Linear Discriminant Analysis

As with the setting of two-sample tests, suppose we have two samples from two populations with the same covariance but different means:

$$\mathcal{N}_p(\vec{\mu}_1, \mathbf{\Sigma}) : \vec{x}_{11}, \dots, \vec{x}_{1n_1},$$

and

$$\mathcal{N}_p(\vec{\mu}_2, \mathbf{\Sigma}) : \vec{x}_{21}, \dots, \vec{x}_{2n_2}.$$

Denote the summary statistics as $\bar{\vec{x}}_1$, $\bar{\vec{x}}_2$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. The inferential task in two-sample test is to test $H_0 : \vec{\mu}_1 = \vec{\mu}_2$, or to find confidence region of $\vec{\mu}_1 - \vec{\mu}_2$, while in discriminant analysis, the goal is to classify a new observation $\vec{x}_0$ to either Class 1 or Class 2.

1 Perspective 1: Comparison of Mahalanobis Distances

The first approach is geometric intuitive. We calculated the Mahalanobis distances $d_M(\vec{x}_0, \bar{\vec{x}}_1)$ and $d_M(\vec{x}_0, \bar{\vec{x}}_2)$, and assign $\vec{x}_0$ to the closer class. Given the assumption that the two classes have the same population covariance, we define the Mahalanobis distance based on the pooled sample covariance matrix

$$\boldsymbol{S}_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_2.$$

Then, we have the following neat rule to classify $\vec{x}_0$:

Definition 1. ***Fisher's rule*** *for classification is defined as follows: any new observation $\vec{x}_0$ is assigned to Class 1 if*

$$(\vec{x}_0 - \bar{\vec{x}}_1)^\top \boldsymbol{S}_{pooled}^{-1}(\vec{x}_0 - \bar{\vec{x}}_1) \leqslant (\vec{x}_0 - \bar{\vec{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\ (\vec{x}_0 - \bar{\vec{x}}_2)$$

or equivalently

$$(\bar{\vec{x}}_1 - \bar{\vec{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\left(\vec{x}_0 - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2)\right) \geq 0,$$

or simply

$$\vec{w}^\top \vec{x}_0 \geq \frac{1}{2}\vec{w}^\top(\bar{\vec{x}}_1 + \bar{\vec{x}}_2),$$

where $\vec{w} = \boldsymbol{S}_{pooled}^{-1}(\bar{\vec{x}}_1 - \bar{\vec{x}}_2)$.

Otherwise, $\vec{x}_0$ is assigned to Class 2.

2 Perspective 2: Bayes Rule

The second perspective for linear discriminant is based on the distributional assumptions. For a new observation $\vec{x}_0$, we assume it is the realization of some random vector $\vec{X}$, which is from a mixture of $\mathcal{N}_p(\vec{\mu}_1, \mathbf{\Sigma})$ and $\mathcal{N}_p(\vec{\mu}_1, \mathbf{\Sigma})$. To be specific, we assume that the prior probabilities on the two classes are half and half, then the pdf of $\vec{X}$ is

$$f_{\vec{X}}(\vec{x}) = \frac{1}{2} f_1(\vec{x}) + \frac{1}{2} f_2(\vec{x}),$$

where f_l is the pdf of $\mathcal{N}_p(\vec{\mu}_l, \mathbf{\Sigma})$ for $l = 1, 2$, i.e.,

$$f_1(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_1)^\top \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu}_1) \right).$$

and

$$f_2(\vec{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu}_2)^\top \mathbf{\Sigma}^{-1} (\vec{x} - \vec{\mu}_2) \right).$$

For the observation $\vec{X} = \vec{x}_0$, it is natural to classify the observation by comparing the conditional (or posterior) probabilities

$$\mathbb{P}(\vec{X} \text{ is sampled from Class } 1 | \vec{X} = \vec{x}_0)$$

and

$$\mathbb{P}(\vec{X} \text{ is sampled from Class } 1 | \vec{X} = \vec{x}_0).$$

By Bayes rule, we have

$$\begin{aligned} &\mathbb{P}(\vec{X} \text{ is sampled from Class } l | \vec{X} = \vec{x}_0) \\ =& \frac{\mathbb{P}(\vec{X} \text{ is sampled from Class } l) f(\vec{x}_0 | \vec{X} \text{ is sampled from Class } l)}{f_{\vec{X}}(\vec{x}_0)} \\ =& \frac{\frac{1}{2} f_l(\vec{x}_0)}{\frac{1}{2} f_1(\vec{x}_0) + \frac{1}{2} f_2(\vec{x}_0)} \\ =& \frac{f_l(\vec{x}_0)}{f_1(\vec{x}_0) + f_2(\vec{x}_0)} \end{aligned}$$

Then, the conditional probability criterion is reduced to the comparison between

$$\frac{f_1(\vec{x}_0)}{f_1(\vec{x}_0) + f_2(\vec{x}_0)} \quad \text{and} \quad \frac{f_2(\vec{x}_0)}{f_1(\vec{x}_0) + f_2(\vec{x}_0)}.$$

Therefore, $\vec{x}_0$ is allocated to Class 1 if and only if $f_1(\vec{x}_0) \geq f_2(\vec{x}_0)$, or equivalently,

$$\begin{aligned} \log \frac{f_1(\vec{x}_0)}{f_2(\vec{x}_0)} &= -\frac{1}{2} (\vec{x}_0 - \vec{\mu}_1)^\top \mathbf{\Sigma}^{-1} (\vec{x}_0 - \vec{\mu}_1) + \frac{1}{2} (\vec{x}_0 - \vec{\mu}_2)^\top \mathbf{\Sigma}^{-1} (\vec{x}_0 - \vec{\mu}_2) \\ &= (\vec{\mu}_1 - \vec{\mu}_2)^\top \mathbf{\Sigma}^{-1} \left(\vec{x}_0 - \frac{1}{2} (\vec{\mu}_1 + \vec{\mu}_2) \right) \geq 0. \end{aligned}$$

In practice, the population mean shift is replaced with the sample mean shift, while the population covariance is replaced with the pooled sample covariance matrix, that is, $\vec{x}_0$ is allocated to Class 1 if and only if

$$(\bar{\vec{x}}_1 - \bar{\vec{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \left(\vec{x}_0 - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2) \right) \geq 0,$$

which is the Fisher's rule derived before.

3 Perspective 3: Mean Difference Maximization after projecting onto one line

Let's come back to the geometrical meaning of the "boundary" of allocation based on the Fisher's rule:

$$\vec{w}^\top \left(\vec{x}_0 - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2) \right) = 0,$$

which is a hyperplane that is orthogonal to $\vec{w}$ and goes through $\frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2)$. In the previous two sections, two methods have been proposed to determine this rule. Here we consider a straightforward approach to find $\vec{w}$ based on the following geometrical idea:

Suppose the two classes have distributions $\mathcal{N}_p(\vec{\mu}_1, \mathbf{\Sigma})$ and $\mathcal{N}_p(\vec{\mu}_2, \mathbf{\Sigma})$, respectively. Looking for a hyperplane to separate the two distributions as much as possible, is equivalent to projecting the two distributions onto the line determined by $\vec{w}$, such that the two "projected" distributions are separated as much as possible.

If $\vec{X} \sim \mathcal{N}_p(\vec{\mu}_1, \mathbf{\Sigma})$, its "projection" on $\vec{w}$, i.e., $\vec{w}^\top \vec{X}$, is distributed in $\mathcal{N}(\vec{w}^\top \vec{\mu}_1, \vec{w}^\top \mathbf{\Sigma} \vec{w})$. Similarly, if $\vec{X} \sim \mathcal{N}_p(\vec{\mu}_2, \mathbf{\Sigma})$, $\vec{w}^\top \vec{X}$ is distributed in $\mathcal{N}(\vec{w}^\top \vec{\mu}_1, \vec{w}^\top \mathbf{\Sigma} \vec{w})$. Our goal becomes to separate the two distributions $\mathcal{N}(\vec{w}^\top \vec{\mu}_1, \vec{w}^\top \mathbf{\Sigma} \vec{w})$ and $\mathcal{N}(\vec{w}^\top \vec{\mu}_1, \vec{w}^\top \mathbf{\Sigma} \vec{w})$.

Notice that these two one-dimensional distributions have different means but the same variance. In order to separate them as much as possible, we want to maximize the mean difference $|\vec{w}^\top (\vec{\mu}_1 - \vec{\mu}_2)|$, and at the same time minimize the variance $\vec{w}^\top \mathbf{\Sigma} \vec{w}$. One idea is to maximize the scaled squared mean difference:

$$\frac{|\vec{w}^\top (\vec{\mu}_1 - \vec{\mu}_2)|^2}{\vec{w}^\top \mathbf{\Sigma} \vec{w}}.$$

By the extended Cauchy-Schwarz inequality we have seen (for the derivation of simultaneous confidence intervals), we have

$$\frac{|\vec{w}^\top (\vec{\mu}_1 - \vec{\mu}_2)|^2}{\vec{w}^\top \mathbf{\Sigma} \vec{w}} \leq \frac{(\vec{w}^\top \mathbf{\Sigma} \vec{w})(\vec{\mu}_1 - \vec{\mu}_2)^\top \mathbf{\Sigma}^{-1} (\vec{\mu}_1 - \vec{\mu}_2)}{\vec{w}^\top \mathbf{\Sigma} \vec{w}} = (\vec{\mu}_1 - \vec{\mu}_2)^\top \mathbf{\Sigma}^{-1} (\vec{\mu}_1 - \vec{\mu}_2),$$

where the equality can be attained by letting $\mathbf{\Sigma}^{\frac{1}{2}} \vec{w} = \mathbf{\Sigma}^{-\frac{1}{2}} (\vec{\mu}_1 - \vec{\mu}_2)$, i.e.,

$$\vec{w} = \mathbf{\Sigma}^{-1} (\vec{\mu}_1 - \vec{\mu}_2).$$

In the sample case, we replace the above formula with the empirical version

$$\vec{w} = \boldsymbol{S}_{pooled}^{-1} (\bar{\vec{x}}_1 - \bar{\vec{x}}_2),$$

which is the same as the Fisher's rule we have derived before.

4 Linear Invariance

Since Fisher's rule can be obtained by comparing Mahalanobis distances, and we have shown that Mahalanobis distances are linearly invariant, this directly implies the linear invariance of LDA. This section gives a detailed proof.

Consider the uniform linear transformation

$$\vec{y}_{k,i} = \boldsymbol{C}\vec{x}_{k,i} + \vec{d},$$

where $\boldsymbol{C}$ is a $p \times p$ nonsingular matrix, $k = 1, 2$, and $i = 1, \ldots, n_k$. For the sample mean, we have

$$\bar{\vec{y}}_1 = \boldsymbol{C}\bar{\vec{x}}_1 + \vec{d}, \quad \bar{\vec{y}}_2 = \boldsymbol{C}\bar{\vec{x}}_2 + \vec{d}.$$

For the sample covariance

$$\boldsymbol{S}_{y,1} = \boldsymbol{C}\boldsymbol{S}_{x,1}\boldsymbol{C}^\top, \quad \boldsymbol{S}_{y,2} = \boldsymbol{C}\boldsymbol{S}_{x,2}\boldsymbol{C}^\top.$$

This gives

$$\begin{aligned}
\boldsymbol{S}_{y,pooled} &= \frac{n_1 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_{y,1} + \frac{n_2 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_{y,2} \\
&= \frac{n_1 - 1}{n_1 + n_2 - 2}\boldsymbol{C}\boldsymbol{S}_{x,1}\boldsymbol{C}^\top + \frac{n_2 - 1}{n_1 + n_2 - 2}\boldsymbol{C}\boldsymbol{S}_{x,2}\boldsymbol{C}^\top \\
&= \boldsymbol{C}\left(\frac{n_1 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_{x,1} + \frac{n_2 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_{x,2}\right)\boldsymbol{C}^\top \\
&= \boldsymbol{C}\boldsymbol{S}_{x,pooled}\boldsymbol{C}^\top
\end{aligned}$$

For a new observation $\vec{x}$, let

$$\vec{y} = \boldsymbol{C}\vec{x} + \vec{d}.$$

Linear invariance for Fisher's rule

$$\begin{aligned}
&(\bar{\vec{y}}_1 - \bar{\vec{y}}_2)^\top \boldsymbol{S}_{y,pooled}^{-1}\left(\vec{y} - \frac{1}{2}(\bar{\vec{y}}_1 + \bar{\vec{y}}_2)\right) \\
&= ((\boldsymbol{C}\bar{\vec{x}}_1 + \vec{d}) - (\boldsymbol{C}\bar{\vec{x}}_2 + \vec{d}))^\top (\boldsymbol{C}\boldsymbol{S}_{x,pooled}\boldsymbol{C}^\top)^{-1}\left((\boldsymbol{C}\vec{x} + \vec{d}) - \frac{1}{2}(\boldsymbol{C}\bar{\vec{x}}_1 + \vec{d} + \boldsymbol{C}\bar{\vec{x}}_2 + \vec{d})\right) \\
&= (\boldsymbol{C}(\bar{\vec{x}}_1 - \bar{\vec{x}}_2))^\top (\boldsymbol{C}\boldsymbol{S}_{x,pooled}\boldsymbol{C}^\top)^{-1}\left(\boldsymbol{C}\left(\vec{x} - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2)\right)\right) \\
&= (\bar{\vec{x}}_1 - \bar{\vec{x}}_2)^\top \boldsymbol{C}^\top (\boldsymbol{C}^\top)^{-1}\boldsymbol{S}_{x,pooled}^{-1}\boldsymbol{C}^{-1}\left(\boldsymbol{C}\left(\vec{x} - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2)\right)\right) \\
&= (\bar{\vec{x}}_1 - \bar{\vec{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\left(\vec{x} - \frac{1}{2}(\bar{\vec{x}}_1 + \bar{\vec{x}}_2)\right).
\end{aligned}$$

5 Training error and validation error

For the concepts of apparent error rates (training error) and estimated actual error rate (validation/testing error), please read the book from Page 598 - 600.