to 95% highest probability density, as well as the class centroids. Notice that the decision boundaries are not the perpendicular bisectors of the line segments joining the centroids. This would be the case if the covariance $\mathbf{\Sigma}$ were spherical $\sigma^2\mathbf{I}$, and the class priors were equal. From (4.9) we see that the ***linear discriminant functions***

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k \qquad (4.10)$$

are an equivalent description of the decision rule, with $G(x) = \arg\max_k \delta_k(x)$.

Note:

How can we get (4.10) from (4.9)?

First, we need to show: $X^T AY = Y^T AX$, when A is a symmetric matrix (and square matrix).

$$
\begin{aligned}
X^T AY &= (X)^T (AY) \\
&= (AY)^T (X) && [\because P^T Q = Q^T P] \\
&= Y^T A^T X && [\because (PQ)^T = Q^T P^T] \\
&= Y^T AX && [\because A \text{ is symmetric, } A = A^T]
\end{aligned}
$$

So, we have shown:

$$X^T AY = Y^T AX \qquad (i)$$

Now we like to show: $(X+Y)^T A(X-Y) = X^T AX - Y^T AY$

$$
\begin{aligned}
(X+Y)^T A(X-Y) &= (X^T + Y^T)A(X-Y) && [\because (P+Q)^T = P^T + Q^T] \\
&= X^T AX + Y^T AX - X^T AY - Y^T AY \\
&= X^T AX + (Y^T AX - X^T AY) - Y^T AY \\
&= X^T AX - Y^T AY && [\because \textit{from}(i),\ (Y^T AX - X^T AY) = 0]
\end{aligned}
$$

So, we have shown:

$$(X+Y)^T A(X-Y) = X^T AX - Y^T AY \qquad (ii)$$

Now, Equation (4.9) is the log-ratio, and it is zero when the ratio is 1 (i.e., log(1) = 0). This is true at the boundary of the two classes.

Thus, $\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) = 0 \qquad (iii)$

If we extend the LHS of (*iii*) and group class-wise, we can describe the discriminant function.

$$
\begin{aligned}
\text{LHS} &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) \\
&= log\, \pi_k - log\, \pi_l - \frac{1}{2}[(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l)] + x^T \Sigma^{-1}(\mu_k - \mu_l) \\
&= log\, \pi_k - log\, \pi_l - \frac{1}{2}[\mu_k^T \Sigma^{-1} \mu_k - \mu_l^T \Sigma^{-1} \mu_l] + x^T \Sigma^{-1}(\mu_k - \mu_l) \\
& \qquad [\text{using } (ii) \text{ for the terms in '[…]'}]
\end{aligned}
$$

$$
\begin{aligned}
&= log\, \pi_k - log\, \pi_l - \frac{1}{2}[\mu_k^T \Sigma^{-1} \mu_k - \mu_l^T \Sigma^{-1} \mu_l] + x^T \Sigma^{-1} \mu_k - x^T \Sigma^{-1} \mu_l \\
&= [x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + log\, \pi_k] \ - \ [x^T \Sigma^{-1} \mu_l - \frac{1}{2}\mu_l^T \Sigma^{-1} \mu_l + log\, \pi_l] \\
&= \delta_k(x) - \delta_l(x) && [\because \delta_l(x) = x^T \Sigma^{-1} \mu_l - \frac{1}{2}\mu_l^T \Sigma^{-1} \mu_l + log\, \pi_l]
\end{aligned}
$$

Following Equation (iii), for the boundary condition we can write:

$$\delta_k(x) = \delta_l(x)$$

Note [Ends]

In practice, we do not know the parameters of the Gaussian distributions, and will need to estimate them using our training data:

- $\hat{\pi}_k = N_k / N$, where N_k is the number of class-k observations;
- $\hat{\mu}_k = \sum_{g_i=k} x_i / N_k$.
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$.

Figure 4.5 (right panel) shows the estimated decision boundaries based on a sample of size 30 each from three Gaussian distributions.

With two classes, there is a simple correspondence between linear discriminant analysis and classification by linear least squares, as in (4.5). The LDA rule classifies to class 2 if:

$$x^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + \log(N_1 / N) - \log(N_2 / N) \quad (4.11)$$

And otherwise, class 1.

Note: How we get Equation (4.11)?

From, Equation (4.10) we got: $\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$

So, we can write:

$$\delta_2(x) = x^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \log(N_2 / N) \quad \ldots (i) \quad [\because \hat{\pi}_k = N_k / N]$$

Also, $$\delta_1(x) = x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \log(N_1 / N) \quad \ldots (ii)$$

To classify to class 2,

$\delta_2(x) > \delta_1(x)$ [Now we use (i) and (ii) to replace by respective RHS]

$$\Rightarrow \quad [\, x^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \log(N_2/N)\,] > [\, x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log(N_1/N)\,]$$

$$\Rightarrow \quad x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}\hat{\mu}_2^T\,\hat{\Sigma}^{-1}\,\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^T\,\hat{\Sigma}^{-1}\,\hat{\mu}_1 + \log(N_1/N) - \log(N_2/N)$$

which is same as Equation (4.11).

Note [Ends]

Quadratic discriminant analysis (QDA)

Now, getting back to the general discriminant problem (4.8), if the Σ_k are not assumed to be equal, then the convenient cancellations in (4.9) do not occur [***Already shown in (A1)***]; in particular, the pieces quadratic in *x* remain. We then get *quadratic discriminant functions* (can easily be seen from Equation (**A0**)),

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \qquad (4.12)$$

The decision boundary between each pair of classes *k* and ℓ is described by a quadratic equation $\{x : \delta_k(x) = \delta_l(x)\}$.

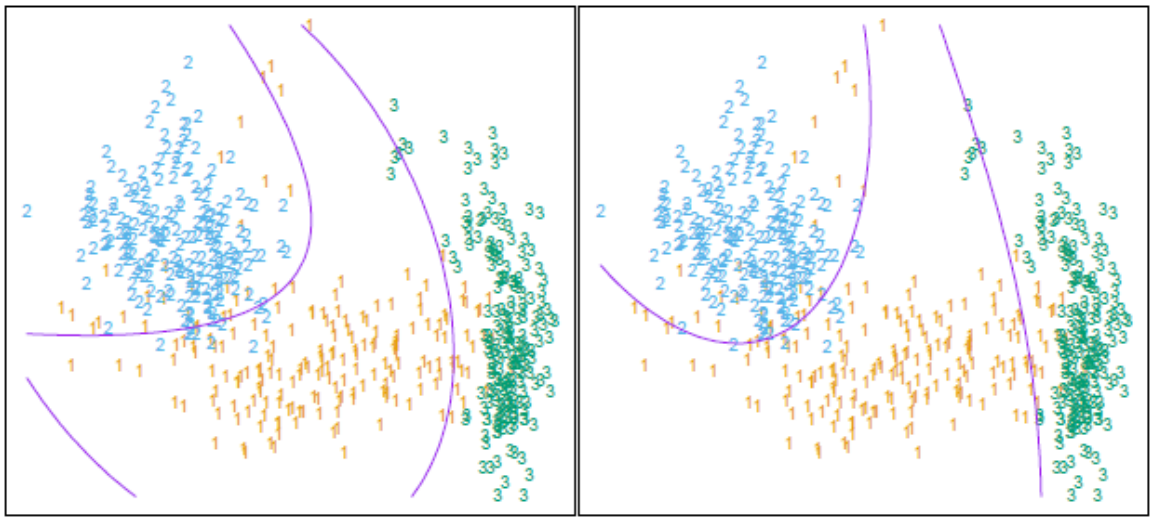

Figure 4.6 [1]: Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space X_1, X_2, X_1X_2, X_1^2, X_2^2). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.

Figure 4.6 shows an example where the three classes are Gaussian mixtures (Section 6.8 of [1]), and the decision boundaries are approximated by quadratic equations in *x*. Here we illustrate two popular ways of fitting these quadratic boundaries. The right plot uses QDA as described here, while the

left plot uses LDA in the enlarged five-dimensional quadratic polynomial space. The differences are generally small; QDA is the preferred approach, with the LDA method a convenient substitute[2].

The estimates for QDA are similar to those for LDA, except that separate covariance matrices must be estimated for each class. When p is large, this can mean a dramatic increase in parameters. Since the decision boundaries are functions of the parameters of the densities, counting the number of parameters must be done with care. For LDA, it seems there are $(K - 1) \times (p + 1)$ parameters, since we only need the differences $\delta_k(x) - \delta_K(x)$ between the discriminant functions where K is some pre-chosen class (here we have chosen the last), and each difference requires p + 1 parameters. Likewise for QDA there will be $(K - 1) \times \{p\,(p + 3)/2 + 1\}$ parameters (Since, for quadratic expansion from p feature can be p(p+1)/2 and (p+1) variables of power 1; therefore p(p+1)/2+(p+1) = $\{p\,(p + 3)/2 + 1\}$. Both LDA and QDA perform well on an amazingly large and diverse set of classification tasks. Both techniques are widely used, and entire books are devoted to LDA. It seems that whatever exotic tools are the rage of the day, we should always have available these two simple tools. The question arises why LDA and QDA have such a good track record. The reason is not likely to be that the data are approximately Gaussian, and, in addition, for LDA, that the covariances are approximately equal. More likely, a reason is that the data can only support simple decision boundaries such as linear or quadratic, and the estimates provided via the Gaussian models are stable. This is a bias-variance tradeoff—we can put up with the bias of a linear decision boundary because it can be estimated with much lower variance than more exotic alternatives. This argument is less believable for QDA since it can have many parameters itself, although perhaps fewer than the non-parametric alternatives.

4.3.1 *Regularized Discriminant Analysis*
4.3.2 Computations for LDA

Optional: read them from the Book# [1], page 112 to 113

Reference:

1. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009: Springer.

---------- ×--------

[2] For this figure and many similar figures in the book we compute the decision boundaries by an exhaustive contouring method. We compute the decision rule on a fine lattice of points, and then use contouring algorithms to compute the boundaries.