Linear discriminant analysis

Linear discriminant analysis (LDA) is a common "plug-in" method for classification which operates by estimating $\pi_k f_{X|Y}(\boldsymbol{x}|k)$ for each class $k = 0, \ldots, K-1$ and then simply plugging these into the formula for the Bayes classifier in order to make a decision. In LDA we make the (strong) assumption that class conditional pdfs are given by the multivariate normal distribution, but with differing means. Mathematically, this corresponds to the assumption that

$$f_{X|Y}(\boldsymbol{x}|k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}$$

for $k = 0, \ldots, K-1$. Note that under this assumption, each class has a distinct mean $\boldsymbol{\mu}_k$, but all classes share the same covariance matrix $\boldsymbol{\Sigma}$.

In LDA, we assume that $\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_{K-1}$ and $\boldsymbol{\Sigma}$, as well as the prior probabilities $\pi_0, \ldots, \pi_{K-1}$ are all unknown, but can be estimated from the data. In particular, we can use the estimates

$$\begin{aligned}
\widehat{\pi}_k &= \frac{|\{i : y_i = k\}|}{n} \\
\widehat{\boldsymbol{\mu}}_k &= \frac{1}{|\{i : y_i = k\}|} \sum_{i : y_k = k} \boldsymbol{x}_i \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n} \sum_{k=0}^{K-1} \sum_{i : y_i = k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T.
\end{aligned}$$

The LDA classifier is then defined by

$$\widehat{h}(\boldsymbol{x}) = \arg\max_k \ \widehat{\pi}_k \cdot \frac{1}{(2\pi)^{d/2}|\widehat{\boldsymbol{\Sigma}}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)}.$$

Georgia Tech ECE 6254 Spring 2017; Notes by M. A. Davenport and J. Romberg. Last updated 21:39, January 24, 2017

Since the log is a monotonic transformation (meaning that if $x > y$ then $\log(x) > \log(y)$), we can equivalently state the classifier as

$$\begin{aligned}
\widehat{h}(\boldsymbol{x}) = & \arg\max_k \log\left(\widehat{\pi}_k\right) + \log\left(\frac{1}{(2\pi)^{d/2}|\widehat{\boldsymbol{\Sigma}}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)}\right) \\
&= \arg\max_k \ \log\left(\widehat{\pi}_k\right) - \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) \\
&= \arg\min_k \ \frac{1}{2}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_k) - \log\left(\widehat{\pi}_k\right)
\end{aligned}$$

where the second equality above follows from the fact that

$$\log\left(\frac{1}{(2\pi)^{d/2}|\widehat{\boldsymbol{\Sigma}}|^{1/2}}\right)$$

is constant across all k and so does not affect which k maximizes the expression.

It is enlightening to consider what happens in the special case of $K = 2$ (i.e., binary classification). In this case, LDA results in a classifier such that $\widehat{h}(\boldsymbol{x}) = 1$ when

$$(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_0) - 2\log\widehat{\pi}_0 \geq (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1) - 2\log\widehat{\pi}_1.$$

We can rewrite this as

$$(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_0) - (\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\widehat{\boldsymbol{\mu}}_1) + 2\log\frac{\widehat{\pi}_1}{\widehat{\pi}_0} \geq 0.$$

Using the fact that $\boldsymbol{\Sigma}$ is symmetric, which implies that we have

$(\boldsymbol{\Sigma}^{-1})^T = \boldsymbol{\Sigma}^{-1}$, we can simplify this rule to

$$
\begin{aligned}
0 &\leq (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_0) - (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1) + 2 \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} \\
&= \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - 2 \widehat{\boldsymbol{\mu}}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \widehat{\boldsymbol{\mu}}_0^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_0 \\
&\quad - \left(\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - 2 \widehat{\boldsymbol{\mu}}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \widehat{\boldsymbol{\mu}}_1^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_1 \right) + 2 \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} \\
&= 2 (\widehat{\boldsymbol{\mu}}_1^T - \widehat{\boldsymbol{\mu}}_0^T) \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \widehat{\boldsymbol{\mu}}_0^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_1 + 2 \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0} \\
&= (\boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0))^T \boldsymbol{x} + \frac{1}{2} \widehat{\boldsymbol{\mu}}_0^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_0 - \frac{1}{2} \widehat{\boldsymbol{\mu}}_1^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_1 + \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0}.
\end{aligned}
$$

Thus, if

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)$$

and

$$b = \frac{1}{2} \widehat{\boldsymbol{\mu}}_0^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_0 - \frac{1}{2} \widehat{\boldsymbol{\mu}}_1^T \boldsymbol{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_1 + \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0},$$

we can re-write this as

$$\boldsymbol{w}^T \boldsymbol{x} + b \geq 0.$$

This is the expression of a simple linear classifier, and thus LDA will always result in a linear classifier.