Derivation of Between-Class Scatter Matrix in LDA

Lu Sun

March 18, 2020

Suppose that we have a dataset with K classes. There are N_k samples in the k-th class, and its i-th sample is denoted by $x_i \in \mathbb{R}^p$. In LDA, we decompose the total scatter matrix $\mathbf{T}$ into a sum of between-class scatter $\mathbf{B}$ and within-class scatter $\mathbf{W}$:

$$\mathbf{T} = \mathbf{B} + \mathbf{W}, \tag{1}$$

in which

$$\begin{aligned}
\mathbf{T} &= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \mu)(x_i - \mu)^T \\
\mathbf{W} &= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \mu_k)(x_i - \mu_k)^T.
\end{aligned} \tag{2}$$

In (2), μ and μ_k denote the global sample mean and class-specific sample mean, respectively,

$$\mu = \frac{1}{N} \sum_{k=1}^{K} \sum_{g_i=k} x_i, \quad \mu_k = \frac{1}{N_k} \sum_{g_i=k} x_i. \tag{3}$$

According to the fact that,

$$x_i - \mu = (x_i - \mu_k) + (\mu_k - \mu), \tag{4}$$

the total scatter can be rewritten by

$$\begin{aligned}
\mathbf{T} &= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \mu)(x_i - \mu)^T \\
&= \sum_{k=1}^{K} \sum_{g_i=k} \left[(x_i - \mu_k)(x_i - \mu_k)^T + 2(x_i - \mu_k)(\mu_k - \mu)^T + (\mu_k - \mu)(\mu_k - \mu)^T \right] \\
&= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \mu_k)(x_i - \mu_k)^T + 2 \sum_{k=1}^{K} \left[(\mu_k - \mu)^T \sum_{g_i=k} (x_i - \mu_k) \right] + \sum_{k=1}^{K} \sum_{g_i=k} (\mu_k - \mu)(\mu_k - \mu)^T \\
&= \mathbf{W} + \mathbf{0} + \sum_{k=1}^{K} N_k (\mu_k - \mu)(\mu_k - \mu)^T.
\end{aligned} \tag{5}$$

The fourth equation in (5) holds because

$$\sum_{g_i=k} (x_i - \mu_k) = \sum_{g_i=k} x_i - N_k \mu_k = N_k \mu_k - N_k \mu_k = 0. \tag{6}$$

Thus, it is reasonable to represent the between-class scatter matrix by

$$\mathbf{B} = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T. \tag{7}$$