STAT 510 Mathematical Statistics Fall 2022

Problem set 3: Due on 11:59pm, Friday, 9/30/2022

1. **K-component Gaussian mixture model.** Let $X_1, \ldots, X_n$ be i.i.d. p-dimensional random variables with mixture distribution of K Gaussians, i.e., the common probability density function of $X_i, i = 1, \ldots, n$ is given by

$$f(x \mid \theta) = \sum_{k=1}^{K} \frac{\pi_k}{\sqrt{\det(2\pi\Sigma_k)}} \exp\left[-\frac{1}{2}(x-\mu_1)^T \Sigma_k^{-1}(x-\mu_1)\right], \quad x \in \mathbb{R}^p,$$

where $\theta = \left\{(\pi_k, \mu_k, \Sigma_k)_{k=1}^K\right\}$ contains the unknown parameters such that $\pi_k \geq 0, \sum_{k=1}^K \pi_k = 1$, $\mu_k \in \mathbb{R}^p$, and $\Sigma_k \succ 0$ is a $p \times p$ positive-definite matrix.

(a) **E-step.** Compute the Q-function defined as $Q(\theta|\theta^{(t)}) = \mathbb{E}[\ell(\theta|X, Z)|X, \theta^{(t)}]$, where $X = (X_1, \ldots, X_n)$ and $Z = (Z_{ik})_{i=1,\ldots,n;k=1,\ldots,K}$ contains the cluster membership hidden variables (i.e., if X_i comes from component k then $Z_{ik} = 1$, and $Z_{ik} = 0$ otherwise), $\theta^{(t)}$ is the current estimate of parameter θ at iteration t, and $\ell(\theta|X, Z)$ is the complete data log-likelihood function.

(b) **M-step.** Maximize $Q(\theta|\theta^{(t)})$ over θ in the above Gaussian mixture model and obtain the explicit updating expression for $\theta^{(t+1)}$.

(c) **Implementation.** Submit your computer program code to implement the EM algorithm for the above K-component Gaussians mixture model.

(d) **Data.** Apply your EM algorithm for clustering the Iris data with $p = 4$ features and 150 data points. Output your estimated θ from the EM algorithm. Meanwhile, report the convergence of the EM algorithm, i.e., look at $\theta^{(t)} - \theta$ versus t, where θ is the true parameter. Iris data is provided in the file `Iris data.txt` and description of the Iris data is provided in `iris description.txt`.

2. **Missing data problem.** Let $X = (X_1, X_2, X_3)$ be a (three-category) multinomial random variable with the following probability mass function

$$f(x \mid \theta) = \frac{(x_1 + x_2 + x_3)!}{x_1!\, x_2!\, x_3!} \left(\frac{4+\theta}{6}\right)^{x_1} \left(\frac{1-\theta}{3}\right)^{x_2} \left(\frac{\theta}{6}\right)^{x_3}, \quad x = (x_1, x_2, x_3),$$

where $\theta \in (0, 1)$ is the unknown parameter.

(a) Derive the EM algorithm for estimating the parameter θ in the above multinomial model (that is, write down the Q-function and maximize it).

(b) Submit your computer program code to implement the EM algorithm derived in part (a). Output your estimated θ and report the convergence of the EM algorithm on the given data $X = (42, 10, 15)$.