Project 4 Question 3 Data Mining

Jan C. Bierowiec

1 Question 3

Given two clusters

$C_1 = (1,1),(2,2),(3,3)$
$C_2 = (5,2),(6,2),(7,2),(8,2),(9,2)$

compute the values in (a) - (f). Use the definition for scattering criteria presented in class. Note that tr in the scattering criterion is referring to the trace of the matrix.

a) The mean vectors m_1 and m_2

Calculating the mean of m_1 and m_2:

$$m_1 = \frac{\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}2\\2\end{pmatrix}+\begin{pmatrix}3\\3\end{pmatrix}}{3} = \frac{\begin{pmatrix}6\\6\end{pmatrix}}{3} = \begin{pmatrix}2\\2\end{pmatrix}$$

$$m_2 = \frac{\begin{pmatrix}5\\2\end{pmatrix}+\begin{pmatrix}6\\2\end{pmatrix}+\begin{pmatrix}7\\2\end{pmatrix}+\begin{pmatrix}8\\2\end{pmatrix}+\begin{pmatrix}9\\2\end{pmatrix}}{5} = \frac{\begin{pmatrix}35\\10\end{pmatrix}}{5} = \begin{pmatrix}7\\2\end{pmatrix}$$

(b) The total mean vector m

$$m = \frac{\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}2\\2\end{pmatrix}+\begin{pmatrix}3\\3\end{pmatrix}+\begin{pmatrix}5\\2\end{pmatrix}+\begin{pmatrix}6\\2\end{pmatrix}+\begin{pmatrix}7\\2\end{pmatrix}+\begin{pmatrix}8\\2\end{pmatrix}+\begin{pmatrix}9\\2\end{pmatrix}}{8} = \frac{\begin{pmatrix}41\\16\end{pmatrix}}{8} = \begin{pmatrix}5.125\\2\end{pmatrix}$$

(c) The scatter matrices S_1 and S_2

$$\left[\begin{pmatrix}1\\1\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]\left[\begin{pmatrix}1\\1\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]^T=\begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}-1 & -1\end{pmatrix}=\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$
$$\left[\begin{pmatrix}2\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]\left[\begin{pmatrix}2\\2\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]^T=\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0 & 0\end{pmatrix}=\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$$
$$\left[\begin{pmatrix}3\\3\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]\left[\begin{pmatrix}3\\3\end{pmatrix}-\begin{pmatrix}2\\2\end{pmatrix}\right]^T=\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1 & 1\end{pmatrix}=\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$

$$S_1=\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}+\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}=\begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}$$

$$\left[\begin{pmatrix}5\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]\left[\begin{pmatrix}5\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]^T=\begin{pmatrix}-2\\0\end{pmatrix}(-20)=\begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}$$
$$\left[\begin{pmatrix}6\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]\left[\begin{pmatrix}6\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]^T=\begin{pmatrix}-1\\0\end{pmatrix}(-10)=\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$$
$$\left[\begin{pmatrix}7\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]\left[\begin{pmatrix}7\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]^T=\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0 & 0\end{pmatrix}=\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$$
$$\left[\begin{pmatrix}8\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]\left[\begin{pmatrix}8\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]^T=\begin{pmatrix}1\\0\end{pmatrix}(10)=\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$$
$$\left[\begin{pmatrix}9\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]\left[\begin{pmatrix}9\\2\end{pmatrix}-\begin{pmatrix}7\\2\end{pmatrix}\right]^T=\begin{pmatrix}2\\0\end{pmatrix}(20)=\begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}$$

$$S_2=\begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}+\begin{bmatrix}4 & 0\\0 & 0\end{bmatrix}=\begin{bmatrix}10 & 0\\0 & 0\end{bmatrix}$$

(d) The within-cluster scatter matrix S_W

$S_W = S_1 + S_2$

$$S_W=\begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}+\begin{bmatrix}10 & 0\\0 & 0\end{bmatrix}=\begin{bmatrix}12 & 2\\2 & 2\end{bmatrix}$$

(e) The between-cluster scatter matrix S_B

$$S_B = \sum_{n=1}^{N} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

$$S_1 = 3\left[\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}5.125\\2\end{pmatrix}\right]\left[\begin{pmatrix}2\\2\end{pmatrix} - \begin{pmatrix}5.125\\2\end{pmatrix}\right]^T$$
$$= 3\begin{pmatrix}-3.125\\0\end{pmatrix}\begin{pmatrix}-3.125 & 0\end{pmatrix} = 3\begin{bmatrix}9.766 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}29.297 & 0\\0 & 0\end{bmatrix}$$

$$S_2 = 5\left[\begin{pmatrix}7\\2\end{pmatrix} - \begin{pmatrix}5.125\\2\end{pmatrix}\right]\left[\begin{pmatrix}7\\2\end{pmatrix} - \begin{pmatrix}5.125\\2\end{pmatrix}\right]^T$$
$$= 5\begin{pmatrix}1.875\\0\end{pmatrix}\begin{pmatrix}1.875 & 0\end{pmatrix} = 5\begin{bmatrix}3.516 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}17.578 & 0\\0 & 0\end{bmatrix}$$

$$S_B = S_1 + S_2 = \begin{bmatrix}29.297 & 0\\0 & 0\end{bmatrix} + \begin{bmatrix}17.578 & 0\\0 & 0\end{bmatrix} = \begin{bmatrix}46.875 & 0\\0 & 0\end{bmatrix}$$

(f) The scatter criterion $\frac{tr(S_B)}{tr(S_W)}$

This measures how good the clustering is. (The higher value, the better)

Scatter Criterion $= \frac{tr(S_B)}{tr(S_W)} = \frac{46.875}{14} = 3.348$