hw5-ans

一、

（1）

数据中心化：

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

计算协方差矩阵：

$$Cov(X) = \frac{1}{N}\hat{X}^T\hat{X} = \begin{bmatrix} 2 & 1.2 \\ 1.2 & 2 \end{bmatrix} \tag{2}$$

通过特征值分解计算投影矩阵：

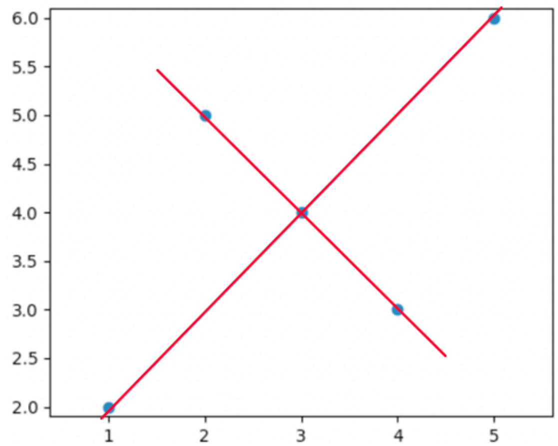
$$W = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \tag{3}$$

投影：

$$X' = \hat{X}W = \begin{bmatrix} 0 & 0 \\ \sqrt{8} & 0 \\ -\sqrt{8} & 0 \\ 0 & -\sqrt{2} \\ 0 & \sqrt{2} \end{bmatrix} \tag{4}$$

（$X' = XW$同样正确）

主成分方向如图：



（2）

由于原始数据的各特征之间可能存在相关性或冗余信息，PCA通过正交变换，将原始特征空间中具有冗余信息的特征转换为一组相互独立的特征（即主成分），有助于去除或减少冗余和噪声，使得新数据一定程度上更加满足独立性假设。

二、

（1）

First Iteration:

Centers: (1, 1), (6, 7)

Clusters: (0, 0, 0, 1, 1, 0)

New centers: (1.75, 2), (6.5, 6.5)

Second Iteration:

Centers: (1.75, 2), (6.5, 6.5)

Clusters: (0, 0, 0, 1, 1, 0)

与前一次迭代cluster相同，因此停止。

因此聚类中心为： (1.75, 2), (6.5, 6.5)，类别为： (0, 0, 0, 1, 1, 0)

（2）

First Iteration:

Centers: (1, 2), (3, 4)

Clusters: (0, 0, 1, 1, 1, 0)

New centers: (4/3, 4/3), (16/3, 17/3)

Second Iteration:

Centers: (4/3, 4/3), (16/3, 17/3)

Clusters: (0, 0, 1, 1, 1, 0)

与前一次迭代cluster相同，因此停止。

因此聚类中心为：(4/3, 4/3), (16/3, 17/3)，类别为：(0, 0, 1, 1, 1, 0)

（3）

Metric：

$$L = \sum_{i \in [n]} \sum_{k \in [K]} \gamma_{ik} ||x_i - \mu_k||^2 \tag{5}$$

K-means并非直接对该metric进行优化，而是分布进行，第一次优化cluster的分配，第二次优化聚类中心坐标，循环往复。虽然每一步都能保证对L的优化，但是不能保证优化到全局最优值。因此，不同的初值可能导致不同的极小值。

评估两次结果：

$L_1 = 9.75$
$L_2 = 14.67$

因此第一次聚类效果更好。

3. 解：令$\sum_k = \epsilon I$，这里ϵ当作一个固定的常数，而不是要重新估计的参数。已知$\epsilon \to 0$，计算后验概率有：

$$\gamma_{ik} = \frac{\pi_k N(x_i|\mu_k, \epsilon I)}{\sum_j \pi_j N(x_i|\mu_j, \epsilon I)} = \frac{\pi_k \exp\{-\frac{|x_i-\mu_k|^2}{2\epsilon}\}}{\sum_j \pi_j \exp\{-\frac{|x_i-\mu_j|^2}{2\epsilon}\}}$$

当$\epsilon \to 0$，在分母中，当 $|x_n - \mu_j|^2$ 最小时衰减地最慢，因此当$k = \mathrm{argmin}_{j\in[k]}|x_i - \mu_j|^2$时，$\gamma_{ik} = 1$；否则都为0。即对每一个数据点分类时，其被分到距离最近的cluster的可能性接近于1。另外对μ_k的估计有$\mu_k = \frac{\sum_{i\in[n]} \gamma_{ik} x_i}{\sum_{i\in[n]} \gamma_{ik}}$。即在被分到cluster k的数据点$x_i$的平均。

因此在这种极限情况下，高斯混合模型等价于K-means算法。

4. 解：（1）$\gamma_k^{(i)} = P\left(z_k^{(i)} = 1 \middle| \mathrm{x}^{(i)}, \pi, \boldsymbol{p}\right) = \frac{P(z_k^{(i)}=1, \mathrm{x}^{(i)}|\pi,\boldsymbol{p})}{P(\mathrm{x}^{(i)}|\pi,\boldsymbol{p})} = \frac{P(\mathrm{x}^{(i)}|z_k^{(i)}=1,\pi,\boldsymbol{p}) \times P(z_k^{(i)}=1|\pi,\boldsymbol{p})}{P(\mathrm{x}^{(i)}|\pi,\boldsymbol{p})}$

已知$P\left(\mathrm{x}^{(i)} \middle| z_k^{(i)} = 1, \pi, \boldsymbol{p}\right) = P(x^{(i)}|p^{(k)})$，$P\left(z_k^{(i)} = 1 \middle| \pi, \boldsymbol{p}\right) = \pi_k$，

$P\left(\mathrm{x}^{(i)} \middle| \pi, \boldsymbol{p}\right) = \sum_{k\in[K]} \pi_k P(x^{(i)}|p^{(k)})$，代入可得：

$\gamma_k^{(i)} = \frac{\pi_k P(x^{(i)}|p^{(k)})}{\sum_{k\in[K]} \pi_k P(x^{(i)}|p^{(k)})} = \frac{\pi_k P(x^{(i)}|p^{(k)})}{\sum_{k\in[K]} \pi_k P(x^{(i)}|p^{(k)})} = \frac{\pi_k P(x^{(i)}|p^{(k)})}{\sum_{k\in[K]} \pi_k P(x^{(i)}|p^{(k)})}$。

（2） 优化目标：$\max \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k^{(i)} (\log \pi_k + \log P(x^{(i)}|p^{(k)})$ s.t. $\sum_{k\in[K]} \pi_k = 1$

利用拉格朗日乘子法，且代入 $\mathrm{P}\left(x^{(i)} \middle| p^{(k)}\right) = \prod_{d=1}^{D} \left(p_d^{(k)}\right)^{x_d^{(i)}} \left(1 - p_d^{(k)}\right)^{1-x_d^{(i)}}$：

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k^{(i)} \left(\log \pi_k + \sum_{d=1}^{D} (x_d^{(i)} \log p_d^{(k)} + \left(1 - x_d^{(i)}\right) \log(1 - p_d^{(k)})\,)\right) + \lambda(1 - \sum_{k\in[K]} \pi_k)$$

对$p^{(k)}$和π_k求导：

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

$\frac{\partial L}{\partial p_d^{(k)}} = \sum_{i=1}^{N} \sum_{d=1}^{D} \gamma_k^{(i)} (\frac{x_d^{(i)}}{p_d^{(k)}} - \frac{1-x_d^{(i)}}{1-p_d^{(k)}}) = 0$，化简得到$p_d^{(k)} = \frac{\sum_{i=1}^{N} \gamma_k^{(i)} x_d^{(i)}}{\sum_{i=1}^{N} \gamma_k^{(i)}}$