课后练习5

1 问题一

1.1

我们已知

$$X = \begin{pmatrix} 3 & 4 \\ 5 & 6 \\ 1 & 2 \\ 4 & 3 \\ 2 & 5 \end{pmatrix} \quad \bar{x} = \begin{pmatrix} 3 & 4 \end{pmatrix}$$

因此

$$\hat{X} = \begin{pmatrix} 0 & 0 \\ 2 & 2 \\ -2 & -2 \\ 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$\Sigma = \frac{1}{n}\hat{X}^T\hat{X} = \begin{pmatrix} 2 & 1.2 \\ 1.2 & 2 \end{pmatrix}$$

对Σ进行特征值分解得到

$$\Sigma = W^T\Lambda W \approx \begin{pmatrix} -0.707 & 0.707 \\ 0.707 & 0.707 \end{pmatrix} \begin{pmatrix} 0.8 & 0 \\ 0 & 3.2 \end{pmatrix} \begin{pmatrix} -0.707 & 0.707 \\ 0.707 & 0.707 \end{pmatrix}$$

投影矩阵W及投影后的数据为

$$W = \begin{pmatrix} -0.707 & 0.707 \\ 0.707 & 0.707 \end{pmatrix}$$

$$X' = XW = \begin{pmatrix} 0.707 & 4.949 \\ 0.707 & 7.777 \\ 0.707 & 2.121 \\ -0.707 & 4.949 \\ -2.121 & 4.949 \end{pmatrix}$$
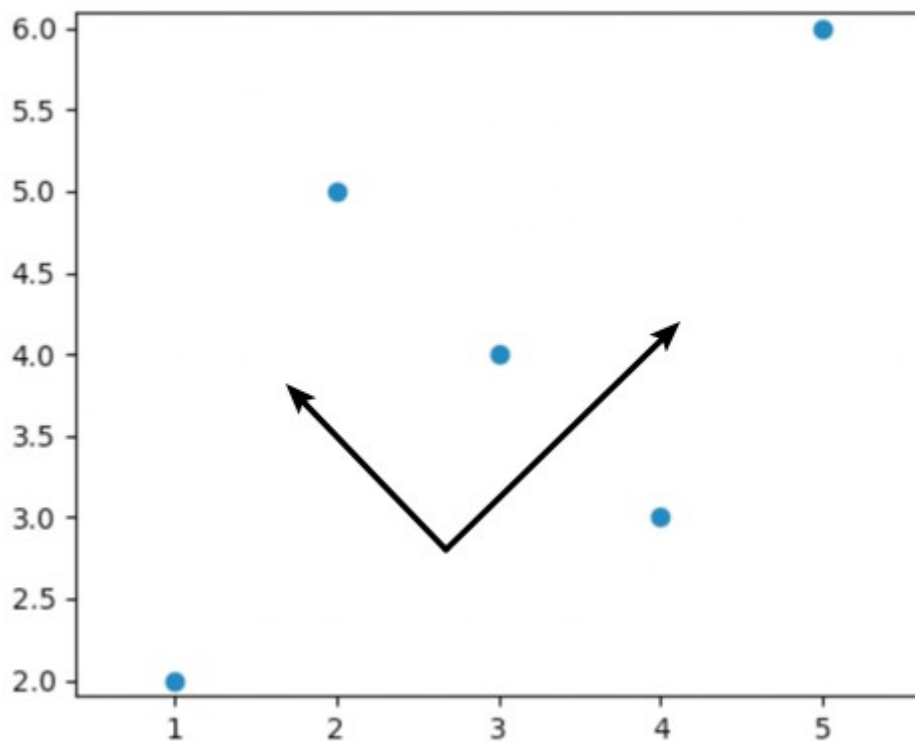
图 1: 1.1

1.2

保证不同维度正交，可以消除不同维度之间的相互影响，使得不同维度之间相互独立，互不影响

2 问题二

2.1

1. $\mu_1 = (1,1)$ $\mu_2 = (6,7)$
2. $(1,1)(1,2)(2,1)(3,4)$被分配到μ_1，$(6,7)(7,6)$被分配到μ_2
3. $\mu_1 = (1.75,2)$ $\mu_2 = (6.5,6.5)$
4. $(1,1)(1,2)(2,1)(3,4)$被分配到μ_1，$(6,7)(7,6)$被分配到μ_2
5. 收敛

2.2

1. $\mu_1 = (1,2)$ $\mu_2 = (3,4)$
2. $(1,1)(1,2)(2,1)$被分配到μ_1，$(3,4)(6,7)(7,6)$被分配到μ_2
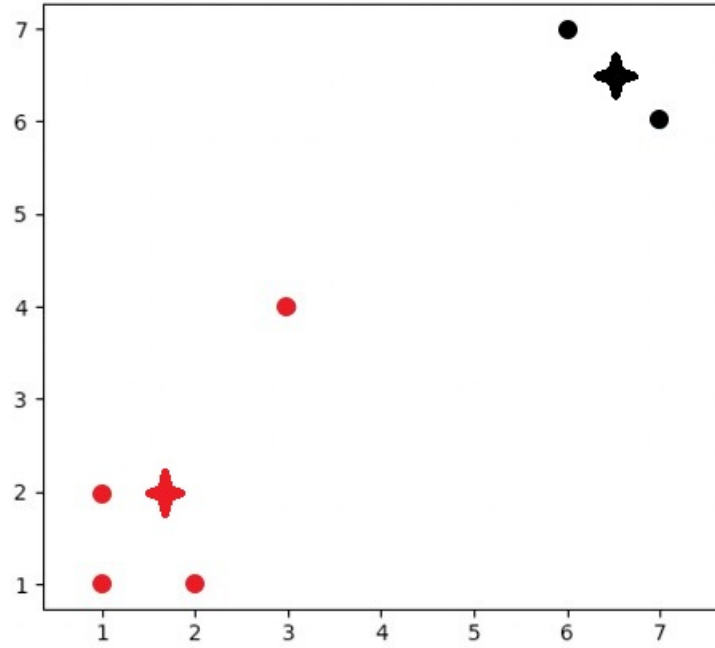3. $\mu_1 = (1.33,1.33)$ $\mu_2 = (5.33,5.67)$

图 2: 2.1

4. $(1,1)(1,2)(2,1)$ 被分配到 μ_1，$(3,4)(6,7)(7,6)$ 被分配到 μ_2

5. 收敛

2.3

初始中心点的选择会影响算法的收敛过程，使得算法陷入局部最优解而不是全局最优解，导致聚类结果的不同。图1的结果更优，因为 $(3,4)$ 确实要离 $(1,1)(1,2)(2,1)$ 这一簇更近一些

3 问题三

我们已知高斯混合模型中对软标签的更新过程为

$$\gamma_{ik} = \frac{\pi_k N(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_k, \Sigma_k)}$$
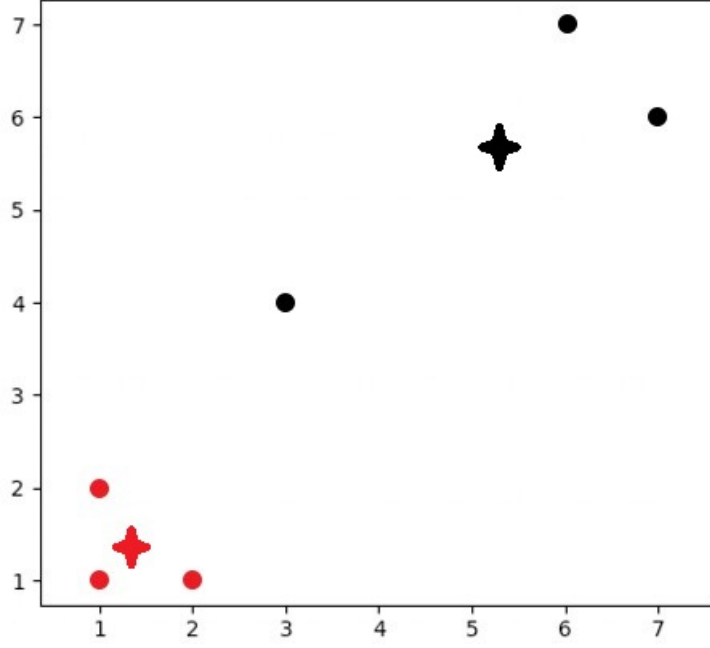
图 3: 2.2

现在$\Sigma = \epsilon I$，展开正态分布

$$\gamma_{ik} = \frac{\pi_k \frac{1}{(2\pi^{n/2}|\Sigma|^{1/2})} \exp(-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1}(x_i - \mu_k))}{\sum_{j=1}^{K} \pi_j \frac{1}{(2\pi^{n/2}|\Sigma|^{1/2})} \exp(-\frac{1}{2}(x_i - \mu_j)^T \Sigma^{-1}(x_i - \mu_j))}$$

$$= \frac{\pi_k \exp(-\frac{1}{2\epsilon}(x_i - \mu_k)^T(x_i - \mu_k))}{\sum_{j=1}^{K} \pi_j \exp(-\frac{1}{2\epsilon}(x_i - \mu_j)^T(x_i - \mu_j))}$$

$$= \frac{\pi_k \exp(-\frac{1}{2\epsilon}\|x_i - \mu_k\|^2)}{\sum_{j=1}^{K} \pi_j \exp(-\frac{1}{2\epsilon}\|x_i - \mu_j\|^2)}$$

当$\epsilon \to 0$时，对于样本点x_i，假设其属于第k类的概率最大，那么$\|x_i - \mu_k\|^2$显然要显著小于$\|x_i - \mu_j\|^2$，对任意$j \neq k$，即

$$\exp(-\frac{1}{2\epsilon}\|x_i - \mu_k\|^2) >> \exp(-\frac{1}{2\epsilon}\|x_i - \mu_j\|^2) \quad \forall j \neq k$$

这表明在分母的求和项中第k项显著大于其他项，又因为

$$\frac{\pi_j \exp(-\frac{1}{2\epsilon}\|x_i - \mu_j\|^2)}{\pi_k \exp(-\frac{1}{2\epsilon}\|x_i - \mu_k\|^2)} \begin{cases} \to 0 & j \neq k \\ = 1 & j = k \end{cases}$$

因此显然有

$$\gamma_{ik} = \begin{cases} 1 & \text{if } k = \underset{j=1,\dots,K}{argmin} \|x_i - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

现在这与K-means的硬标签完全相同，可以证明当$\epsilon \to 0$时，高斯混合模型与K-means等价

4 问题四

4.1

$$\gamma_k^{(i)} = P(z_k^{(i)} = 1 | x^{(i)}, \pi, \mathbf{p})$$

$$= \frac{P(x^{(i)} | z_k^i = 1) P(z_k^{(i)} = 1)}{\sum_{j=1}^K P(x^{(i)} | z_j^i = 1) P(z_j^{(i)} = 1)}$$

$$= \frac{\pi_k P(x^{(i)} | p^{(k)})}{\sum_{j=1}^K \pi_j P(x^{(i)} | p^{(j)})}$$

4.2

优化目标为

$$\max_{\mathbf{p},\pi} \sum_{i=1}^N \sum_{k=1}^K \gamma_k^{(i)} (\log \pi_k + \log P(x^{(i)} | p^{(k)})) \quad \text{s.t.} \sum_{k=1}^K \pi_k = 1$$

构造拉格朗日方程

$$L(\mathbf{p}, \pi, \lambda) = \sum_{i=1}^N \sum_{k=1}^K \gamma_k^{(i)} (\log \pi_k + \log P(x^{(i)} | p^{(k)})) + \lambda(1 - \sum_{k=1}^K \pi_k)$$

计算π_k

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^N \frac{\gamma_k^{(i)}}{\pi_k} - \lambda = 0$$

$$\pi_k = \frac{\sum_{i=1}^N \gamma_k^{(i)}}{\lambda}$$

由于已知

$$\lambda \sum_{k=1}^K \pi_k = \sum_{i=1}^N \sum_{k=1}^K \gamma_{ik} = N$$

$$\Rightarrow \quad \lambda = N$$

因此

$$\pi_k = \frac{\sum_{i=1}^N \gamma_k^{(i)}}{N}$$

计算$p^{(k)}$的第d个分量$p_d^{(k)}$

$$\frac{\partial L}{\partial p_d^{(k)}} = \sum_{i=1}^{N} \gamma_k^{(i)} \frac{\partial \log P(x_d^{(i)}|p_d^{(k)})}{\partial p_d^{(k)}}$$

$$= \sum_{i=1}^{N} \gamma_k^{(i)} \frac{\partial \log\left((p_d^{(k)})^{x_d^{(i)}}(1-p_d^{(k)})^{1-x_d^{(i)}}\right)}{\partial p_d^{(k)}}$$

$$= \sum_{i=1}^{N} \gamma_k^{(i)} \left(\frac{x_d^{(i)}}{p_d^{(k)}} - \frac{1-x_d^{(i)}}{1-p_d^{(k)}}\right)$$

$$= \sum_{i=1}^{N} \gamma_k^{(i)}(x_d^{(i)} - p_d^{(k)}) = 0$$

$$p_d^{(k)} = \frac{\sum_{i=1}^{n} \gamma_k^{(i)} x_d^{(i)}}{\sum_{i=1}^{n} \gamma_k^{(i)}}$$

因此$p^{(k)}$为

$$p^{(k)} = \frac{\sum_{i=1}^{n} \gamma_k^{(i)} x^{(i)}}{\sum_{i=1}^{n} \gamma_k^{(i)}}$$