Introduction to Machine Learning

Project Phase 1

Ilia Hashemi Rad

MohammadPouya Toroghi

99102456 400109179

Theory Questions

Question 1

Solution:

To better understand this question we need to keep in mind that regardless of the convexity status of $l(\theta)$, finding a function $Q(\theta, \theta^{(t)})$ that provides the necessary conditions is enough to satisfy the following inequality $l(\theta^{(t)}) = Q(\theta^{(t)}, \theta^{(t)}) \leq Q(\theta^{(t+1)}, \theta^{(t)}) \leq l(\theta^{(t+1)})$. Which means that no matter what $l(\theta)$ is, the MM Algorithm converges to one of its maximums, however, for non-convex $l(\theta)$, the converging $\theta^{(t)}$ might not be the global maximum of $l(\theta)$, but the local maximum. In which case, we could come up with a different surrogate function $Q'(\theta, \theta^{(t)})$ to continue the algorithm after $Q(\theta, \theta^{(t)})$'s convergence. And this is possible because the surrogate function is not unique and we can use many approaches to come up with different surrogate functions for some objective function $l(\theta)$, such as $Jensen's$ inequality, Cauchy-Schwartz inequality, Inequality of arithmetic and geometric means, second order Taylor expansion of twice-differentiable functions with bounded curvature, or many others. Of course it will be easier to perform the MM algorithm using a convex function $Q(\theta, \theta^{(t)})$ because finding $\arg\max_{\theta} Q(\theta, \theta^{(t)})$ is easier for convex function. Now, finding convex $Q(\theta, \theta^{(t)})$ is case-relative but a useful method, for some cases, is the aforementioned $Jensen's$ Inequality.

Question 2

Solution:

$$p_Y(\mathbf{y};\theta) = \prod_{i=1}^{N} p_Y(y^{(i)};\theta) = \prod_{i=1}^{N}(\sum_{k=1}^{K} p_{Y,Z}(y^{(i)}, Z^{(i)} = k;\theta)) = \sum_{k=1}^{K}(\prod_{i=1}^{N} p_{Y,Z}(y^{(i)}, Z^{(i)} = k;\theta))$$

$$= \sum_{k=1}^{K} p_{Y,Z}(\mathbf{y}, Z = z_k;\theta) = \sum_{k=1}^{K} p_Z(z_k;\theta)\frac{p_{Y,Z}(\mathbf{y}, Z = z_k;\theta)}{p_Z(z_k;\theta)} = \sum_{k=1}^{K} p_Z(z_k;\theta) p_{Y|Z}(\mathbf{y}|Z = z_k;\theta)$$

We need to keep in mind that y_n may have come from an arbitrary distribution, which is unknown to us. Hence, calculating $p_{\mathbf{Y}}(y_n;\theta)$ is rather difficult, even impossible. So, we need to fit a family of common distributions to let us handle the given data more easily; and that's the idea behind mixture models. Now, having a family of known distributions, calculating $p_{\mathbf{Y},\mathbf{Z}}(y_n, z_n;\theta)$ is much easier. Consequently, optimizing $p_{\mathbf{Y},\mathbf{Z}}(y_n, z_n;\theta)$ is easier than optimizing $p_{\mathbf{Y}}(y_n;\theta)$.

Question 3

Solution:

While both, EM and VI, aim to approximate a true posterior distribution, they have different mathematical approaches. As we learned previously in the project PDF, EM maximizes the likelihood function of the observed data, iteratively, using 2 steps; Expectation step(E-step), where the expected values of the latent variables are computed given the current parameter estimates, and a Maximization step(M-step), where the parameters are updated to maximize the expected log-likelihood. EM assumes that the latent variables are unobserved, and the goal is to estimate them along with the parameters. VI, however, is a deterministic optimization method that minimizes the $Kullback - Leibler$(KL) divergence between the true posterior distribution and a simpler, tractable distribution(often a Gaussian). VI assumes that the latent variables are observed, and the goal is to estimate the posterior distribution over the parameters given the observed data and the latent variables.
Another aspect of the difference between these two algorithms, which was mentioned in $Wikipedia$, is that VI can be seen as an extension of EM from MAP estimation of the single most probable value of each parameter to fully $Bayesian$ estimation which computes (an approximation to) the entire posterior distribution of the parameters and latent variables. As in EM, it finds a set of optimal parameter values.

Question 4

Solution:

1.

Assume that we have a data set of N observed d-dimensional points $\{\boldsymbol{x}_i\}_{i=1}^N$.
the data set follows a mixture of K Multivariate Normal distributions such that:

$$p(\boldsymbol{x}|\Theta) = \sum_k \pi_k \, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

In which π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the prior probability, mean vector and the covariance matrix of the kth Multivariate Gaussian model.
Our task is to estimate the set of parameters $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for each Multivariate Gaussian model using the data set.

Clearly at the first step, we should initialize mentioned parameters. Here we choose an algorithm to do so. In this algorithm, we initialize the parameters π_k uniformly:

$$\pi_k^0 = \frac{1}{K} \qquad \forall k; \quad 1 \le k \le K$$

Which means each point in the data set is equally likely to belong to the kth Multivariate Gaussian model. Also, in order to initialize the Gaussian models' parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we first divide the data set randomly into K clusters with $M = \frac{N}{K}$ data points. (Notice if the value of M wasn't integer, we would divide the data set to clusters such that all the clusters approximately have the same sizes.)Then using MLE method we would have:

$$\begin{cases} \boldsymbol{\mu}_k^0 = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{x}_j^{(k)} \\ \boldsymbol{\Sigma}_k^0 = \frac{1}{M}\sum_{j=1}^{M} (\boldsymbol{x}_j^{(k)} - \boldsymbol{\mu}_k^0)(\boldsymbol{x}_j^{(k)} - \boldsymbol{\mu}_k^0)^T \end{cases} \qquad \forall k; \quad 1 \le k \le K$$

Where, $\boldsymbol{x}_j^{(k)}$ denotes the jth point of the kth cluster.

2.

In this section we are supposed to determine complete data set likelihood. According to descriptions at the previous part we have:

$$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) = \prod_i p(\boldsymbol{x}_i, z_i|\boldsymbol{\Theta}) = \prod_i \sum_k p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta}) = \prod_i \sum_k p(z_i = k)p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta}) = \prod_i \sum_k \pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$= \prod_i \sum_k p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta}) \frac{\pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta})} = \prod_i \sum_k \omega_{i,k} \frac{\pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\omega_{i,k}}$$

In which, using $Bayes' \ Rule$ we have:

$$\omega_{i,k} \triangleq p^*(z_i = k|\boldsymbol{x}_i, \boldsymbol{\Theta}) = \frac{p(z_i = k)p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta})}{\sum_{k'} p(z_i = k')p(\boldsymbol{x}_i|z_i = k', \boldsymbol{\Theta})} = \frac{\pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

Therefore, the complete log likelihood function is determined in this way:(this part is for the next section)

$$L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) = \sum_i \log \sum_k \omega_{i,k} \frac{\pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\omega_{i,k}} \ge \sum_i \sum_k \omega_{i,k} \log \frac{\pi_k \, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\omega_{i,k}} \qquad *$$

3.

Assume that the latent parameters at the iteration t, is defined as calculated before

$$\omega_{i,k}^t \triangleq p^*(z_i = k|\boldsymbol{x}_i, \boldsymbol{\Theta}^t) = \frac{\pi_k^t \ \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t)}{\sum\limits_{k'} \pi_{k'}^t \ \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}^t, \boldsymbol{\Sigma}_{k'}^t)}$$

Using the results in the previous parts and the source book, we can define an instance iterative target function(the lower bound function) to work with:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \sum_i \sum_k \omega_{i,k}^t \log \frac{\pi_k \ \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\omega_{i,k}^t}$$

$$= \sum_i \sum_k \omega_{i,k}^t \log \left(\frac{\pi_k}{\omega_{i,k}^t \sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \ \exp\left[-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \ \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right] \right)$$

Which obviously holds the two constraint of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) \leq L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$ (based on *) and $Q(\boldsymbol{\Theta}^t, \boldsymbol{\Theta}^t) = L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}^t)$ (based on the proof in source). We expand the result more, to use that in further steps

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \sum_i \sum_k \omega_{i,k}^t [\log \pi_k - \log \omega_{i,k}^t - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \ \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)] \tag{1}$$

Now we go for the E-step and M-step

E-step:

The latent parameters to be estimated in each iteration of E-step have been determined in a closed-form:

$$\omega_{i,k}^t = \frac{\pi_k^t \ \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t)}{\sum\limits_{k'} \pi_{k'}^t \ \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}^t, \boldsymbol{\Sigma}_{k'}^t)}$$

Where, ω is a $N \times K$ matrix.

M-step:

Here we are going to update the Θ parameters using maximization of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$ by parameters:

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$$

Using the expanded equation in (1) we conclude:

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \sum_i \sum_k \omega_{i,k}^t [\log \pi_k - \log \omega_{i,k}^t - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \ \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)]$$

Updating $\boldsymbol{\mu}_k$:

the target is to find

$$\boldsymbol{\mu}_k^{t+1} = \arg\max_{\boldsymbol{\mu}_k} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$$

To do so, taking partial derivative $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$ with respect to $\boldsymbol{\mu}_k$ we have:

$$\left.\frac{\partial Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)}{\partial \boldsymbol{\mu}_k}\right|_{\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^{t+1}} = \sum_i \omega_{i,k}^t \left.\frac{\partial [-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \ \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)]}{\partial \boldsymbol{\mu}_k}\right|_{\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^{t+1}} = 0 \tag{2}$$

On the other hand, it can be proven that if $\boldsymbol{Y}$ is a symmetric matrix for any two vectors of x and a we have:

$$\frac{\partial(\boldsymbol{x}-\boldsymbol{a})^T\boldsymbol{Y}(\boldsymbol{x}-\boldsymbol{a})}{\partial\boldsymbol{x}} = -2\boldsymbol{Y}(\boldsymbol{x}-\boldsymbol{a})$$

So, using the fact that the covariance matrix and consequently its inverse are symmetric, from (2) we conclude:

$$\sum_i \omega_{i,k}^t \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{t+1}) = 0 \Rightarrow \sum_i \omega_{i,k}^t \boldsymbol{x}_i = \boldsymbol{\mu}_k^{t+1}\sum_i \omega_{i,k}^t$$

$$\Rightarrow \boldsymbol{\mu}_k^{t+1} = \frac{\sum\limits_i \omega_{i,k}^t \boldsymbol{x}_i}{\sum\limits_i \omega_{i,k}^t}$$

Updating $\boldsymbol{\Sigma}_k$:

the target is to find

$$\boldsymbol{\Sigma}_k^{t+1} = \arg\max_{\boldsymbol{\Sigma}_k} Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^t)$$

To do so, taking partial derivative $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^t)$ with respect to $\boldsymbol{\Sigma}_k^{-1}$ we have:

$$\left.\frac{\partial Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^t)}{\partial \boldsymbol{\Sigma}_k^{-1}}\right|_{\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}_k^{t+1}} = \sum_i \omega_{i,k}^t \left.\frac{\partial[-\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\ \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]}{\partial\boldsymbol{\Sigma}_k^{-1}}\right|_{\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}_k^{t+1}} = 0 \qquad (3)$$

On the other hand for the symmetric matrices of $\boldsymbol{\Sigma}$ and its inverse and every vector a, we have:

$$\frac{\partial\log|\boldsymbol{\Sigma}|}{\partial\boldsymbol{\Sigma}^{-1}} = \frac{\partial\log|\boldsymbol{\Sigma}|}{\partial|\boldsymbol{\Sigma}|}\frac{\partial|\boldsymbol{\Sigma}|}{\partial\boldsymbol{\Sigma}^{-1}} = \frac{1}{|\boldsymbol{\Sigma}|}(-|\boldsymbol{\Sigma}|\boldsymbol{\Sigma}^T) = -\boldsymbol{\Sigma}^T$$

$$\frac{\partial\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}}{\partial\boldsymbol{\Sigma}^{-1}} = \boldsymbol{a}\boldsymbol{a}^T$$

Therefore, from (3) we conclude:

$$\left.\frac{\partial Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^t)}{\partial\boldsymbol{\Sigma}_k^{-1}}\right|_{\boldsymbol{\Sigma}_k=\boldsymbol{\Sigma}_k^{t+1}} = \frac{1}{2}\sum_i \omega_{i,k}^t[\boldsymbol{\Sigma}_k^{t+1} - (\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T] = 0 \Rightarrow \sum_i \omega_{i,k}^t\boldsymbol{\Sigma}_k^{t+1} = \sum_i \omega_{i,k}^t(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T$$

Clearly, the result is dependent on $\boldsymbol{\mu}_k$. So, for the optimization, we should substitute that with the updated value of $\boldsymbol{\mu}_k^{t+1}$ that was determined before:

$$\boldsymbol{\Sigma}_k^{t+1} = \frac{\sum\limits_i \omega_{i,k}^t(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{t+1})(\boldsymbol{x}_i-\boldsymbol{\mu}_k^{t+1})^T}{\sum\limits_i \omega_{i,k}^t}$$

Updating π_k :

the target is to find

$$\pi_k^{t+1} = \arg\max_{\pi_k} Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^t) = \arg\max_{\pi_k} \sum_i\sum_k \omega_{i,k}^t \log\pi_k$$

In which, π_k is constrained by $\sum\limits_k \pi_k = 1$. So, to find the local maxima in this problem we use the method of *Lagrange Multipliers*. Constructing the needed Lagrangian we have:

$$\mathcal{L}(\pi_k,\lambda) = \sum_i\sum_k \omega_{i,k}^t\log\pi_k + \lambda(\sum_k \pi_k - 1)$$

$$\frac{\partial \mathcal{L}(\pi_k, \lambda)}{\partial \pi_k} = 0 \Rightarrow \sum_i \omega_{i,k}^t \frac{1}{\pi_k} + \lambda = 0 \Rightarrow \pi_k = -\frac{\sum_i \omega_{i,k}^t}{\lambda} \qquad (4)$$

Using the constraint we have:

$$\sum_k \pi_k = 1 \Rightarrow -\frac{\sum_i \sum_k \omega_{i,k}^t}{\lambda} = -\frac{\sum_i 1}{\lambda} = 1 \Rightarrow \lambda = -N$$

So, from (4) we conclude:

$$\pi_k^{t+1} = \frac{\sum_i \omega_{i,k}^t}{N}$$

Question 5

Solution:

1.

Assume that we have a data set of N observed d-dimensional points $\{\boldsymbol{x}_i\}_{i=1}^N$.
the data set follows a mixture of K Categorical distributions such that:

$$p(\boldsymbol{x}|\Theta) = \sum_k \pi_k \ \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}_k)$$

In which π_k and $\boldsymbol{\theta}_k$ are the prior probability and the parameter vector of the kth Categorical model. Where each $\boldsymbol{\theta} = (\theta_{k,1}, \cdots, \theta_{k,d})$ is a probability distribution over the labels $\boldsymbol{\mathcal{A}} = \{1, \cdots, d\}$.
For example, if $x_i = c; \quad 1 \leq c \leq d$, then we can say $p(x_i|\boldsymbol{\theta}_k) = \theta_{k,c}$.
But we assumed that each data point is d-dimensional. So, we can represent $\boldsymbol{x}_i$ as a d-dimensional vector such that for example if $x_i = 3$, we can show that as turning on the third component of the data point vector.$(\boldsymbol{x}_i = (0, 0, 1, 0, \cdots, 0) \ and \ x_{i,3} = 1)$

Our task is to estimate the set of parameters $(\pi_k, \boldsymbol{\theta}_k)$ for each Categorical model using the data set.

Clearly at the first step, we should initialize mentioned parameters. Here we choose an algorithm to do so. In this algorithm, we initialize the parameters π_k uniformly:

$$\pi_k^0 = \frac{1}{K} \qquad \forall k; \quad 1 \leq k \leq K$$

Which means each point in the data set is equally likely to belong to the kth Categorical model. Also, in order to initialize the Categorical models' parameter vector $\boldsymbol{\theta}_k$, we first divide the data set randomly into K clusters with $M = \frac{N}{K}$ data points. (Notice if the value of M wasn't integer, we would divide the data set to clusters such that all the clusters approximately have the same sizes.)Then using MLE method we would have:

$$\theta_{k,j}^0 = \frac{N_j^{(k)}}{M} \qquad \forall k; \quad 1 \leq k \leq K$$

Where, $N_j^{(k)}$ denotes the number of times that the jth component of the data point vectors in the kth cluster, is turned on.
equivalently

$$\boldsymbol{\theta}_k^0 = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{x}_i^{(k)}$$

2.

In this section we are supposed to determine complete data set likelihood. According to descriptions at the previous part we have:

$$p(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) = \prod_i p(\boldsymbol{x}_i, z_i|\boldsymbol{\Theta}) = \prod_i \sum_k p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta}) = \prod_i \sum_k p(z_i = k) p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta}) = \prod_i \sum_k \pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$$

$$= \prod_i \sum_k p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta}) \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{p(\boldsymbol{x}_i, z_i = k|\boldsymbol{\Theta})} = \prod_i \sum_k \omega_{i,k} \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\omega_{i,k}}$$

In which, using $Bayes' \ Rule$ we have:

$$\omega_{i,k} \triangleq p^*(z_i = k|\boldsymbol{x}_i, \boldsymbol{\Theta}) = \frac{p(z_i = k) p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta})}{\sum_{k'} p(z_i = k') p(\boldsymbol{x}_i|z_i = k', \boldsymbol{\Theta})} = \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\sum_{k'} \pi_{k'} \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_{k'})}$$

Therefore, the complete log likelihood function is determined in this way:(this part is for the next section)

$$L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}) = \sum_i \log \sum_k \omega_{i,k} \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\omega_{i,k}} \geq \sum_i \sum_k \omega_{i,k} \log \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\omega_{i,k}} \qquad *$$

3.

Assume that the latent parameters at the iteration t, is defined as calculated before

$$\omega^t_{i,k} \triangleq p^*(z_i = k|\boldsymbol{x}_i, \boldsymbol{\Theta}^t) = \frac{\pi^t_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}^t_k))}{\sum_{k'} \pi^t_{k'} \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}^t_{k'})}$$

Using the results in the previous parts and the source book, we can define an instance iterative target function(the lower bound function) to work with:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \sum_i \sum_k \omega^t_{i,k} \log \frac{\pi_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{\omega^t_{i,k}}$$

$$= \sum_i \sum_k \omega^t_{i,k} \log \left(\frac{\pi_k \ \prod_{j=1}^{d} (\theta_{k,j})^{x_{i,j}}}{\omega^t_{i,k}} \right)$$

Which obviously holds the two constraint of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) \leq L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta})$ (based on $*$) and $Q(\boldsymbol{\Theta}^t, \boldsymbol{\Theta}^t) = L(\boldsymbol{\mathcal{D}}|\boldsymbol{\Theta}^t)$ (based on the proof in source). We expand the result more, to use that in further steps

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \sum_i \sum_k \omega^t_{i,k} [\log \pi_k - \log \omega^t_{i,k} + \underbrace{\sum_{j=1}^{d} x_{i,j} \log \theta_{k,j}]}_{\boldsymbol{x}_i \cdot \log \boldsymbol{\theta}_k} \qquad (1)$$

Now we go for the E-step and M-step

E-step:

The latent parameters to be estimated in each iteration of E-step have been determined in a closed-form:

$$\omega^t_{i,k} = \frac{\pi^t_k \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}^t_k))}{\sum_{k'} \pi^t_{k'} \ \text{Cat}(\boldsymbol{x}_i|\boldsymbol{\theta}^t_{k'})}$$

Where, ω is a $N \times K$ matrix.

M-step:
Here we are going to update the Θ parameters using maximization of $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$ by parameters:

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t)$$

Using the expanded equation in (1) we conclude:

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} \sum_i \sum_k \omega_{i,k}^t [\log \pi_k - \log \omega_{i,k}^t + \sum_{j=1}^d x_{i,j} \log \theta_{k,j}]$$

Updating π_k :
the target is to find

$$\pi_k^{t+1} = \arg\max_{\pi_k} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \arg\max_{\pi_k} \sum_i \sum_k \omega_{i,k}^t \log \pi_k$$

In which, π_k is constrained by $\sum\limits_k \pi_k = 1$.
As you can see, the optimization problem is exactly like the one in the previous question(naturally, we mean the Updating π_k part of that.); So, the result is the same:

$$\pi_k^{t+1} = \frac{\sum\limits_i \omega_{i,k}^t}{N}$$

Updating $\boldsymbol{\theta}_k$:
the target is to find

$$\theta_{k,j}^{t+1} = \arg\max_{\theta_{k,j}} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^t) = \arg\max_{\theta_{k,j}} \sum_i \sum_k \omega_{i,k}^t \sum_{j=1}^d x_{i,j} \log \theta_{k,j} = \arg\max_{\theta_{k,j}} \sum_i \omega_{i,k}^t \ x_{i,j} \log \theta_{k,j}$$

In which, $\theta_{k,j}$ is constrained by $\sum_{j=1}^d \theta_{k,j} = 1$.
So, to find the local maxima in this problem we use the method of *Lagrange Multipliers*. Constructing the needed Lagrangian we have:

$$\mathcal{L}(\theta_{k,j}, \lambda) = \sum_i \omega_{i,k}^t \ x_{i,j} \log \theta_{k,j} + \lambda(\sum_{j=1}^d \theta_{k,j} - 1)$$

$$\frac{\partial \mathcal{L}(\theta_{k,j}, \lambda)}{\partial \theta_{k,j}} = 0 \Rightarrow \sum_i \omega_{i,k}^t \ x_{i,j} \frac{1}{\theta_{k,j}} + \lambda = 0 \Rightarrow \theta_{k,j} = -\frac{\sum\limits_i \omega_{i,k}^t \ x_{i,j}}{\lambda} \tag{4}$$

Using the constraint we have:

$$\sum_{j=1}^d \theta_{k,j} = 1 \Rightarrow -\frac{\sum\limits_i \sum\limits_{j=1}^d \omega_{i,k}^t \ x_{i,j}}{\lambda} = -\frac{\sum\limits_i \omega_{i,k}^t \sum\limits_{j=1}^d x_{i,j}}{\lambda} = -\frac{\sum\limits_i \omega_{i,k}^t}{\lambda} = 1 \Rightarrow \lambda = -\sum_i \omega_{i,k}^t$$

So, from (4) we conclude:

$$\theta_{k,j}^{t+1} = \frac{\sum\limits_i \omega_{i,k}^t \ x_{i,j}}{\sum\limits_i \omega_{i,k}^t}$$