THU-70250043-0, Pattern Recognition (Spring 2021) **Homework: 7**

Feature Extraction and Selection

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Hong Zhao vzhao@tsinghua.edu.cn

Student: Jingxuan Yang yangjx20@mails.tsinghua.edu.cn

Fisher Criterion

1. It's interesting to see that under some circumstances, the Fisher criterion can be obtained as a special case of the least squares. Consider the binary classification problem, let's unify the expression at the very beginning for convenience of the following steps, and you are required to obey the notations given below.

Suppose we have N_1 points of class $\mathcal{C}_1$ and N_2 of class $\mathcal{C}_2$, then the mean vectors of the two classes are given by

$$\boldsymbol{m}_1 = \frac{1}{N_1}\sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n, \quad \boldsymbol{m}_2 = \frac{1}{N_2}\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n. \tag{1}$$

In the lecture notes, we have defined *between-class* covariance matrix and *within-class* covariance matrix

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top, \tag{2}$$

$$S_W = \sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^\top + \sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^\top. \tag{3}$$

Now, let's turn to the least square problem. We take the targets for $\mathcal{C}_1$ to be N/N_1 and $\mathcal{C}_2$ to be $-N/N_2$ where $N = N_1 + N_2$ (This may be a little confusing, but you will see the reasons of doing so in a short time). Then the sum-of-square error function can be written as

$$E = \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{w}^\top\boldsymbol{x}_n + w_0 - t_n)^2, \tag{4}$$

where, $(\boldsymbol{x}_n, t_n)$ are the points we have. Target t_n equals to N/N_1 or $-N/N_2$ according to its class. Our goal is to estimate $\boldsymbol{w}$ and w_0.

1.1. Show that the optimal w_0 is $w_0 = -\boldsymbol{w}^\top\boldsymbol{m}$, where

$$\boldsymbol{m} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n. \tag{5}$$

解: 最小二乘表达式对 w_0 求导得

$$\frac{\partial E}{\partial w_0} = \sum_{n=1}^{N} (\boldsymbol{w}^\top \boldsymbol{x}_n + w_0 - t_n) \tag{6}$$

令此偏导数为 0, 有

$$\begin{aligned} w_0 &= -\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}^\top \boldsymbol{x}_n - t_n) \\ &= -\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{w}^\top \boldsymbol{x}_n + \frac{1}{N}\sum_{n\in\mathcal{C}_1} t_n + \frac{1}{N}\sum_{n\in\mathcal{C}_2} t_n \\ &= -\boldsymbol{w}^\top \boldsymbol{m} + \frac{1}{N}\sum_{n\in\mathcal{C}_1}\frac{N}{N_1} - \frac{1}{N}\sum_{n\in\mathcal{C}_2}\frac{N}{N_2} \\ &= -\boldsymbol{w}^\top \boldsymbol{m} + 1 - 1 \\ &= -\boldsymbol{w}^\top \boldsymbol{m} \end{aligned} \tag{7}$$

1.2. Derive the equation that the optimal $\boldsymbol{w}$ should obey

$$\left(S_W + \frac{N_1 N_2}{N} S_B\right)\boldsymbol{w} = N(\boldsymbol{m}_1 - \boldsymbol{m}_2). \tag{8}$$

解: 由 $w_0 = -\boldsymbol{w}^\top \boldsymbol{m}$ 以及 $N\boldsymbol{m} = N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2$ 可知最小二乘表达式对 $\boldsymbol{w}$ 求导为

$$\begin{aligned} \frac{\partial E}{\partial \boldsymbol{w}} &= \sum_{n=1}^{N}\boldsymbol{x}_n(\boldsymbol{w}^\top \boldsymbol{x}_n + w_0 - t_n) \\ &= \sum_{n=1}^{N}\boldsymbol{x}_n(\boldsymbol{w}^\top \boldsymbol{x}_n - \boldsymbol{w}^\top \boldsymbol{m}) - \sum_{n\in\mathcal{C}_1} t_n\boldsymbol{x}_n - \sum_{n\in\mathcal{C}_2} t_n\boldsymbol{x}_n \\ &= \sum_{n=1}^{N}\boldsymbol{x}_n(\boldsymbol{x}_n^\top \boldsymbol{w} - \boldsymbol{m}^\top \boldsymbol{w}) - \sum_{n\in\mathcal{C}_1}\frac{N}{N_1}\boldsymbol{x}_n + \sum_{n\in\mathcal{C}_2}\frac{N}{N_2}\boldsymbol{x}_n \\ &= \sum_{n=1}^{N}(\boldsymbol{x}_n\boldsymbol{x}_n^\top - \boldsymbol{x}_n\boldsymbol{m}^\top)\boldsymbol{w} - N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \\ &= \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top - \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{m}^\top\right)\boldsymbol{w} - N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \\ &= \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top - N\boldsymbol{m}\boldsymbol{m}^\top\right)\boldsymbol{w} - N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \\ &= \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top - \frac{N\boldsymbol{m}(N\boldsymbol{m})^\top}{N}\right)\boldsymbol{w} - N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \\ &= \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top - \frac{(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1 + N_2\boldsymbol{m}_2)^\top}{N}\right)\boldsymbol{w} - N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \end{aligned} \tag{9}$$

又由 S_W 定义可知

$$\begin{aligned}
S_W &= \sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n-\boldsymbol{m}_1)(\boldsymbol{x}_n-\boldsymbol{m}_1)^\top+\sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n-\boldsymbol{m}_2)(\boldsymbol{x}_n-\boldsymbol{m}_2)^\top\\
&= \sum_{n\in\mathcal{C}_1}(\boldsymbol{x}_n\boldsymbol{x}_n^\top-\boldsymbol{x}_n\boldsymbol{m}_1^\top-\boldsymbol{m}_1\boldsymbol{x}_n^\top+\boldsymbol{m}_1\boldsymbol{m}_1^\top)+\sum_{n\in\mathcal{C}_2}(\boldsymbol{x}_n\boldsymbol{x}_n^\top-\boldsymbol{x}_n\boldsymbol{m}_2^\top-\boldsymbol{m}_2\boldsymbol{x}_n^\top+\boldsymbol{m}_2\boldsymbol{m}_2^\top)\\
&= \sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n\boldsymbol{x}_n^\top-\sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n\boldsymbol{m}_1^\top-\sum_{n\in\mathcal{C}_1}\boldsymbol{m}_1\boldsymbol{x}_n^\top+N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n\boldsymbol{x}_n^\top-\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n\boldsymbol{m}_2^\top-\sum_{n\in\mathcal{C}_2}\boldsymbol{m}_2\boldsymbol{x}_n^\top+N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top\\
&= \sum_{n\in\mathcal{C}_1}\boldsymbol{x}_n\boldsymbol{x}_n^\top-N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top-N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+\sum_{n\in\mathcal{C}_2}\boldsymbol{x}_n\boldsymbol{x}_n^\top-N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top-N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top+N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top\\
&= \sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top-N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top-N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top
\end{aligned}\tag{10}$$

即

$$\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top=S_W+N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top\tag{11}$$

代入偏导数表达式可得

$$\begin{aligned}
\frac{\partial E}{\partial\boldsymbol{w}} &= \left(\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top-\frac{(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)^\top}{N}\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top-\frac{(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)^\top}{N}\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+\frac{NN_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+NN_2\boldsymbol{m}_2\boldsymbol{m}_2^\top-(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)(N_1\boldsymbol{m}_1+N_2\boldsymbol{m}_2)^\top}{N}\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+\frac{N_2N_1\boldsymbol{m}_1\boldsymbol{m}_1^\top+N_1N_2\boldsymbol{m}_2\boldsymbol{m}_2^\top-N_1N_2\boldsymbol{m}_1\boldsymbol{m}_2^\top-N_2N_1\boldsymbol{m}_2\boldsymbol{m}_1^\top}{N}\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+\frac{N_1N_2}{N}(\boldsymbol{m}_1\boldsymbol{m}_1^\top-\boldsymbol{m}_1\boldsymbol{m}_2^\top-\boldsymbol{m}_2\boldsymbol{m}_1^\top+\boldsymbol{m}_2\boldsymbol{m}_2^\top)\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+\frac{N_1N_2}{N}(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^\top\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)\\
&= \left(S_W+\frac{N_1N_2}{N}S_B\right)\boldsymbol{w}-N(\boldsymbol{m}_1-\boldsymbol{m}_2)
\end{aligned}\tag{12}$$

令此偏导数为 0, 则有

$$\left(S_W+\frac{N_1N_2}{N}S_B\right)\boldsymbol{w}=N(\boldsymbol{m}_1-\boldsymbol{m}_2)\tag{13}$$

1.3. Show that $\boldsymbol{w}$ satisfies: $\boldsymbol{w}\propto S_W^{-1}(\boldsymbol{m}_2-\boldsymbol{m}_1)$ from equation (8), which means we've got the same form as that of Fisher criterion.

解: 记

$$R\triangleq(\boldsymbol{m}_2-\boldsymbol{m}_1)^\top\boldsymbol{w}\tag{14}$$

由式 (8) 可知

$$S_W\boldsymbol{w} + \frac{N_1N_2}{N}S_B\boldsymbol{w} = -N(\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{15}$$

移项并代入 S_B 定义可得

$$\begin{aligned} S_W\boldsymbol{w} &= -N(\boldsymbol{m}_2 - \boldsymbol{m}_1) - \frac{N_1N_2}{N}S_B\boldsymbol{w} \\ &= -N(\boldsymbol{m}_2 - \boldsymbol{m}_1) - \frac{N_1N_2}{N}(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top\boldsymbol{w} \\ &= -N(\boldsymbol{m}_2 - \boldsymbol{m}_1) - \frac{N_1N_2}{N}(\boldsymbol{m}_2 - \boldsymbol{m}_1)R \\ &= \left(-N - \frac{RN_1N_2}{N}\right)(\boldsymbol{m}_2 - \boldsymbol{m}_1) \end{aligned} \tag{16}$$

假设矩阵 S_W 可逆, 则有

$$\begin{aligned} \boldsymbol{w} &= \left(-N - \frac{RN_1N_2}{N}\right)S_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \\ &\propto S_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \end{aligned} \tag{17}$$

Linear Discriminant Analysis (LDA)

2. Consider the generalization of Fisher discriminant to $K > 2$ classes, and assume that the dimensionality of the input space is greater than the number K of classes. Next, we introduce linear features $y_k = \boldsymbol{w}_k^\top\boldsymbol{x}$. The weight vectors $\{\boldsymbol{w}_k\}$ can be considered to be the columns of a matrix $\boldsymbol{W}$, so that:

$$\boldsymbol{y} = \boldsymbol{W}^\top\boldsymbol{x}, \tag{18}$$

where, $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^{D'}$. By this means, we have projected the D-dimensional $\boldsymbol{x}$-space onto the D'-dimensional $\boldsymbol{y}$-space, in which we can better separate the data.

The generalization of the *within-class* covariance matrix to the case of K classes is:

$$S_W = \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top, \tag{19}$$

where

$$\boldsymbol{m}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\boldsymbol{x}_n. \tag{20}$$

The total covariance matrix is:

$$S_T = \sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^\top, \tag{21}$$

where

$$\boldsymbol{m} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n = \frac{1}{N}\sum_{k=1}^{K}N_k\boldsymbol{m}_k. \tag{22}$$

2.1. Decompose the total covariance matrix S_T into *within-class* covariance matrix S_W and *between-class* covariance matrix S_B, and show that S_B has the form:

$$S_B = \sum_{k=1}^{K} N_k(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top. \tag{23}$$

解: 由矩阵 S_T 定义可得

$$\begin{aligned}
S_T &= \sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^\top \\
&= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^\top \\
&= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{x}_n - \boldsymbol{m}_k + \boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m}_k + \boldsymbol{m}_k - \boldsymbol{m})^\top \\
&= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}\Big[(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top + (\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{m}_k - \boldsymbol{m})^\top + (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top \\
&\qquad\qquad + (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top\Big] \\
&= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top + \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}\Big[(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{m}_k - \boldsymbol{m})^\top + (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top\Big] \\
&\quad + \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top \\
&= S_W + \sum_{k=1}^{K}\left\{\sum_{n\in\mathcal{C}_k}\Big[(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{m}_k - \boldsymbol{m})^\top + (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top\Big]\right\} + S_B \\
&= S_W + \sum_{k=1}^{K}\Big[(N_k\boldsymbol{m}_k - N_k\boldsymbol{m}_k)(\boldsymbol{m}_k - \boldsymbol{m})^\top + (\boldsymbol{m}_k - \boldsymbol{m})(N_k\boldsymbol{m}_k - N_k\boldsymbol{m}_k)^\top\Big] + S_B \\
&= S_W + \sum_{k=1}^{K}\Big[\mathbf{0}(\boldsymbol{m}_k - \boldsymbol{m})^\top + (\boldsymbol{m}_k - \boldsymbol{m})\mathbf{0}^\top\Big] + S_B \\
&= S_W + S_B
\end{aligned} \tag{24}$$

其中类间协方差矩阵为

$$\begin{aligned}
S_B &= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k}(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top \\
&= \sum_{k=1}^{K} N_k(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top
\end{aligned} \tag{25}$$

2.2. Write down the *within-class* covariance matrix s_W and *between-class* covariance matrix s_B of the projected D'-dimensional $\boldsymbol{y}$-space.

解: 在 $\boldsymbol{y}$ 空间中, $\boldsymbol{y}_n = \boldsymbol{W}^\top \boldsymbol{x}_n$, $\forall\ n = 1, 2, \ldots, N$, 且

$$\tilde{\boldsymbol{m}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{y}_n = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{W}^\top \boldsymbol{x}_n = \boldsymbol{W}^\top \boldsymbol{m} \tag{26}$$

$$\tilde{\boldsymbol{m}}_k = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k} \boldsymbol{y}_n = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k} \boldsymbol{W}^\top \boldsymbol{x}_n = \boldsymbol{W}^\top \boldsymbol{m}_k, \quad \forall\ k = 1, 2, \ldots, K \tag{27}$$

故类内协方差矩阵为

$$\begin{aligned} s_W &= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k} (\boldsymbol{y}_n - \tilde{\boldsymbol{m}}_k)(\boldsymbol{y}_n - \tilde{\boldsymbol{m}}_k)^\top \\ &= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k} (\boldsymbol{W}^\top \boldsymbol{x}_n - \boldsymbol{W}^\top \boldsymbol{m}_k)(\boldsymbol{W}^\top \boldsymbol{x}_n - \boldsymbol{W}^\top \boldsymbol{m}_k)^\top \\ &= \sum_{k=1}^{K}\sum_{n\in\mathcal{C}_k} \boldsymbol{W}^\top (\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top \boldsymbol{W} \\ &= \boldsymbol{W}^\top S_W \boldsymbol{W} \end{aligned} \tag{28}$$

类间协方差矩阵为

$$\begin{aligned} s_B &= \sum_{k=1}^{K} N_k (\tilde{\boldsymbol{m}}_k - \tilde{\boldsymbol{m}})(\tilde{\boldsymbol{m}}_k - \tilde{\boldsymbol{m}})^\top \\ &= \sum_{k=1}^{K} N_k (\boldsymbol{W}^\top \boldsymbol{m}_k - \boldsymbol{W}^\top \boldsymbol{m})(\boldsymbol{W}^\top \boldsymbol{m}_k - \boldsymbol{W}^\top \boldsymbol{m})^\top \\ &= \sum_{k=1}^{K} N_k \boldsymbol{W}^\top (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top \boldsymbol{W} \\ &= \boldsymbol{W}^\top S_B \boldsymbol{W} \end{aligned} \tag{29}$$

2.3. Many possible choices of criterion can be implemented to decide the weight matrix $\boldsymbol{W}$, more than 5 examples are shown in lecture slide. Here, we are using another criterion

$$J(\boldsymbol{W}) = \frac{\prod_{\text{diag}} s_B}{\prod_{\text{diag}} s_W}, \tag{30}$$

where, $\prod_{\text{diag}} \boldsymbol{A}$ means multiplication of the diagonal elements of matrix $\boldsymbol{A}$. Represent $J(\boldsymbol{W})$ explicitly with $\boldsymbol{W}$, S_W and S_B.

解: 由上题可得

$$J(\boldsymbol{W}) = \frac{\prod_{\text{diag}} s_B}{\prod_{\text{diag}} s_W} = \frac{\prod_{\text{diag}} \boldsymbol{W}^\top S_B \boldsymbol{W}}{\prod_{\text{diag}} \boldsymbol{W}^\top S_W \boldsymbol{W}} = \frac{\displaystyle\prod_{k=1}^{D'} \boldsymbol{w}_k^\top S_B \boldsymbol{w}_k}{\displaystyle\prod_{k=1}^{D'} \boldsymbol{w}_k^\top S_W \boldsymbol{w}_k} = \prod_{k=1}^{D'} \frac{\boldsymbol{w}_k^\top S_B \boldsymbol{w}_k}{\boldsymbol{w}_k^\top S_W \boldsymbol{w}_k} \tag{31}$$

2.4. As is stated above, we now want to project the original data space onto a space with D' dimensions, while at the same time trying to maximize $J(\boldsymbol{W})$ represented by equation (30). Write down the equations that columns of weight matrix $\boldsymbol{W}$ should obey (which means the selected projection directions).

解: 记

$$J_k(\boldsymbol{w}_k) \triangleq \frac{\boldsymbol{w}_k^\top S_B \boldsymbol{w}_k}{\boldsymbol{w}_k^\top S_W \boldsymbol{w}_k}, \quad \forall\ k = 1, 2, \ldots, D' \tag{32}$$

由上题可得

$$J(\boldsymbol{W}) = \prod_{k=1}^{D'} \frac{\boldsymbol{w}_k^\top S_B \boldsymbol{w}_k}{\boldsymbol{w}_k^\top S_W \boldsymbol{w}_k} = \prod_{k=1}^{D'} J_k(\boldsymbol{w}_k) \tag{33}$$

则

$$\max_{\boldsymbol{W}} J(\boldsymbol{W}) \Leftrightarrow \max_{\boldsymbol{w}_k} J_k(\boldsymbol{w}_k) = \frac{\boldsymbol{w}_k^\top S_B \boldsymbol{w}_k}{\boldsymbol{w}_k^\top S_W \boldsymbol{w}_k}, \quad k = 1, 2, \ldots, D' \tag{34}$$

由于 $J_k(\boldsymbol{w}_k)$ 为广义 Rayleigh 商, 则其优化问题可以转换为

$$\begin{aligned} &\max_{\boldsymbol{w}_k}\ \ \boldsymbol{w}_k^\top S_B \boldsymbol{w}_k \\ &\text{s.t.} \quad \boldsymbol{w}_k^\top S_W \boldsymbol{w}_k = c \neq 0 \end{aligned} \tag{35}$$

引入 Lagrange 乘子 λ_k 则有 Lagrange 函数

$$L(\boldsymbol{w}_k, \lambda_k) = \boldsymbol{w}_k^\top S_B \boldsymbol{w}_k - \lambda_k(\boldsymbol{w}_k^\top S_W \boldsymbol{w}_k - c) \tag{36}$$

其对 $\boldsymbol{w}_k$ 的偏导数为

$$\frac{\partial L(\boldsymbol{w}_k, \lambda_k)}{\partial \boldsymbol{w}_k} = 2S_B \boldsymbol{w}_k - 2\lambda_k S_W \boldsymbol{w}_k \tag{37}$$

令此偏导数为 0, 可知 $\boldsymbol{w}_k$ 需满足

$$S_B \boldsymbol{w}_k - \lambda_k S_W \boldsymbol{w}_k = 0 \tag{38}$$

若矩阵 S_W 可逆, 则有

$$S_W^{-1} S_B \boldsymbol{w}_k = \lambda_k \boldsymbol{w}_k \tag{39}$$

即此时 $\boldsymbol{w}_k$ 是矩阵 $S_W^{-1} S_B$ 的特征向量.

2.5. As is stated in the problem, we have K classes in all, and we are trying to find linear features (or projection directions) by maximizing $J(\boldsymbol{W})$. How many such features at most are we able to find? Explain your reason.

解: 最多可取 $K-1$ 个特征, 证明如下. 由上题可知 $\boldsymbol{w}_k$ 是矩阵 $S_W^{-1} S_B$ 的特征向量, 当 $\boldsymbol{w}_k$ 取线性无关的特征向量时, 最多取得的特征向量数为 $\text{rank}(S_W^{-1} S_B)$. 易知 $\text{rank}(S_W^{-1} S_B) \leqslant \text{rank}(S_B)$, 则问题转化为求矩阵 S_B 的秩的最大值.

由矩阵 S_B 定义

$$S_B = \sum_{k=1}^{K} N_k(\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^\top \tag{40}$$

可知

$$\operatorname{rank}(S_B) = \operatorname{rank}(\boldsymbol{m}_1 - \boldsymbol{m}, \boldsymbol{m}_2 - \boldsymbol{m}, \ldots, \boldsymbol{m}_K - \boldsymbol{m}) \tag{41}$$

又

$$\begin{aligned}
\sum_{k=1}^{K} N_k(\boldsymbol{m}_k - \boldsymbol{m}) &= \sum_{k=1}^{K} N_k \left(\frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \boldsymbol{x}_n - \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \right) \\
&= \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} \boldsymbol{x}_n - \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{N_k}{N} \boldsymbol{x}_n \\
&= \sum_{n=1}^{N} \boldsymbol{x}_n - \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{x}_n \\
&= \sum_{n=1}^{N} \boldsymbol{x}_n - \sum_{n=1}^{N} \boldsymbol{x}_n \\
&= \mathbf{0}
\end{aligned} \tag{42}$$

则 $\{\boldsymbol{m}_1 - \boldsymbol{m}, \boldsymbol{m}_2 - \boldsymbol{m}, \ldots, \boldsymbol{m}_K - \boldsymbol{m}\}$ 线性相关, 所以

$$\operatorname{rank}(\boldsymbol{m}_1 - \boldsymbol{m}, \boldsymbol{m}_2 - \boldsymbol{m}, \ldots, \boldsymbol{m}_K - \boldsymbol{m}) \leqslant K - 1 \tag{43}$$

因此 $\operatorname{rank}(S_W^{-1} S_B) \leqslant \operatorname{rank}(S_B) \leqslant K - 1$, 即最多可选取 $K - 1$ 个特征.

Feature selection and error rate

An intuitive understanding between features and error rate.

3. Let's review the definition of binary-class Bayesian error rate at first. In classification problems, our goal is always to make as few misclassifications as possible. We need a rule that assigns each $\boldsymbol{x}$ to one of the available classes. Such a rule will divide the input space into regions $\mathcal{R}_k$ called *decision regions*, one for each class, such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$. Take binary classification as an example: A mistake occurs when an input vector belonging to class $\mathcal{C}_1$ is assigned to class $\mathcal{C}_2$ or vice versa. The error rate is then given by

$$\begin{aligned}
p(\text{mistake}) &= p(\boldsymbol{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\boldsymbol{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\boldsymbol{x}, \mathcal{C}_2) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_2} p(\boldsymbol{x}, \mathcal{C}_1) \mathrm{d}\boldsymbol{x}.
\end{aligned} \tag{44}$$

To minimize the above error rate, we have to make use of posterior distribution: If $p(\mathcal{C}_1|\boldsymbol{x}) > p(\mathcal{C}_2|\boldsymbol{x})$, then we assign that $\boldsymbol{x}$ to class $\mathcal{C}_1$, and vice versa. Thus leading to the Bayesian error rate.

3.1. Suppose we consider a K-class problem, derive the corresponding error rate as that of equation (44).

解: 令 $\mathcal{R}$ 表示输入空间 (input space), 即

$$\mathcal{R} \triangleq \bigcup_{k=1}^{K} \mathcal{R}_k \tag{45}$$

则错误率为

$$P(\text{mistake}) = \sum_{k=1}^{K} p(\boldsymbol{x} \in \mathcal{R} \setminus \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^{K} \int_{\mathcal{R}\setminus\mathcal{R}_k} p(\boldsymbol{x}, \mathcal{C}_k) \mathrm{d}\boldsymbol{x} \tag{46}$$

3.2. Let x_i, $i = 1, 2, 3$ be independent binary-valued features, and $P(x_i = 1|w_1) = \alpha_i$, $P(x_i = 1|w_2) = \beta_i$, $P(w_1) = P(w_2)$. Assume that $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3$ and $\alpha_i < \beta_i$, $\forall\, i = 1, 2, 3$. Prove that the Bayesian error rate with only one feature will satisfy $e(x_1) < e(x_2) < e(x_3)$. Give your explanation of this phenomenon based on the three features.

解: 由题意可得 $\forall\, i = 1, 2, 3$, 有

$$\begin{aligned} P(x_i = 1|w_1)P(w_1) = \alpha_i P(w_1) < P(x_i = 1|w_2)P(w_2) = \beta_i P(w_2) \\ P(x_i = 0|w_1)P(w_1) = (1-\alpha_i)P(w_1) > P(x_i = 0|w_2)P(w_2) = (1-\beta_i)P(w_2) \end{aligned} \tag{47}$$

则 $\forall\, i = 1, 2, 3$, Bayes 决策规则为

$$\begin{aligned} x_i = 1 \rightarrow x_i \in w_2 \\ x_i = 0 \rightarrow x_i \in w_1 \end{aligned} \tag{48}$$

所以错误率为

$$\begin{aligned} e(x_i) &= P(x_i = 1, w_1) + P(x_i = 0, w_2) \\ &= P(x_i = 1|w_1)P(w_1) + P(x_i = 0|w_2)P(w_2) \\ &= \frac{1}{2}\alpha_i + \frac{1}{2}(1 - \beta_i) \\ &= \frac{1}{2} - \frac{1}{2}(\beta_i - \alpha_i), \quad \forall\, i = 1, 2, 3 \end{aligned} \tag{49}$$

由 $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3$ 可知

$$e(x_1) < e(x_2) < e(x_3) \tag{50}$$

对每个特征, w_1 类的均值为

$$\mu_{i1} = 0 \cdot P(x_i = 0|w_1) + 1 \cdot P(x_i = 1|w_1) = \alpha_i, \quad \forall\, i = 1, 2, 3 \tag{51}$$

w_2 类的均值为

$$\mu_{i2} = 0 \cdot P(x_i = 0|w_2) + 1 \cdot P(x_i = 1|w_2) = \beta_i, \quad \forall\, i = 1, 2, 3 \tag{52}$$

两类的均值差为

$$d_i \triangleq \mu_{i2} - \mu_{i1} = \beta_i - \alpha_i, \quad \forall\ i = 1, 2, 3 \tag{53}$$

则由 $\beta_1 - \alpha_1 > \beta_2 - \alpha_2 > \beta_3 - \alpha_3$ 可知两类均值差

$$d_1 > d_2 > d_3 \tag{54}$$

即特征 x_1 的两类均值差最大, 特征 x_3 的两类均值差最小, 而均值差越大则在该特征上两类分离得越开, 从而使用该特征进行分类的错误率越小, 因此有 $e(x_1) < e(x_2) < e(x_3)$.

3.3. With the following parameters:

$$\alpha_1 = 0.1, \quad \alpha_2 = 0.05, \quad \alpha_3 = 0.01, \quad \beta_1 = 0.9, \quad \beta_2 = 0.8, \quad \beta_3 = 0.7, \tag{55}$$

calculate $e(x_1),\ e(x_2),\ e(x_3);\ e(x_1, x_2),\ e(x_1, x_3),\ e(x_2, x_3)$. Compare the values of different error rate and present your explanation from the view of feature selection.

解: 由上题可得

$$e(x_1) = \frac{1}{2} - \frac{1}{2}(\beta_1 - \alpha_1) = \frac{1}{2} - \frac{1}{2} \times (0.9 - 0.1) = 0.1 \tag{56}$$

$$e(x_2) = \frac{1}{2} - \frac{1}{2}(\beta_2 - \alpha_2) = \frac{1}{2} - \frac{1}{2} \times (0.8 - 0.05) = 0.125 \tag{57}$$

$$e(x_3) = \frac{1}{2} - \frac{1}{2}(\beta_3 - \alpha_3) = \frac{1}{2} - \frac{1}{2} \times (0.7 - 0.01) = 0.155 \tag{58}$$

若使用两个特征 (x_1, x_2) 进行分类, 注意到特征相互独立可得

$$\begin{gathered} P(x_1=1, x_2=1|w_1)P(w_1) = \alpha_1\alpha_2 P(w_1) < P(x_1=1, x_2=1|w_2)P(w_2) = \beta_1\beta_2 P(w_2) \\ P(x_1=1, x_2=0|w_1)P(w_1) = \alpha_1(1-\alpha_2)P(w_1) < P(x_1=1, x_2=0|w_2)P(w_2) = \beta_1(1-\beta_2)P(w_2) \\ P(x_1=0, x_2=1|w_1)P(w_1) = (1-\alpha_1)\alpha_2 P(w_1) < P(x_1=0, x_2=1|w_2)P(w_2) = (1-\beta_1)\beta_2 P(w_2) \\ P(x_1=0, x_2=0|w_1)P(w_1) = (1-\alpha_1)(1-\alpha_2)P(w_1) > P(x_1=0, x_2=0|w_2)P(w_2) = (1-\beta_1)(1-\beta_2)P(w_2) \end{gathered} \tag{59}$$

因此 Bayes 决策规则为

$$\begin{aligned} x_1 = 1,\ x_2 = 1 &\to (x_1, x_2) \in w_2 \\ x_1 = 1,\ x_2 = 0 &\to (x_1, x_2) \in w_2 \\ x_1 = 0,\ x_2 = 1 &\to (x_1, x_2) \in w_2 \\ x_1 = 0,\ x_2 = 0 &\to (x_1, x_2) \in w_1 \end{aligned} \tag{60}$$

同理可得, 使用特征 (x_1, x_2) 进行分类的 Bayes 决策规则为

$$\begin{aligned}
&x_1 = 1,\ x_3 = 1 \rightarrow (x_1, x_3) \in w_2 \\
&x_1 = 1,\ x_3 = 0 \rightarrow (x_1, x_3) \in w_2 \\
&x_1 = 0,\ x_3 = 1 \rightarrow (x_1, x_3) \in w_2 \\
&x_1 = 0,\ x_3 = 0 \rightarrow (x_1, x_3) \in w_1
\end{aligned} \tag{61}$$

以及使用特征 (x_2, x_3) 进行分类的 Bayes 决策规则为

$$\begin{aligned}
&x_2 = 1,\ x_3 = 1 \rightarrow (x_2, x_3) \in w_2 \\
&x_2 = 1,\ x_3 = 0 \rightarrow (x_2, x_3) \in w_2 \\
&x_2 = 0,\ x_3 = 1 \rightarrow (x_2, x_3) \in w_2 \\
&x_2 = 0,\ x_3 = 0 \rightarrow (x_2, x_3) \in w_1
\end{aligned} \tag{62}$$

所以错误率为

$$\begin{aligned}
e(x_1, x_2) &= P(x_1 = 1, x_2 = 1, w_1) + P(x_1 = 1, x_2 = 0, w_1) + P(x_1 = 0, x_2 = 1, w_1) + P(x_1 = 0, x_2 = 0, w_2) \\
&= [1 - P(x_1 = 0, x_2 = 0|w_1)]P(w_1) + P(x_1 = 0, x_2 = 0|w_2)P(w_2) \\
&= [1 - P(x_1 = 0|w_1)P(x_2 = 0|w_1)]P(w_1) + P(x_1 = 0|w_2)P(x_2 = 0|w_2)P(w_2) \\
&= \frac{1}{2}[1 - (1 - \alpha_1)(1 - \alpha_2)] + \frac{1}{2}(1 - \beta_1)(1 - \beta_2) \\
&= \frac{1}{2} - \frac{1}{2}(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_2 - \alpha_2) + \frac{1}{2}(\beta_1\beta_2 - \alpha_1\alpha_2) \\
&= 0.0825
\end{aligned} \tag{63}$$

同理可得

$$e(x_1, x_3) = \frac{1}{2} - \frac{1}{2}(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_3 - \alpha_3) + \frac{1}{2}(\beta_1\beta_3 - \alpha_1\alpha_3) = 0.0695 \tag{64}$$

$$e(x_2, x_3) = \frac{1}{2} - \frac{1}{2}(\beta_2 - \alpha_2) - \frac{1}{2}(\beta_3 - \alpha_3) + \frac{1}{2}(\beta_2\beta_3 - \alpha_2\alpha_3) = 0.05975 \tag{65}$$

因此不同错误率的大小关系为

$$e(x_3) > e(x_2) > e(x_1) > e(x_1, x_2) > e(x_1, x_3) > e(x_2, x_3) \tag{66}$$

从特征选择的角度来看, 根据错误率的大小关系可知若使用单个特征进行分类, 则 x_1 最好, x_2 次之, x_3 最差. 使用两个特征进行分类的错误率均要小于使用单个特征的错误率, 可知增加特征可以使错误率减小, 提高分类准确率. 但是, 取最好的两个特征 x_1, x_2 进行组合的效果却并不是最好的, 反而是错误率最高的, 说明特征之间的相互作用关系对分类错误率有着很重要的影响.

Programming: Relief

4. Please read at least one of the papers about relief [1, 2]. Then implement a relief-based feature selection method and analyze the result on the dataset in the file `watermelon_3.csv` [3]. Finally, design a classifier on the selected feature space.

解: 数据集 `watermelon_3.csv` 如表 1 所示, 其中有 8 个特征, 共 17 个样本.

表 1: 西瓜数据表格

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.46	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.36	0.37	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

应用 Relief 算法得到 8 个特征的权重为

$$\boldsymbol{w} = [-0.1275,\ 0.0294,\ -0.0294,\ 0.3235,\ 0.1275,\ -0.0588,\ 0.0001,\ 0.1495] \tag{67}$$

按照权重大小对 8 个特征进行排序为

$$w_4 > w_8 > w_5 > w_2 > w_7 > w_3 > w_6 > w_1 \tag{68}$$

取排序第一的第四个特征纹理 x_4 作为特征空间进行分类器设计, 令 $x_4 = 1, 2, 3$ 分别表示纹理清晰, 稍糊和模糊, 假设好瓜与否的先验概率相等 $P(\omega_1) = P(\omega_2)$, 其中 ω_1 表示好瓜, ω_2 表示坏瓜, 则可设计 Bayes 分类器.

由数据集可计算得

$$\begin{aligned} P(x_4=1|\omega_1)=\frac{7}{8} > P(x_4=1|\omega_2)=\frac{2}{9} \\ P(x_4=2|\omega_1)=\frac{1}{8} < P(x_4=2|\omega_2)=\frac{4}{9} \\ P(x_4=3|\omega_1)=\frac{0}{8} < P(x_4=3|\omega_2)=\frac{3}{9} \end{aligned} \tag{69}$$

则 Bayes 分类器规则为

$$\begin{aligned} x_4=1 \rightarrow x_4 \in \omega_1 \\ x_4=2 \rightarrow x_4 \in \omega_2 \\ x_4=3 \rightarrow x_4 \in \omega_2 \end{aligned} \tag{70}$$

若按照权重排序取前两个特征纹理 x_4 和含糖率 x_8 作为特征空间进行 SVM 分类器设计, 则可得到分类边界如图 1 所示, 可以看出分类边界是一竖直直线, 只有特征 x_4 起到了实际的分类作用, 与上述 Bayes 分类器等价.

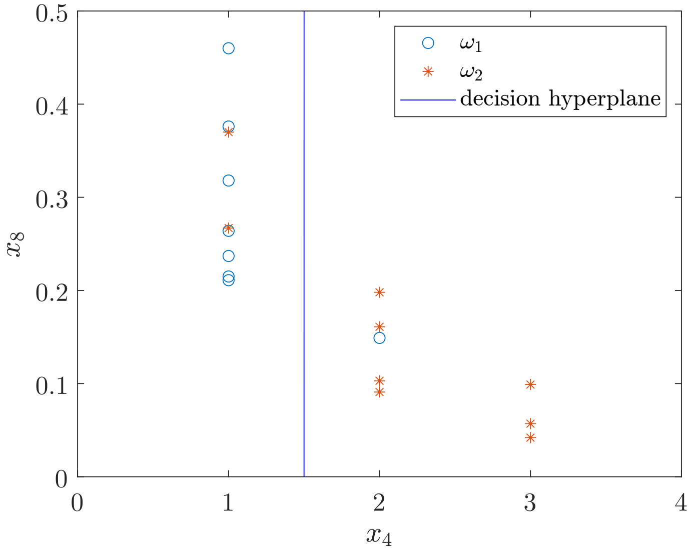

图 1: SVM 分类器

参考文献

[1] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF[J]. Machine learning, 2003, 53(1): 23-69.

[2] Urbanowicz R J, Meeker M, La Cava W, et al. Relief-based feature selection: Introduction and review[J]. Journal of biomedical informatics, 2018, 85: 189-203.

[3] Zhihua Zhou. Machine Learning[J]. 2016. ISBN 978-730-24-2327-8.