Linear Discriminant Analysis Explained

James Taylor

We assume that our data come from k populations $\pi_1, \ldots, \pi_k$ with respective prior probabilities $\gamma_1, \ldots, \gamma_k$. We further assume that in the population π_i the pdf of $X = (X_1, \ldots, X_p)$ is multivariate normal with mean vector μ_i and covariance matrix Σ. That is, all populations have their own mean vector, but they all share the same covariance matrix. Thus,

$$f_i(x) := p(X = x \,|\, Y = i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right]$$

The LDA classifier assigns x to the population π_i for which $p(Y = i \,|\, X = x)$ is largest. By Bayes theorem,

$$\begin{aligned} p(Y = i \,|\, X = x) &= \frac{p(X = x \,|\, Y = i)p(Y = i)}{p(X = x)} \\ &= \frac{\gamma_i}{p(X = x)} f_i(x) \end{aligned}$$

Since $p(X = x)$ does not depend on i, it can be treated as a constant. So x will be assigned to the population π_i for which $\gamma_i f_i(x)$, or equivalently, $\log \gamma_i f_i(x)$ is largest. Now,

$$\begin{aligned} \log(\gamma_i f_i(x)) &= \log(f_i(x)) + \log(\gamma_i) \\ &= \log\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right]\right) + \log(\gamma_i) \\ &= \log\left(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right) + \left[-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right] + \log(\gamma_i) \end{aligned}$$

Again, notice that the first term on the last line does not depend on i and thus can also be treated as a constant. Therefore, x will be assigned to the population π_i for which

$$-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \log(\gamma_i)$$

is the largest. We can simplify this further:

$$\begin{aligned} -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \log(\gamma_i) &= -\frac{1}{2}\left[x^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i - x^T\Sigma^{-1}\mu_i - \mu_i^T\Sigma^{-1}x\right] + \log(\gamma_i) \\ &= -\frac{1}{2}\left[x^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i\right] + \mu_i^T\Sigma^{-1}x + \log(\gamma_i) \end{aligned}$$

and since $x^T\Sigma^{-1}x$ doesn't depend on i, it suffices to assign x to the population π_i whose linear discriminant function

$$d_i^L(x) = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \mu_i^T\Sigma^{-1}x + \log(\gamma_i)$$

is largest when evaluated at x.

Note: Setting $d_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i$ and $d_{ij} =$ jth element of $\mu_i^T\Sigma^{-1}$ we have

$$\begin{aligned} d_i^L(x) &= d_{i0} + d_{i1}x_1 + \cdots + d_{ip}x_p \\ &= d_i \cdot (1, x) \end{aligned}$$

Note: We will need to estimate the prior probabilities γ_i, the population mean vectors μ_i, and the covariance matrix Σ. The latter two are estimated by the sample mean vectors and the pooled covariance matrix, respectively.