In The Name of Almighty

Statistical Pattern Recognition-**HW#3**

The EM algorithm for mixtures of Gaussians

Exercise 1 (E step). Consider a mixture of Gaussians with K component. Assume that we are given N data samples $\{\mathbf{x}_n\}_{n=1}^N$ and a current guess of parameters $\boldsymbol{\theta}^{\text{old}} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that

$$p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \triangleq \gamma_k(\mathbf{x}_n) \qquad \text{for all } n = 1, \ldots, N \, .$$

Exercise 2 (M step). Assume a mixture of Gaussians with K component and N data samples $\{\mathbf{x}_n\}_{n=1}^N$. The *log-likelihood* function is given as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_k(\mathbf{x}_n) \big(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) \, .$$

a) Show that the optimal choice with respect to the *mean vectors* $\boldsymbol{\mu}_k$ for all $k = 1, \ldots, K$ is given as

$$\arg\max_{\boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = \frac{\sum_{n=1}^N \gamma_k(\mathbf{x}_n) \mathbf{x}_n}{\sum_{m=1}^N \gamma_k(\mathbf{x}_m)} \, .$$

Hint: for a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A}\mathbf{x} \, .$$

b) Show that the optimal choice with respect to the *covariance matrices* $\boldsymbol{\Sigma}_k$ for all $k = 1, \ldots, K$ is given as

$$\arg\max_{\boldsymbol{\Sigma}_k} \mathcal{L}(\boldsymbol{\theta}) = \frac{\sum_{n=1}^N \gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^N \gamma_k(\mathbf{x}_m)} \, .$$

Hint: for a symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{a}^T \mathbf{X}^{-1} \mathbf{b} = -\mathbf{X}^{-T} \mathbf{a} \mathbf{b}^T \mathbf{X}^{-T} \, ,$$

and for a non-singular matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$,

$$\frac{\partial}{\partial \mathbf{X}} |\mathbf{X}| = |\mathbf{X}| \, \mathbf{X}^{-1} \, .$$