PRML chapter9

1. K-means clustering
   - our purpose is to deivide $\{x_1, x_2, \dots, x_N\}$ into K groups
   - Objective function is
   $$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \boldsymbol{\mu}_k \right\|^2$$
   - $\mu_k$is a center of the cluster, and r_{nk} is what cluster $\mathbf{x}_n$ in
   - First we initialize μ_k, then minimize J with respect to the r_{nk}, keeping the μ_k fixed. And then minimize J with respect to the μ_k, keeping r_{nk} fixed.
   - J can be written ad follows
   $$J = \sum_{k=1}^{K} r_{1k} \left\| x_1 - \mu_k \right\|^2 + \cdots + \sum_{k=1}^{K} r_{Nk} \left\| x_N - \mu_k \right\|^2$$
   $$r_{nk} = \begin{cases} 1 & k = \arg\min_j \left\| x_n - \mu_j \right\|^2 \\ 0 & \end{cases}$$
   - minimize J with respect to the μ_k, keeping r_{nk} fixed.
   $$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\sum_{n=1}^{N} r_{nk} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_k \right\|^2 \right) = 2 \sum_{n=1}^{N} r_{nk} \left(\boldsymbol{x}_n - \boldsymbol{\mu}_k \right) = 0$$
   - gain
   $$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$
   - K-medoids algorithm is to make objective function
   $$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}\left(\mathbf{x}_n, \boldsymbol{\mu}_k \right)$$
2. Mixtures of Gaussians
   - the conditional distribution of x given a particular value for z is
   $$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$
   -
   $$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- we are able to work with the joint distribution p(x, z) instead of the marginal distribution p(x) which is written in chapter 2.9

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} = \prod_{k=1}^{K} (\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_k}$$

- the conditional probability of z given x plays an important role

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

(a) Maximum likelihood

- the log of the likelihood function is

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- there is a problem that distribution mighht be too shape aroud the mean to cause over-fitting → Baysean approach prevent it

(b) EM for Gaussian mixtures

-

$$0 = \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) = \sum_{n=1}^{N} \frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k)}{\sum_{n=1}^{K} \pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}$$

- then we gain

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$
$$(under N_k = \sum_{n=1}^{N} \gamma(z_{nk}))$$

- deriviate it with respect to $\boldsymbol{\Sigma}_k$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

- maximizw it with respect to the mixing coefficient π_k, using lagrangian function

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

we gain

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

- repeat this three steps unitil convergence criterion is satisfied (EM algorithm)

3. An Alternative View of EM
   - first, we chose an initial value of parameters $\boldsymbol{\theta}^{\mathrm{old}}$
   - Second, we take, $\mathbf{E} step Evaluate p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$
   - let Q be

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

   and take M step,

$$\theta^{new} = \arg\max \mathcal{Q}\left(\theta, \theta^{old}\right)$$

- repeat this step until convergence criterion is satisfied

(a) Gaussian mixtures revisited

- the likelihood for the complete data set X, Z

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_{nk}}$$

- posterior distribution of Z is

$$p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right]^{z_{nk}}$$

- we gain

$$E\left[z_{nk}\right] = \frac{\sum_{z_{nk}} z_{nk} \left[\pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right]^{z_{nk}}}{\sum_{z_{nj}} \left[\pi_j \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)\right]^{z_{nj}}} = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)} = \gamma\left(z_{nk}\right)$$

(b) Relation to K-means

- EM algorithm makes a soft assignment based on the posterior probabilities (K-means are hard)

(c) Mixtures of Bernoulli distributions

- Consider the a finite mixture of Bernoulli distributions. let x be a variable, μ be a parameter, then

$$p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x} | \boldsymbol{\mu}_k)$$

- the log likelihood function is

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{\sum_{k=1}^{K} \pi_k p\left(\mathbf{x}_n | \boldsymbol{\mu}_k\right)\right\}$$

- prior distribution for the latent value is

$$p(\mathbf{z} | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- since

$$p(x | z, \mu) = \prod_{k=1}^{K} p(x | \mu_k)^{z_k} = \prod_{k=1}^{K} \left(\prod_{i=1}^{D} \mu_{ki}^{x_i} \left(1 - \mu_{ki}\right)^{1 - x_i}\right)^{z_k}$$

  , we could gain

$$p(\boldsymbol{x}, \boldsymbol{z} | \boldsymbol{\mu}, \boldsymbol{\pi}) = p(\boldsymbol{x} | \mathbf{z}, \boldsymbol{\mu}) p(\mathbf{z} | \boldsymbol{\pi}) = \prod_{k=1}^{K} \left(\pi_k p(\boldsymbol{x} | \boldsymbol{\mu}_k)\right)^{z_k} = \prod_{k=1}^{K} \left(\pi_k \left(\prod_{i=1}^{D} \mu_i^{x_i} \left(1 - \mu_i\right)^{1 - x_i}\right)\right)^{z_k}$$

- using

$$\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi}\right) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \pi_k \left(\prod_{t=1}^{D} \mu_{ki}^{x_{ni}} \left(1 - \mu_{ki}\right)^{1 - x_{nt}}\right),$$

$$E_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right) \left\{\ln \pi_k + \sum_{i=1}^{D} \left[x_{ni} \ln \mu_{ki} + \left(1 - x_{ni}\right) \ln \left(1 - \mu_{ki}\right)\right]\right\}$$

- deriviate it with respect to $\boldsymbol{\mu}_k$ and π_k

4. The EM algorithm in general
   - Here we give a very general treatment of the EM algorithm and in the process provide a proof that the EM algorithm derived heuristically in Sections 9.2 and 9.3 for Gaussian mixtures does indeed maximize the likelihood function
   - consider
   $$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$
   - we could decompose log likelihood as
   $$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p),$$
   under
   $$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$
   $$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$