Gaussian Models for Class Densities

When X is a real vector of length p, we might model the density for class k, which I'll write $p(x|k)$ — the book uses $f_k(x)$ — as a Gaussian (normal) distribution.

Letting μ_k be the mean vector, and Σ_k the covariance matrix:

$$p(x|k) \;=\; (2\pi)^{-p/2}\,|\Sigma_k|^{-1/2}\,\exp(-(1/2)(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))$$

The obvious way to estimate the parameters μ_k and Σ_k from the training data is

$$\hat{\mu}_k \;=\; \frac{1}{N_k}\sum_{i:g_i=k} x_i$$

$$\hat{\Sigma}_k \;=\; \frac{1}{N_k-1}\sum_{i:g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T$$

where N_k is the number of training cases (out of N) in class k.

We can also estimate the class proportions by

$$\hat{\pi}_k \;=\; N_k/N$$

Result: Quadratic Discriminant Functions

We should classify a test case with inputs x as class k if $\pi_k \, p(x|k)$ is greater for this k than for any other. Taking logs, we see that this corresponds to using the following discriminant functions:

$$\delta_k(x) \;\; = \;\; \log \hat{\pi}_k \;\; - \;\; (1/2) \log |\hat{\Sigma}_k| \;\; - \;\; (1/2)(x - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (x - \hat{\mu}_k)$$

This leaves out the constant 2π parts, that don't affect the comparisons of $\delta_k(x)$ with $\delta_{k'}(x)$.

Note that $\delta_k(x)$ is a quadratic function of x. So the boundaries between adjacent classes (where $\delta_k(x) = \delta_{k'}(x)$) are given by quadratic equations.

This method is called *Quadratic Discriminant Analysis (QDA)*.

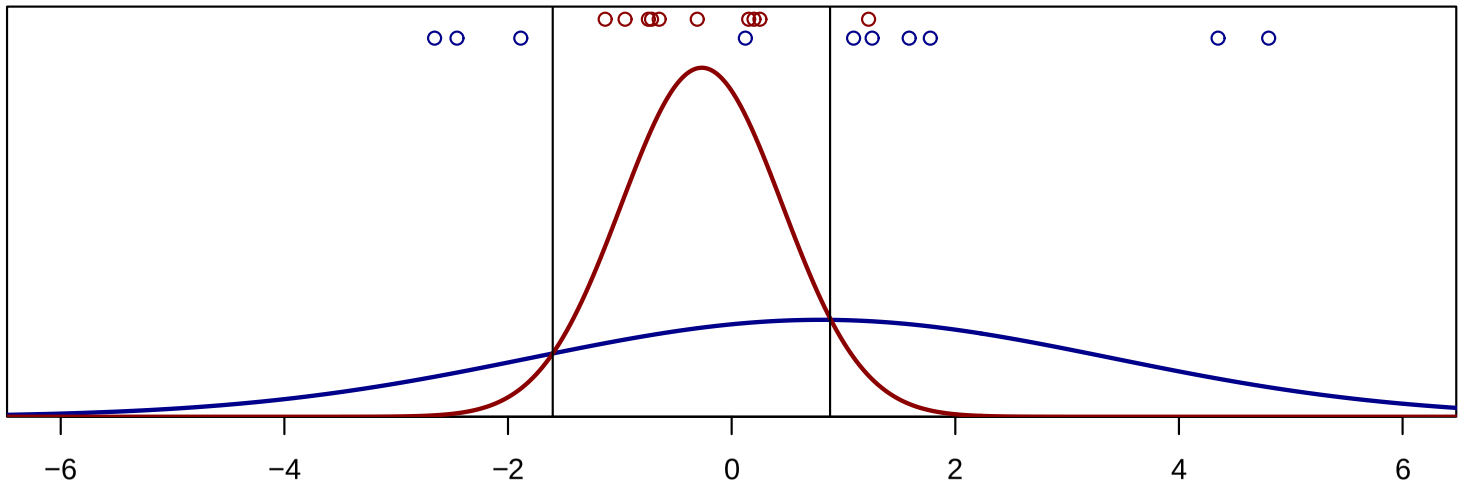
QDA With One Input Variable

If we have only one input variable, the sample covariance "matrix" is just the sample variance, $\hat{\sigma}_k^2$. The discriminant functions are then

$$\delta_k(x) \;=\; \log \hat{\pi}_k \;-\; (1/2)\log(\hat{\sigma}_k^2) \;-\; (1/2\hat{\sigma}^2)(x-\hat{\mu}_k)^2$$

With only two classes, the class boundaries are the (generally two) points where $\delta_0(x) = \delta_1(x)$.

The example below shows ten training points from each of two classes, with the estimated normal density functions of each class. The class boundaries are where they cross (since here $\hat{\pi}_0 = \hat{\pi}_1$).



Linear Discriminant Analysis (LDA)

Rather than estimate the Σ_k separately for each class, we might assume that they are all the same. We would then estimate this common covariance matrix by

$$\hat{\Sigma} \;=\; \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

The discriminant functions are then

$$\delta_k(x) \;=\; \log \hat{\pi}_k \;-\; (1/2) \log |\hat{\Sigma}| \;-\; (1/2)(x - \hat{\mu}_k)^T \hat{\Sigma}^{-1} (x - \hat{\mu}_k)$$

By multipying out, and removing parts that are the same for all k, we can simplify the discriminant functions to

$$\delta_k(x) \;=\; \log \hat{\pi}_k \;-\; (1/2) \mu_k^T \hat{\Sigma}^{-1} \mu_k \;+\; x^T \hat{\Sigma}^{-1} \hat{\mu}_k$$

Note that this is a *linear* function of x. Class boundaries are hyperplanes.

Relationship of LDA and Logistic Regression

When there are just two classes,

$$P(G=1 \,|\, X=x) \;=\; \frac{\pi_1 \, p(x|1)}{\pi_0 \, p(x|0) + \pi_0 \, p(x|1)}$$

According to the Gaussian model with equal covariances used by LDA, we can write the logit of this as

$$\begin{aligned}
\text{logit}(P(G=1 \,|\, X=x)) \;&=\; \log\left(\frac{P(G=1 \,|\, X=x)}{1-P(G=1 \,|\, X=x)}\right) \;=\; \log\left(\frac{\pi_1 \, p(x|1)}{\pi_0 \, p(x|0)}\right) \\
&= \; [\log \pi_1 - (1/2)\mu_1^T \Sigma^{-1} \mu_1 + x^T \Sigma^{-1} \mu_1] \;-\; [\log \pi_0 - (1/2)\mu_0^T \Sigma^{-1} \mu_0 + x^T \Sigma^{-1} \mu_0] \\
&= \; \log(\pi_1/\pi_0) - (1/2)(\mu_1 - \mu_0)^T \Sigma^{-1} (\mu_1 - \mu_0) + x^T \Sigma^{-1} (\mu_1 - \mu_0)
\end{aligned}$$

Notice that this is a linear function of x, just as in logistic regression. The same holds for more than two classes.

However, the ways the coefficients are estimated are **not** equivalent.