DATA130004: Homework 10

董晴园 14300680173

2018.1.3

1. Refer to the lecture notes about mixture models. Now suppose Y follows p-component Gaussian mixture density,

$$f(y|\theta) = \sum_{r=1}^{p} \pi_r f(y|\mu_r, \sigma_r^2)$$

where $\theta = \{\pi_r, \mu_r, \sigma_r^2 : 1 \leq r \leq p\}$. Follow the steps to derive the EM algorithm.

   (a) Let U be the latent variable indicating which subpopulation Y comes from. Derive the log-likelihood function for complete data log $f(y, u|\theta)$.

   (b) The conditional distribution of $U|Y$,θ is a discrete distribution. Given another θ' which is different from θ, calculate $P(U = r|Y = y, \theta')$ and denote it by $w_r(y, \theta')$.

   (c) (E-step) When the complete data $(y_1, u_1), ..., (y_n, u_n)$ are assumed to be i.i.d., show that

$$\begin{aligned} Q(\theta, \theta') &= E\{logf(U, Y|\theta)|Y = y, \theta'\} \\ &= \sum_{r=1}^{p} \left\{ \sum_{j=1}^{n} w_r(y_j, \theta') \right\} log\pi_r + \sum_{r=1}^{p} \sum_{j=1}^{n} w_r(y_j, \theta') logf(y_j|\mu_r, \sigma_r^2) \end{aligned}$$

   (d) (M-step) Maximize $Q(\theta, \theta')$ with respect to θ when fixing θ'. Use † to indicate the maximizer. Show that

$$\begin{aligned} \pi_r^\dagger &= \frac{1}{n} \sum_{j=1}^{n} w_r(y_j, \theta') \\ \mu_r^\dagger &= \frac{\sum_{j=1}^{n} w_r(y_j, \theta') y_j}{\sum_{j=1}^{n} w_r(y_j, \theta')} \\ \sigma_r^{2\dagger} &= \frac{\sum_{j=1}^{n} w_r(y_j, \theta')(y_j - \mu_r^\dagger)^2}{\sum_{j=1}^{n} w_r(y_j, \theta')} \end{aligned}$$