Notes on Fisher's Linear Discriminant

Zhihan Yang

March 27, 2020

1 Binary classification problem

There are two classes, C_1 and C_2, each containing a non-zero number of data vectors, $\boldsymbol{x}^n$. For each class, C_k, we define some properties that will be used later on in this document.

The number of vectors in class k is given by N_k.

The mean vector of class k is given by:

$$\boldsymbol{m}_k = \frac{1}{N_k} \sum_{n \in C_1} \boldsymbol{x}^n$$

The projected mean of class k is given by:

$$m_k = \boldsymbol{w}^T \boldsymbol{m}_k$$

The within-class variance of class k is given by:

$$s_k = \sum_{n \in C_k} (y^n - m_k)^2$$

2 Fisher's criterion

A linear combination between input variables, x_i, and weights w_i, can be written in vector form as:

$$y = \boldsymbol{w}^T \boldsymbol{x}$$

where y can be interpreted as the projection of $\boldsymbol{x}$ onto $\boldsymbol{w}$.

Given two classes, C_1 and C_2, we would like to select a $\boldsymbol{w}$ such that the projection of $\boldsymbol{x}$ maximizes class separation. Intuitively, this means maximizing between-class separation while minimizing within-class separation. Mathematically, we define between-class separation as $(m_2 - m_1)^2$, and define within-class separation as $s_1^2 + s_2^2$. We can complete the two optimization tasks simultaneously by maximizing the following ratio, known as the Fisher criterion:

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

Now, the question remains: how do we find the optimal $\boldsymbol{w}$ to maximize $J(\boldsymbol{w})$?

3 The Fisher criterion in matrix form

It turns out that the Fisher criterion is more conveniently optimized in matrix form. In this section, we convert the algebraic form derived in the last section into its matrix form.

Expand the numerator:

$$\begin{aligned} \text{numerator}(J(\boldsymbol{w})) &= (m_2 - m_1)^2 && (1)\\ &= (m_2 - m_1)(m_2 - m_1) && (2)\\ &= (\vec{w}^T\vec{x} - \vec{w}^T\vec{m_1})(\vec{w}^T\vec{x} - \vec{w}^T\vec{m_2}) && (3) \end{aligned}$$

where both factors are scalars, so the order of multiplication does not matter within each factor,

$$= \vec{w}^T(\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_2)^T\vec{w} \quad (4)$$

Since matrix multiplication is associative, i.e., $A(BC) = (AB)C$, we can rewrite the numerator as:

$$\text{numerator}(J(\boldsymbol{w})) = \vec{w}^T S_B \vec{w}$$

where $S_B = (\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_2)^T$ and is called the between-class covariance matrix.

Expand the denominator (somewhat tedious):

$$\begin{aligned}
\text{denominator}(J(\vec{w})) &= s_1^2 + s_2^2 && (5)\\
&= \sum_{n \in C_1} (y^n - m_1)^2 + \sum_{n \in C_2} (y^n - m_2)^2 && (6)\\
&= \sum_{n \in C_1} (\vec{w}^T \vec{x} - \vec{w}^T \vec{m}_1)^2 + \sum_{n \in C_2} (\vec{w}^T \vec{x} - \vec{x}^T \vec{m}_2)^2 && (7)\\
&= \sum_{n \in C_1} \vec{w}^T (\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_1)^T \vec{w} + \sum_{n \in C_2} \vec{w}^T (\vec{x} - \vec{m}_2)(\vec{x} - \vec{m}_2)^T \vec{w} && (8)\\
&= \vec{w}^T \{ \sum_{n \in C_1} (\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_1)^T \} \vec{w} + \vec{w}^T \{ \sum_{n \in C_2} (\vec{x} - \vec{m}_2)(\vec{x} - \vec{m}_2)^T \} \vec{w} && (9)\\
&= \vec{w}^T \{ \sum_{n \in C_1} (\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_1)^T + \sum_{n \in C_2} (\vec{x} - \vec{m}_2)(\vec{x} - \vec{m}_2)^T \} \vec{w} && (10)\\
&= \vec{w}^T S_W \vec{w} && (11)
\end{aligned}$$

where $S_W = \sum_{n \in C_1} (\vec{x} - \vec{m}_1)(\vec{x} - \vec{m}_1)^T + \sum_{n \in C_2} (\vec{x} - \vec{m}_2)(\vec{x} - \vec{m}_2)^T$ and is called the within-class covariance matrix.

Putting the numerator and the denominator together, obtain:

$$J(\vec{w}) = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$$

4 Fisher's linear discriminant

Since $J(\vec{w})$ involves $\vec{w}$ in a quadratic fashion, $\frac{\partial J(\vec{w})}{\partial \vec{w}}$ involves $\vec{w}$ in a linear fashion and hence $\frac{\partial J(\vec{w})}{\partial \vec{w}} = 0$ has a unique solution for $\vec{w}$. First, let's find $\frac{\partial J(\vec{w})}{\partial \vec{w}}$.

$$J(\vec{w}) = \frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}} \quad (12)$$

In order to obtain $\frac{\partial J(\vec{w})}{\partial \vec{w}}$, we need the quotient rule for matrix calculus, which requires us to define the numerator, u, the denominator, v, the derivative of numerator, u', and the derivative of denominator, v'. Let's define them below:

$$u = \vec{w}^T S_B \vec{w}$$
$$v = \vec{w}^T S_W \vec{w}$$

According to the Matrix Cookbook,

$$u' = \frac{\partial u}{\partial \vec{w}} = (S_B + S_B^T)\vec{w}$$

Derivation of $(S_B + S_B^T)\vec{w} = 2S_B\vec{w}$:

$$\begin{aligned} S_B &= (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T \\ S_B^\top &= \left((\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T\right)^T \\ &= (\vec{m}_2 - \vec{m}_1)^{TT}(\vec{m}_2 - \vec{m}_1)^T \\ &= (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T \end{aligned}$$

Therefore, $S_B^T = S_B$ and $u' = 2S_B\vec{w}$.

Again, according to the Matrix Cookbook,

$$v' = \frac{\partial u}{\partial \vec{w}} = (S_W + S_W^T)\vec{w}$$

Derivation of $(S_W + S_W^T)\vec{w} = 2S_W\vec{w}$:

$$S_W = \sum_{n \in C_1} (\vec{x}^n - \vec{m}_1)(\vec{x}^n - \vec{m}_1)^T + \sum_{n \in C_2} (\vec{x}^n - \vec{m}_2)(\vec{x}^n - \vec{m}_2)^T$$

Since a transpose of sums is a sum of transposes,

$$\begin{aligned} S_W^T &= \sum_{n \in C_1} \left[(\vec{x}^n - \vec{m}_1)(\vec{x}^n - \vec{m}_1)^T\right]^T + \sum_{n \in C_2} \left[(\vec{x}^n - m_2)(\vec{x}^n - \vec{m}_2)^T\right]^T \\ &= \sum_{n \in C_1} (\vec{x}^n - \vec{m}_1)(\vec{x}^n - \vec{m}_1)^T + \sum_{n \in C_2} (\vec{x}^n - \vec{m}_2)(\vec{x}^n - \vec{m}_2)^T \end{aligned}$$

Therefore, $S_W^T = S_W$ and $u' = 2S_W\vec{w}$.

Using u, u', v and v' and the quotient rule to find $\frac{\partial J(\vec{w})}{\partial \vec{w}}$:

$$\frac{\partial J(\vec{w})}{\partial \vec{w}} = \frac{u'v - uv'}{v^2} \tag{13}$$

$$= \frac{(2S_B\vec{w})\left(\vec{w}^\top S_W \vec{w}\right) - \left(\vec{w}^\top S_B \vec{w}\right)(2S_w\vec{w})}{(\vec{w}^\top S_W \vec{w})\,(\vec{w}^\top S_W \vec{w})} \tag{14}$$

$$= \frac{2S_B\vec{w}}{\vec{w}^\top S_W \vec{w}} - \frac{\left(\vec{w}^\top S_B \vec{w}\right)(2S_W\vec{w})}{(\vec{w}^\top S_W \vec{w})\,(\vec{w}^\top S_W \vec{w})} \tag{15}$$

Setting $\frac{\partial J(\vec{w})}{\partial \vec{w}} = 0$ and solving for this equation:

$$\frac{\partial J(\vec{w})}{\partial \vec{w}} = 0 \tag{16}$$

$$\frac{2S_B\vec{w}}{\vec{w}^T S_w \vec{w}} = \frac{\left(\vec{w}^T S_B \vec{w}\right)(2S_w\vec{w})}{(\vec{w}^T S_w \vec{w})\,(\vec{w}^T S_w \vec{w})} \tag{17}$$

$$\underbrace{2\left(\vec{w}^\top S_w \vec{w}\right)}_{\text{scalar}}(S_B\vec{w}) = \underbrace{2\left(\vec{w}^\top S_B \vec{w}\right)}_{\text{scalar}}(S_w\vec{w}) \tag{18}$$

Since we do not care about the magnitude of $\vec{w}$, only its direction, we remove the scalars:

$$S_B\vec{w} \propto S_W\vec{w}.$$

Since $S_B = (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T$ and $(\vec{m}_2 - \vec{m}_1)^T\vec{w}$ is a scalar, $S_B\vec{w}$ is always in the direction of $\vec{m}_2 - \vec{m}_1$:

$$(\vec{m}_2 - \vec{m}_1) \propto S_W\vec{w}$$

Multiplying both sides by S_W^{-1}:

$$\vec{w} \propto S_W^{-1}(\vec{m}_2 - \vec{m}_1)$$

"This choice of $\vec{w}$ is known as Fisher's linear discriminant, although it is strictly not a discriminant but rather a specific choice of direction for projection of data down to one dimension." (Bishop, 1995)