Homework 3

Aukkawut Ammartayakun
CS 539 Machine Learning

Spring 2023

Problem 1

Consider three binary variables $a, b, c \in \{0, 1\}$ having the joint distribution given in the above table. Show

a	b	c	$p(a, b, c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

by direct evaluation that this distribution has the property that a and b are marginally dependent, so that $p(a, b) \neq p(a)p(b)$, but that they become independent when conditioned on c, so that $p(a, b|c) = p(a|c)p(b|c)$ for both $c = 0$ and $c = 1$.

Solution. We want to show that $p(a, b) \neq p(a)p(b)$ and $p(a, b|c) = p(a|c)p(b|c)$ for both $c = 0$ and $c = 1$. First, $p(a, b) = p(a, b|c = 0) + p(a, b|c = 1)$ in which from the table has 4 different values depending on a and b whereas $p(a) = \sum_{b,c} p(a, b, c)$ and $p(b) = \sum_{a,c} p(a, b, c)$ both have 2 different values. Hence, $p(a, b) \neq p(a)p(b)$.

Second, using the same logic as previous one to get $p(c = 0) = \sum_{a,b} p(a, b, c = 0) = 0.48$ and $p(c = 1) = 1 - p(c = 0) = 0.52$.
Using Bayes' rule, one can see that

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \begin{cases} 0.192/0.48 & (a, b, c) = (0, 0, 0) \\ 0.144/0.52 & (a, b, c) = (0, 0, 1) \\ 0.048/0.48 & (a, b, c) = (0, 1, 0) \\ 0.216/0.52 & (a, b, c) = (0, 1, 1) \\ 0.192/0.48 & (a, b, c) = (1, 0, 0) \\ 0.064/0.52 & (a, b, c) = (1, 0, 1) \\ 0.048/0.48 & (a, b, c) = (1, 1, 0) \\ 0.096/0.52 & (a, b, c) = (1, 1, 1) \end{cases}$$

with $p(b|c) = \frac{p(b,c)}{p(c)} = \begin{cases} 0.384/0.48 & (b, c) = (0, 0) \\ 0.208/0.52 & (b, c) = (0, 1) \\ 0.096/0.48 & (b, c) = (1, 0) \\ 0.312/0.52 & (b, c) = (1, 1) \end{cases}$ and $p(a|c) = \frac{p(a,c)}{p(c)} = \begin{cases} 0.240/0.48 & (a, c) = (0, 0) \\ 0.360/0.52 & (a, c) = (0, 1) \\ 0.240/0.48 & (a, c) = (1, 0) \\ 0.160/0.52 & (a, c) = (1, 1) \end{cases}$.

Multiplication shows that $p(a|c)p(b|c) = \begin{cases} (0.384/0.48)(0.240/0.48) & (a,b,c) = (0,0,0) \\ (0.208/0.52)(0.360/0.52) & (a,b,c) = (0,0,1) \\ (0.384/0.48)(0.240/0.48) & (a,b,c) = (0,1,0) \\ \vdots & \vdots \\ (0.312/0.52)(0.160/0.52) & (a,b,c) = (1,1,1) \end{cases}$ which all of them are equal. Thus, we can conclude that $p(a,b|c) = p(a|c)p(b|c)$ for both $c = 0$ and $c = 1$.

Problem 2

Evaluate the distributions $p(a)$, $p(b|c)$, and $p(c|a)$ corresponding to the joint distribution given in the above table. Hence show by direct evaluation that $p(a,b,c) = p(a)p(c|a)p(b|c)$. Draw the corresponding directed graph.

Solution. We want to show that $p(a,b,c) = p(a)p(c|a)p(b|c)$.

First, $p(a) = \sum_{b,c} p(a,b,c) = \begin{cases} 0.6 & a = 0 \\ 0.4 & a = 1 \end{cases}$ and $p(b|c) = \frac{p(b,c)}{p(c)} = \begin{cases} 0.384/0.48 & (b,c) = (0,0) \\ 0.208/0.52 & (b,c) = (0,1) \\ 0.096/0.48 & (b,c) = (1,0) \\ 0.312/0.52 & (b,c) = (1,1) \end{cases}$. Lastly, $p(c|a) = \frac{p(a,c)}{p(a)} = \begin{cases} 0.240/0.6 & (a,c) = (0,0) \\ 0.360/0.4 & (a,c) = (0,1) \\ 0.240/0.6 & (a,c) = (1,0) \\ 0.160/0.4 & (a,c) = (1,1) \end{cases}$. Evaluating the product yields the same result. Thus, the relationship is as followed

$$a \to c \to b$$

Problem 3

Consider two discrete variables x and y each having three possible states, for example $x, y \in \{0,1,2\}$. Construct a joint distribution $p(x,y)$ over these variables having the property that the value $\hat{x}$ that maximizes the marginal $p(x)$, along with the value $\hat{y}$ that maximizes the marginal $p(y)$, together have probability zero under the joint distribution, so that $p(\hat{x},\hat{y}) = 0$.

	$x=0$	$x=1$	$x=2$
$y=0$	0	0.4	0
$y=1$	0.3	0	0
$y=2$	0.3	0	0

Solution. $\arg\max_x p(x) = 0$, $\arg\max_y p(y) = 0$ and, $p(\hat{x},\hat{y}) = 0$.

Problem 4

Suppose we wish to use the EM algorithm to maximize the posterior distribution over parameters $p(\boldsymbol{\theta}|\mathbf{X})$ for a model containing latent variables, where X is the observed data set. Show that the E step remains the same as in the maximum likelihood case, whereas in the M step the quantity to be maximized is given by $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) + \ln p(\boldsymbol{\theta})$ where $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ is defined by

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{1}$$

Solution. As in the normal EM algorithm, the difference between the normal one and this one is instead of maximizing the likelihood function, we maximize the posterior distribution. From

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

for the sake of generality, let say $p(\boldsymbol{\theta}|\mathbf{X}) = kp(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Then, we have

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

This shows that the only difference in comparison to the given $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ is the term $\ln p(\boldsymbol{\theta})$ which the complete-data log likelihood function is only used in M step. Thus, E step remains the same but $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ is replaced by $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) + \ln p(\boldsymbol{\theta})$ from linear separability of logarithm function as shown above.

Problem 5

Consider a special case of a Gaussian mixture model in which the covariance matrices Σ_k of the components are all constrained to have a common value Σ. Derive the EM equations for maximizing the likelihood function under such a model.

Solution. Let say we constraint the covariance to be homogenous. Since the only change here is the covariance, we can only differentiate this with respect to $\boldsymbol{\Sigma}$ in order to chage the updated value of $\boldsymbol{\Sigma}$ (which is the only thing that changed from normal EM). The result equation for updating $\boldsymbol{\Sigma}$ is

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) = -\frac{N}{2} \ln |\boldsymbol{\Sigma}| + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)\Sigma^{-1}(\mathbf{x}_n - \mu_k)^\top$$

As mentioned that the covariance are homogenous, we can factor it out, differentiate that with respect to the $\boldsymbol{\Sigma}$ and set it to zero to obtain:

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^\top$$

Problem 6

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 vectors μ_k given by the beta distribution (2.13),

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2}$$

and a Dirichlet prior $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$ given by (2.38)

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \tag{3}$$

Derive the EM algorithm for maximizing the posterior probability $p(\boldsymbol{\mu}, \boldsymbol{\pi}|\mathbf{X})$.

Solution. In the similar manner to Problem 4, we can see that at M step, we need to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}}) + \ln p(\boldsymbol{\theta})$. As discussed in the section 9.3.3 that

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left[\ln \pi_k + \sum_{i=1}^{D}(x_{ni}\ln \mu_{ki} + (1 - x_{n_i}\ln(1-\mu_{ki}))\right] \tag{4}$$

given the prior, we can altered the equation to add the log-transformed prior term

$$\ln p(\boldsymbol{\theta}) = \sum_{j=1}^{K}\sum_{m=1}^{D}[(a_j - 1)\ln \mu_{jm} + (b_j - 1)\ln(1-\mu_{jm})] + \sum_{l=1}^{K}(\alpha_l - 1)\ln \pi_l$$

Differentiating the above notion with respect to μ_{ki} and set it to zero to get μ_{ki} update. Now, maximizing the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{\text{old}})$ to get π_k update.