Hao Tong 1749248

9.1 Consider the K-means algorithm discussed in Section 9.1. Show that as a consequence of there being a finite number of possible assignments for the set of discrete indicator variables r_{nk}, and that for each such assignment there is a unique optimum for the $\{\mu_k\}$, the K-means algorithm must converge after a finite number of iterations.

Sol. Since both E- and the M-step minise and distortion measure (9.1), the algorithm will never change from a particular assignment of data points to prototypes, unless the new assignment has a lower value for (9.1)

Since there is a finite number of possible assignments, each with a corresponding unique minimum of (9.1) with respect to the prototypes $\{\mu_k\}$, the K-means algorithm will ~~convergence~~ converge after a finite number of steps, when no re-assignment of data points to prototypes will result in a decrease of (9.1). When no-reassignment takes place, there also will ~~not~~ not be any change in $\{\mu_k\}$.

9.3. Consider a Gaussian mixture model in which the marginal distribution $p(z)$ for the latent variable is given by (9.10), and the conditional distribution $p(x|z)$ for the observed variable is given by (9.11). Show that the marginal distribution $p(x)$, obtained by summing $p(z)p(x|z)$ over all possibles values of z, is a Gaussian mixture of the form (9.7).

Sol. $p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k} \quad z_k \in \{0,1\}$

$p(x|z_k=1) = N(x|\mu_k, \Sigma_k)$

$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}. \qquad p(z_k=1) = \pi_k$

$\therefore\ p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^{K} p(z_k=1)\, p(x|z_k=1) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k).$

9.7. Verify that maximization of the complete-data log likelihood (9.36) for a Gaussian mixture model lead to the result that the means and covariances of each component are fitted independently to the corresponding group of data points, and the mixing coeffients are given by the ~~function~~ fractions of points in each group.

Sol. from (9.36) we can get

$$\ln p(X, Z|\mu, \Sigma, \pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln N(x_n|\mu_k, \Sigma_k)\}$$

$$\frac{\partial \ln p(X, Z|\mu, \Sigma, \pi)}{\partial \mu_k} = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot \frac{\Sigma^{-1}(x_n-\mu_k)}{N(x_n|\mu_k, \Sigma_k)} = 0$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \cdot \frac{x_n-\mu_k}{N(x_n|\mu_k, \Sigma_k)} = \sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} \frac{x_n-\mu_k}{N(x_n|\mu_k, \Sigma_k)} = 0$$

So, for each point, x_n, it has a specific class, so for each z_n, only has one component is 1, others are all 0. So, for each component in ~~N~~ mixture model, it only correlated with points ~~of~~ which are belong to this group. So, the means of each component are fitted independently to the corresponding group of data points.

And, same for covariance,

$$\frac{\partial \ln p(X, Z|\mu, \Sigma, \pi)}{\partial \Sigma_k} = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \frac{(x_n-\mu_k)(x_n-\mu_k)^T}{N(x_n|\mu_k, \Sigma_k)}$$

$$= \sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} \frac{(x_n-\mu_k)(x_n-\mu_k)^T}{N(x_n|\mu_k, \Sigma_k)}$$

each component are independently.

Finally, for π_k, it must add the constraint condition.

$$\frac{\partial \ln p(X, Z|\mu, \Sigma, \pi) + \lambda(\sum_{k=1}^{K}\pi_k - 1)}{\partial \pi_k} = 0 \Rightarrow \ln p(X, Z|\mu, \Sigma, \pi) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln N(x_n|\mu_k, \Sigma_k)\} + \lambda \sum_{k=1}^{K}\pi_k - \lambda$$

we only get the term related with π_k. so

$$\frac{\partial \ln p(X, Z|\mu, \Sigma, \pi) + \lambda(\sum_{k=1}^{K}\pi_k - 1)}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}\left(\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\ln \pi_k + \lambda \cdot \sum_{k=1}^{K}\pi_k\right) = \frac{\partial}{\partial \pi_k}\left\{\sum_{k=1}^{K}\ln \pi_k \cdot \sum_{n=1}^{N} z_{nk} + \lambda \cdot \sum_{k=1}^{K}\pi_k\right\}$$

$$= \frac{\partial}{\partial \pi_k}\left\{\sum_{k=1}^{K}\ln \pi_k \cdot N_k + \lambda \cdot \sum_{k=1}^{K}\pi_k\right\} = \frac{N_k}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = -\frac{N_k}{\lambda}$$

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K} -\frac{N_k}{\lambda} = -\frac{1}{\lambda}\sum_{k=1}^{K}N_k = -\frac{1}{\lambda}N = 1 \quad \therefore \lambda = -N \Rightarrow \pi_k = \frac{N_k}{N}.$$

where N_k is the number of data points in group X_k.

9.10. Consider a density model given by a mixture distribution $p(x)=\sum_{k=1}^{K}\pi_k p(x|k)$ and suppose that we partition the vector x into two parts so that $x=(x_a, x_b)$. Show that the conditional density $p(x_b|x_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and for the component densities.

Sol. $p(x)=\sum_{k=1}^{K}\pi_k p(x|k)$

$$P(x_b|x_a)=\frac{P(x_a, x_b)}{P(x_a)}$$

$x=(x_a, x_b)$

$P(x_a)=\sum_{k=1}^{K}\pi_k P(x_a|k)$

$$\therefore P(x_b|x_a)=\frac{\sum_{k=1}^{K}\pi_k P(x_a, x_b|k)}{\sum_{k=1}^{K}\pi_k P(x_a|k)}$$

$P(x_a, x_b)=\sum_{k=1}^{K}\pi_k P(x_a, x_b|k)$

$p(x_a, x_b|k)=P(x_b|x_a, k)\cdot P(x_a|k)$

$$\therefore P(x_b|x_a)=\frac{\sum_{k=1}^{K}\pi_k\cdot P(x_b|x_a, k)\cdot P(x_a|k)}{\sum_{k=1}^{K}\pi_k P(x_a|k)}$$

$$\therefore P(x_b|x_a)=\sum_{k=1}^{K}P(x_b|x_a, k)\cdot\frac{P(x_a|k)}{\sum_{k=1}^{K}\pi_k p(x_a|k)}$$

$\frac{P(x_a|k)}{\sum_{k=1}^{K}\pi_k p(x_a|k)}$ denoted as s_k

$$\therefore P(x_b|x_a)=\sum_{k=1}^{K}s_k P(x_b|x_a, k)$$

9.18. Consider a Bernoulli mixture model as discussed in Section 9.3.3, together with a prior distribution $p(\mu_k|a_k, b_k)$ over each of the parameter vetors μ_k given by the beta distribution (2.13), and a Dirichlet prior $P(\pi|\alpha)$ given by (2.38). Derive the EM algorithm for maximizing the posterior probability $P(\mu, \pi|X)$.

Sol. $p(\theta|x)=\frac{P(x|\theta)P(\theta)}{P(x)}$, θ is parameters

the E-M algorithm want to maximum the $P(x|\theta)$. and if it add the ~~pos~~ prior distribution. We will be aim to maximum $p(x|\theta)p(\theta)$.

So. ~~log~~ $\ln p(x|\theta)p(\theta)=\ln p(x|\theta)+\ln p(\theta)$

We need to ~~inform the~~ import the latent variables. Z.

$\ln p(x|\theta)=\ln\{\sum_z p(x, z|\theta)\}$

~~ln~~ in the M-step. $\ln p(x|\theta)=\sum_z p(z|x, \theta^{old})\ln p(x, z|\theta)$

from, (9.55).

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

So, the M-step for ~~posterior~~ maximum posterior probability can be written as

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\Gamma(z_{nk})\left\{\ln\pi_k+\sum_{i=1}^{D}[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})]\right\} + \sum_{j=1}^{K}\sum_{i'=1}^{D}\{(a_j-1)\ln\mu_{ji'}+(b_j-1)\ln(1-\mu_{ji'})\}$$

$$+\sum_{l=1}^{K}(a_l-1)\ln\pi_l.$$ $\Gamma(z_{nk})$ is the form as (9.56)

terms in prior distributions independent of $\{\mu_k\}$ and π has been dropped.

derivate above formula respect to μ_{ki}.

$$\sum_{n=1}^{N}\Gamma(z_{nk})\left\{\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right\} + \frac{a_k-1}{\mu_{ki}} + \frac{1-b_k}{1-\mu_{ki}}$$

$$= \frac{\sum_{n=1}^{N}\Gamma(z_{nk})\cdot x_{ni}}{\mu_{ki}} - \frac{\sum_{n=1}^{N}\Gamma(z_{nk})(1-x_{ni})}{1-\mu_{ki}} + \frac{a_k-1}{\mu_{ki}} + \frac{1-b_k}{1-\mu_{ki}}$$

$$= \frac{N_k\bar{x}_{ki}}{\mu_{ki}} - \frac{N_k(1-\bar{x}_{ki})}{1-\mu_{ki}} + \frac{a_k-1}{\mu_{ki}} + \frac{1-b_k}{1-\mu_{ki}}$$

$$= \frac{N_k\bar{x}_{ki}+a_k-1}{\mu_{ki}} - \frac{N_k-N_k\bar{x}_{ki}+b_k-1}{1-\mu_{ki}} = 0$$

So. $$\mu_{ki} = \frac{N_k\bar{x}_{ki}+a_k-1}{N_k+a_k+b_k-2}$$

And then, we need to derive the π_k for M-step. and the constraint is $\sum_{k}^{K}\pi_k=1$

So, the maximum posterior probability can be written without terms indepent of π.

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\Gamma(z_{nk})\ln\pi_k + \sum_{l=1}^{K}(a_l-1)\ln\pi_l + \lambda\left(\sum_{j=1}^{K}\pi_j-1\right)$$

derivate it get below formula.

$$\sum_{n=1}^{N}\frac{\Gamma(z_{nk})}{\pi_k} + \frac{a_k-1}{\pi_k} + \lambda = 0 \Rightarrow \frac{N_k+a_k-1}{\pi_k} = -\lambda \Rightarrow \pi_k = \frac{N_k+a_k-1}{-\lambda}$$

$$\sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K}\frac{N_k+a_k-1}{-\lambda} = \frac{N+a_0-K}{-\lambda} = 1 \qquad \therefore -\lambda = N+a_0-K$$

$$\therefore \pi_k = \frac{N_k+a_k-1}{-\lambda} = \frac{N_k+a_k-1}{N+a_0-K}.$$

So, in the M-step. μ_{ki} and π_k can be updated by above two formula.

in the E-step. use μ^{old}, π^{old} to estimate the latent variables' posterior probability distribution.

9.19. Consider a D-dimensional variable x each of whose components i is itself a multinomial variable of degree M so that X is a binary vector with components x_{ij} where $i=1,\ldots,D$ and $j=1,\ldots,M$, subject to the ~~constant~~ constraint that $\sum_j x_{ij}=1$ for all i. Suppose that the distribution of these variables is described by a mixture of the discrete multinomial distributions considered in Section 2.2 so that $p(x)=\sum_{k=1}^{K}\pi_k p(x|\mu_k)$ where $p(x|\mu_k)=\prod_{i=1}^{D}\prod_{j=1}^{M}\mu_{kij}^{x_{ij}}$. The parameters μ_{kij} represent the probabilities $p(x_{ij}=1|\mu_k)$ and must satisfy $0\le\mu_{kij}\le 1$ together with the constraint $\sum_j \mu_{kij}=1$ for all values of k and i. Given an observed data set $\{x_n\}$, where $n=1,\ldots,N$, derive the E and M step equations of the EM algorithm for opeimizing the mixing coeffients π_k and the compnet parameters μ_{kij} of this distribution by maximum likelihood.

Sol. In order to maximum likelihood. we introduce a ~~vect~~ latent variable. Z

$$p(X|\theta)=\ln\{\sum_Z p(X,Z|\theta)\}\quad \theta \text{ is parameter.}$$

the E-step in E algorithm is to estimate the posterior probability distribution of latent Varible Z. so.

$$r(z_{nk})\equiv p(z_{nk}=1|x_n)=\frac{p(z_k=1)\,p(x_n|z_k)}{\sum_z p(z_k=1)\,p(x_n|z_k)}=\frac{\pi_k\cdot\prod_{i=1}^{D}\prod_{j=1}^{M}\mu_{kij}^{x_{ij}}}{\sum_{k=1}^{K}\pi_k\cdot\prod_{i=1}^{D}\prod_{j=1}^{M}\mu_{kij}^{x_{ij}}}$$

And then for the M-step

$$Q(\theta,\theta^{old})=\sum_Z p(Z|X,\theta^{old})\ln\{p(X,Z|\theta)\}.$$

$$Q(\theta,\theta^{old})=\sum_{n=1}^{N}\sum_{k=1}^{K}r(z_{nk})\cdot\left(\ln\pi_k+\ln\prod_{i=1}^{D}\prod_{j=1}^{M}\mu_{kij}^{x_{nij}}\right)=\sum_{n=1}^{N}\sum_{k=1}^{K}r(z_{nk})\left(\ln\pi_k+\sum_{i=1}^{D}\sum_{j=1}^{M}x_{ij}\ln\mu_{kij}\right)$$

$\cancel{\frac{\partial Q(\theta,\theta^{old})}{\partial(\mu_{kij})}=\sum_{n=1}^{N}}$ because of the constraints, we must add the constraints to use Lagrange.

$$\text{get } L=\sum_{n=1}^{N}\sum_{k=1}^{K}r(z_{nk})\left(\ln\pi_k+\sum_{i=1}^{D}\sum_{j=1}^{M}x_{nij}\ln\mu_{kij}\right)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)+\sum_{k=1}^{K}\sum_{i=1}^{D}\eta_{ki}\left(\sum_{j=1}^{M}\mu_{kij}-1\right)$$

$$\frac{\partial L}{\partial\pi_k}=\sum_{n=1}^{N}r(z_{nk})\cdot\frac{1}{\pi_k}+\lambda=0\Rightarrow-\lambda=\frac{N_k}{\pi_k}\Rightarrow\sum_{k=1}^{K}\pi_k=\sum_{k=1}^{K}\frac{N_k}{-\lambda}=\frac{N}{-\lambda}=1\Rightarrow-\lambda=N$$

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

Then. $\frac{\partial L}{\partial \mu_{kij}} = \sum_{n=1}^{N} \gamma(z_{nk}) \frac{x_{nij}}{\mu_{kij}} + \eta_{ki} = 0 \quad \Rightarrow \quad -\eta_{ki} = \frac{\sum_{n=1}^{N} \gamma_{z_{nk}} x_{nij}}{\mu_{kij}}$

$$\sum_j \mu_{kij} = \sum_j -\frac{\sum_{n=1}^{N} \gamma_{z_{nk}} x_{nij}}{\eta_{ki}} = -\frac{\sum_{n=1}^{N}\sum_{j=1}^{M} \gamma_{z_{nk}} x_{nij}}{\eta_{ki}} = -\frac{\sum_{n=1}^{N} \gamma_{z_{nk}} \sum_{j=1}^{M} x_{nij}}{\eta_{ki}} = -\frac{\sum_{n=1}^{N} \gamma_{z_{nk}}}{\eta_{ki}} = -\frac{N_k}{\eta_{ki}} = 1$$

$\therefore \; -\eta_{ki} = N_k$

$\therefore \; \mu_{kij} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{z_{nk}} x_{nij}$.

So, we derive the formula how parameters ~~upate~~ update in M-step.

9.20 Show that maximization of the expected complete-data log likelihood function (9.62) for the Bayesian linear regression model leads to the M step re-estimation result (9.63) for α.

Sol. ~~for~~ from 9.62. $E[\ln p(t, w|\alpha, \beta)] = \frac{M}{2}\ln(\frac{\alpha}{2\pi}) - \frac{\alpha}{2}E[w^T w] + \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{n=1}^{N} E[(t - w^T\phi_n)]^2$

$$\frac{\partial E[\ln p(t, w|\alpha, \beta)]}{\partial \alpha} = \frac{M}{2}\cdot\frac{1}{\alpha} - \frac{1}{2}E[w^T w] = 0$$

$\therefore \; \alpha = \frac{M}{E[w^T w]} = \frac{M}{m_N^T m_N + Tr(S_N)}$

9.21. Using the evidence framework of Section 3.5, derive the M-step re-estimation equations for the parameter β in the Bayesian linear regression model, analogous to the result (9.63) for α.

Sol. from 9.62. $E[\ln p(t, w|\alpha, \beta)] = \frac{M}{2}\ln(\frac{\alpha}{2\pi}) - \frac{\alpha}{2}E[w^T w] + \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\sum_{n=1}^{N} E[(t - w^T\phi_n)]^2$

$$\frac{\partial E[\ln p(t, w|\alpha, \beta)]}{\partial \alpha} = \frac{N}{2}\cdot\frac{1}{\beta} - \frac{1}{2}\sum_{n=1}^{N} E[(t_n - w^T\phi_n)^2]^2 = 0$$

$$\begin{aligned} E[(t_n - w^T\phi_n)^2] &= E[t_n \cdot t_n - 2t_n w^T\phi_n + w^T\phi_n \cdot w^T\phi_n] \\ &= t_n^2 - 2t_n w^T\phi_n + Tr[\phi_n\phi_n^T(m_N m_N^T + S_N)] \\ &= (t_n - m_N^T\phi_n)^2 + Tr[\phi_n\phi_n^T S_N] \end{aligned}$$

$\therefore \; \beta^{-1} = \frac{1}{N}[(t_n - m_N^T\phi_n)^2 + Tr[\phi^T\phi S_N]]$

9.25. Show the lower bound $L(q,\theta)$ given by 9.71. with $q(z)=p(z|X,\theta^{old})$. has the same gradient with respect to θ as the loglikelihood function $\ln p(X|\theta)$ at the point $\theta=\theta^{(old)}$.

Sol. Because of $q(z)=p(z|X,\theta^{old})$ So. $KL(q\|p)=0$

So. $\ln p(X|\theta)=L(q,\theta)+KL(q\|p)=L(q,\theta)$

$\therefore$ so the gradient of $L(q,\theta)$ r equal to $\ln p(X|\theta)$

9.26. Consider the incremental form of the EM algorithm for a mixture of Gaussians, in which the ~~sep~~ responsibilities are recomputed only for a specific data point x_m. Starting from the M-step formulae (9.17) and (9.18), derive the results 9.78 and 9.79 for updating the component means.

Sol. from 9.18. $N_k^{old}=\sum_{n=1}^{N} r(z_{nk})$.

for a specific data point x_m, we recomputing the responsibities. $r(z_{mk})$

$$\therefore N_k^{new}=\sum_{n\neq m} r(z_{nk})+r(z_{mk})$$

$$\therefore N_k^{new}=N_k^{old}+r(z_{mk})-r^{old}(z_{mk})$$

from (9.6).

$$\begin{cases}\sum_{n=1}^{N} r_{z_{nk}}^{old}(x_n-\mu_k^{old})=\sum_{n\neq m} r_{z_{nk}}^{old}(x_n-\mu_k^{old})+r_{z_{mk}}^{old}(x_m-\mu_k^{old})=0\\ \sum_{n\neq m} r_{z_{nk}}^{old}(x_n-\mu_k^{new})+r_{z_{mk}}^{new}(x_n-\mu_k^{new})=0\end{cases}$$

$$\Rightarrow\begin{cases}\sum_{n\neq m} r_{z_{nk}}^{old}x_n+r_{z_{mk}}^{old}x_m-\sum_{n\neq m} r_{z_{nk}}^{old}\mu_k^{old}-r_{z_{mk}}^{old}\mu_k^{old}=0\\ \sum_{n\neq m} r_{z_{nk}}^{old}x_n+r_{z_{mk}}^{new}x_n-\sum_{n\neq m} r_{z_{nk}}^{old}\mu_k^{new}-r_{z_{mk}}^{new}\mu_k^{new}=0\end{cases}\Rightarrow\begin{cases}\sum_{n\neq m} r_{z_{nk}}^{old}x_n+r_{z_{mk}}^{old}x_m=N_k^{old}\mu_k^{old}\\ \sum_{n\neq m} r_{z_{nk}}^{old}x_n+r_{z_{mk}}^{new}x_m=N_k^{new}\mu_k^{new}\end{cases}$$

$$\therefore N_k^{new}\mu_k^{new}-N_k^{old}\mu_k^{old}=r_{z_{mk}}^{new}x_m-r_{z_{mk}}^{old}x_m$$

$$\therefore \mu_k^{new}=\frac{N_k^{old}}{N_k^{new}}\mu_k^{old}+\frac{r_{z_{mk}}^{new}-r_{z_{mk}}^{old}}{N_k^{new}}x_m \qquad \therefore \mu_k^{new}=\left(1-\frac{r_{z_{mk}}^{new}-r_{z_{mk}}^{old}}{N_k^{new}}\right)\mu_k^{old}+\frac{r_{z_{mk}}^{new}-r_{z_{mk}}^{old}}{N_k^{new}}x_m$$

$$\frac{N_k^{old}}{N_k^{new}}=1-\frac{r_{z_{mk}}^{new}-r_{z_{mk}}^{old}}{N_k^{new}} \qquad =\mu_k^{old}+\frac{r_{z_{mk}}^{new}-r_{z_{mk}}^{old}}{N_k^{new}}(x_m-\mu_k^{old})$$