Chapter 9 Mixture Models and EM

Yue Yu

9.1 K-means Clustering

(9.5)

Recall the distortion function

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2.$$

Setting the derivative of J with respect to $\boldsymbol{\mu}_k$ to $\mathbf{0}$, we have

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} J = -2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = \mathbf{0},$$

which implies that

$$-\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = \mathbb{E}[-r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)] = \mathbf{0}.$$

Applying the Robbings-Monro algorithm while setting $a^{\text{old}} r_{nk} = \eta_n$, we obtain

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}}).$$

9.2 Mixtures of Gaussians

(9.12)

$$\begin{aligned} p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) \\ &= \sum_{k=1}^{K} p(z_k = 1) p(\mathbf{x}|z_k = 1) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \end{aligned}$$

(9.22)

As (9.21) indicates, by setting the derivative of the Lagrangian with respect to π_k to 0, we have

$$\frac{\partial}{\partial \pi_k} L = \sum_{n=1}^{N} \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda = 0.$$

Multiplying both sides by π_k and sum over k, we have

$$\begin{aligned} 0 &= \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \sum_{k=1}^{K} \lambda \pi_k \\ &= \sum_{n=1}^{N} \frac{\sum_k \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda \sum_{k=1}^{K} \pi_k \\ &= N + \lambda, \end{aligned}$$

which implies that

$$\lambda = -N.$$

Plugging it into

$$\underbrace{\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{N_k} + \lambda \pi_k = 0,$$

we obtain

$$\pi_k = \frac{N_k}{N}.$$

9.3 An Alternative View of EM

(9.33)

As is stated in the book, the EM algorithm can also be used to find MAP solutions. In this case, the log likelihood function becomes

$$\begin{aligned} \ln p(\boldsymbol{\theta}|\mathbf{X}) &= \ln \left\{ p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \right\} + \text{const} \\ &= \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + \text{const} \\ &= \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\} + \ln p(\boldsymbol{\theta}) + \text{const}. \end{aligned}$$

Hence, the corresponding expectation is given by

$$\begin{aligned} \mathcal{Q}'(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln \left\{ p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \right\} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta}) \end{aligned}$$

where $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ is defined by (9.30).

(9.39)

$$\begin{aligned} \mathbb{E}[z_{nk}] &= \sum_{\mathbf{z}_n} z_{nk} p(\mathbf{z}_n|\mathbf{x}_n) \\ &= 1 \cdot p(z_{nk} = 1|\mathbf{x}_n) \\ &= p(z_{nk} = 1|\mathbf{x}_n) \\ &= \gamma(z_{nk}). \end{aligned}$$

(9.40)

According to (9.36) and (9.39),

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] &= \mathbb{E}\Big[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}\Big] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}]\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\}.\end{aligned}$$

(9.46)

Because the variables x_i's are independent, the covariance matrix must be diagonal, with each diagonal element being $\mu_i(1-\mu_i)$.

(9.47)

We can recover (9.47) by introducing a 1-of-K binary latent variable $\mathbf{z} = [z_1, \dots z_K]^{\mathrm{T}}$. To be more specific,

$$\begin{aligned}p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi}) &= \sum_{\mathbf{z}} p(\mathbf{z}|\boldsymbol{\pi})p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu}) \\ &= \sum_{k=1}^{K} p(z_k = 1|\boldsymbol{\pi})p(\mathbf{x}|z_k = 1,\boldsymbol{\mu}) \\ &= \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k).\end{aligned}$$

(9.49)

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi}) \\ &= \sum_{\mathbf{x}} \mathbf{x} \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \\ &= \sum_{k=1}^{K} \pi_k \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}|\boldsymbol{\mu}_k) \\ &= \sum_{k=1}^{K} \pi_k \mathbb{E}_k[\mathbf{x}] \\ &= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k.\end{aligned}$$

(9.50)

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] &= \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^{\mathrm{T}} p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\pi}) \\ &= \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^{\mathrm{T}} \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \\ &= \sum_{k=1}^{K} \pi_k \sum_{\mathbf{x}} \mathbf{x}\mathbf{x}^{\mathrm{T}} p(\mathbf{x}|\boldsymbol{\mu}_k) \\ &= \sum_{k=1}^{K} \pi_k \mathbb{E}_k[\mathbf{x}\mathbf{x}^{\mathrm{T}}] \\ &= \sum_{k=1}^{K} \pi_k (\mathrm{cov}_k[\mathbf{x}] + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}) \\ &= \sum_{k=1}^{K} \pi_k \big\{ \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}} \big\}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{cov}[\mathbf{x}] &= \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}} \\ &= \sum_{k=1}^{K} \pi_k \big\{ \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}} \big\} - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^{\mathrm{T}}.\end{aligned}$$

(9.59)

Setting the derivative of the expectation of the log likelihood with respect to μ_{ki} to 0, we have

$$\frac{\partial}{\partial \mu_{ki}} \mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N} \gamma(z_{nk}) \left\{ x_{ni}\frac{1}{\mu_{ki}} - (1-x_{ni})\frac{1}{1-x_{ni}} \right\} = 0.$$

Solving for μ_{ki}, we obtain

$$\begin{aligned}\mu_{ki} &= \frac{1}{\sum_{n=1}^{N}\gamma(z_{nk})} \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} \\ &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni}.\end{aligned}$$

Therefore,

$$\begin{aligned}\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\ &= \overline{\mathbf{x}}_k.\end{aligned}$$

(9.60)

Making use of the constraint $\sum_k \pi_k = 1$, the Lagrangian of the expectation of log likelihood is given by

$$L(\pi_k, \lambda) = \mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] + \lambda\bigg(\sum_{k=1}^{K} \pi_k - 1 \bigg).$$

Setting the derivative with respect to π_k to 0, we have

$$\frac{\partial}{\partial \pi_k} L(\pi_k, \lambda) = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{\pi_k} + \lambda = 0. \tag{*}$$

By multiplying π_k on both sides and summing over k, the equation can be reduced by

$$\begin{aligned}
0 &= \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) + \lambda \sum_{k=1}^{K} \pi_k \\
&= \sum_{k=1}^{K} N_k + \lambda \\
&= N + \lambda
\end{aligned}$$

where in the second step we used the constraint $\sum_{k=1}^{K} \pi_k = 1$. Hence,

$$\lambda = -N.$$

Plugging back into (*) and solving for π_k, we obtain

$$\begin{aligned}
\pi_k &= \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \\
&= \frac{N_k}{N}.
\end{aligned}$$

(9.63)

$$\begin{aligned}
\alpha &= \arg_\alpha \left(\frac{\partial}{\partial \alpha} \mathbb{E}[\ln p(\mathbf{t}, \mathbf{w} | \alpha, \beta)] = 0 \right) \\
&= \arg_\alpha \left(\frac{M}{2} \cdot \frac{1}{2\pi} \cdot \frac{2\pi}{\alpha} - \frac{1}{2} \mathbb{E}[\mathbf{w}^{\mathrm{T}} \mathbf{w}] = 0 \right) \\
&= \frac{M}{\mathbb{E}[\mathbf{w}^{\mathrm{T}} \mathbf{w}]} \\
&= \frac{M}{\mathbf{m}_N^{\mathrm{T}} \mathbf{m}_N + \mathbf{Tr}(\mathbf{S}_N)}
\end{aligned}$$

where in the last step we used the property that if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$, then $\mathbb{E}[\mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}] = \mathrm{Tr}(\mathbf{A}\mathbf{\Sigma}) + \mathbf{m}^{\mathrm{T}} \mathbf{A} \mathbf{m}$.

9.4 The EM Algorithm in General

N/A