Problem 2: Gaussian Mixture Model and EM (10pts+5pts Bonus)

In class, we applied EM to learn Gaussian Mixture Models (GMMs) and showed the M-Step without a proof. Now, it is time that you prove it.

Consider a GMM with the following PDF of $\mathbf{x}_i$:

$|\Sigma_j| = (\sigma_j^2)^d \cdot I \qquad -\frac{1}{2\sigma_j^2}\|x_i - \mu_j\|_2^2$

$$p(\mathbf{x}_i) = \sum_{j=1}^{k} \pi_j N(\mathbf{x}_i \mid \mu_j, \Sigma_j) = \sum_{j=1}^{k} \frac{\pi_j}{(\sqrt{2\pi})^d \underline{|\Sigma_j|^{\frac{1}{2}}}} \exp\left(\overbrace{-\frac{1}{2}(\mathbf{x}_i - \mu_j)^T \underline{\underline{\Sigma_j^{-1}}}(\mathbf{x}_i - \mu_j)}\right)$$

where k is the number of Gaussian components, d is dimension of a data point $\mathbf{x}_i$ and N is the usual Gaussian pdf ($|\Sigma|$ in the pdf denotes the determinant of matrix Σ). This GMM has k tuples of model parameters $\{(\mu_j, \Sigma_j, \pi_j)\}_{j=1}^k$, where the parameters represent the mean vector, covariance matrix, and component weight of the j-th Gaussian component. For simplicity, we further assume that all components are isotropic Gaussian, i.e., $\underline{\underline{\Sigma_j = \sigma_j^2 I}}$. $\quad \Sigma_j^{-1} = \frac{1}{\sigma_j^2} I$

2.1 (10 pts) Find the MLE of *the expected complete log-likelihood*. Equivalently, find the optimal solution to the following optimization problem.

$$\begin{aligned} \operatorname*{argmax}_{\pi_j, \mu_j, \Sigma_j} \; & \sum_i \sum_j \gamma_{ij} \ln \pi_j + \sum_i \sum_j \gamma_{ij} \ln N(\mathbf{x}_i \mid \mu_j, \Sigma_j) \\ \text{s.t. } & \pi_j \geq 0 \\ & \sum_{j=1}^{k} \pi_j = 1 \end{aligned}$$

where γ_{ij} is the posterior of latent variables computed from the E-Step.

You can use the following fact: Given $a_1, \ldots, a_k \in \mathbb{R}^+$, the solution to the following optimization problem over $q_1, \ldots, q_k$:

$$\begin{aligned} \operatorname*{argmax}_{q_j} \; & \sum_{j=1}^{k} a_j \ln q_j, \\ \text{s.t. } & q_j \geq 0, \\ & \sum_{j=1}^{k} q_j = 1. \end{aligned} \qquad a_j = \sum_i \gamma_{ij}$$

is given by:

$$q_j^* = \frac{a_j}{\sum_{k'} a_{k'}}$$

To find $\pi_1, \ldots, \pi_k$, we simply solve

$$\begin{aligned} \operatorname*{argmax}_{\boldsymbol{\pi}} \; & \sum_j \sum_i \gamma_{ij} \ln \pi_j. \\ \text{s.t. } & \pi_j \geq 0 \\ & \sum_{j=1}^{k} \pi_j = 1 \qquad \text{(2 points)} \end{aligned}$$

The solution is

$$\pi_j^* = \frac{\sum_i \gamma_{ij}}{\sum_j \sum_i \gamma_{ij}} = \frac{\sum_i \gamma_{ij}}{\sum_i 1} = \frac{\sum_i \gamma_{ij}}{n}. \qquad \text{(1 points)}$$

To find μ_j and σ_j, we solve for each j

$$\underset{\mu_j,\sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \ln N(\mathbf{x}_i \mid \mu_j, \sigma_j) = \underset{\mu_j,\sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \ln \left[\frac{1}{(\sqrt{2\pi}\sigma_j)^d} \exp\left(-\frac{1}{2\sigma_j^2} \|\mathbf{x}_i - \mu_j\|^2 \right) \right]$$

$$= \underset{\mu_j,\sigma_j}{\operatorname{argmax}} \sum_i \gamma_{ij} \left(-d \underbrace{\ln \sigma_j}_{\frac{1}{\sigma_j}} - \frac{\|\mathbf{x}_i - \mu_j\|^2}{2\sigma_j^2} \right). \qquad \underset{\mu_j}{\arg\min} \sum_i \gamma_{ij} \|x_i - \mu_j\|^2$$

$$\frac{1}{\sigma_j^2} \to \frac{-2}{\sigma_j^3}$$

(3 points)

First we set the derivative w.r.t. μ_j to 0:

$$\frac{1}{\sigma_j^2} \sum_i \gamma_{ij} (\mathbf{x}_i - \mu_j) = 0,$$

which gives

$$\mu_j^* = \frac{\sum_i \gamma_{ij} \mathbf{x}_i}{\sum_i \gamma_{ij}} \qquad \text{(2 points)}$$

Next we set the derivative w.r.t. σ_j to 0:

$$\sum_i \gamma_{ij} \left(-\frac{d}{\sigma_j} + \frac{\|\mathbf{x}_i - \mu_j\|^2}{\sigma_j^3} \right) = 0.$$

Solving for σ_j gives

$$(\sigma_j^*)^2 = \frac{\sum_i \gamma_{ij} \|\mathbf{x}_i - \mu_j^*\|^2}{d \sum_i \gamma_{ij}}. \qquad \text{(2 points)}$$

2.2 (Bonus) (5 pts) The posterior probability of z in GMM can be seen as a *soft* assignment to the clusters; in contrast, k-means assign each data point to one cluster at each iteration (*hard* assignment). Show that if we set $\{\sigma_j, \pi_j\}_{j=1}^k$ in a particular way in the GMM model, then the cluster assignments given by the GMM reduce in the limit to the k-means clusters assignment (where the cluster centers $\{\mu_j\}_{j=1}^k$ remain the same for both the models). To verify your answer, you should derive $p(z_i = j|\boldsymbol{x}_i)$ for your choice.

Set all $\sigma_j = \sigma \to 0$ and $\pi_j = \frac{1}{k}$, we have

$$p(\boldsymbol{x}_i, z_i = j) \propto \pi_j \exp\left(-\frac{1}{2\sigma_j^2} \|\mathbf{x}_i - \mu_j\|^2 \right),$$

where constant terms are ignored. (2 points)

The posterior then becomes

$$p(z_i = j|\mathbf{x}_i) = \lim_{\sigma \to 0} \frac{\pi_j \exp\{-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mu_j\|^2\}}{\sum_j \pi_j \exp\{-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mu_j\|^2\}} \to \begin{cases} 1, & \text{if } j = \arg_c \min \|\mathbf{x}_i - \mu_c\|^2 \\ 0, & \text{otherwise.} \end{cases}$$

(3 points)

$$\frac{\pi_j \sim}{\pi_j \sim + 0 \ldots} = 1$$