Assignment 1: EM for Categorical Data Advanced Signal Processing

Daniel Barrejón Moreno

January 14, 2019

1 Problem formulation

In this project we want to study a set of documents with a model following a mixture of categorical distributions. For a document $\mathbf{x}$ belonging to a set of documents $\mathbf{X}$, the marginal li,kelihood is expressed as follows

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \text{Cat}(x_j|\boldsymbol{\theta}_k), \tag{1}$$

where K is the number of mixtures (in our case the topics), D is the number of words that appear in a certain document and $\text{Cat}(x_j|\boldsymbol{\theta}_k)$ is the categorical distribution with I categories and parameter $\boldsymbol{\theta}_k$ which represents the probability that a certain category appears. Bear in mind that the model parameters are defined as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, \boldsymbol{\pi}\}$. Since $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,m}, \ldots, \theta_{k,I})$ is a vector with the probabilities of topic k probability, it must satisfy two constraints:

$$0 < \theta_{k,m} < 1 \tag{2}$$

$$\text{and} \quad \sum_{m=1}^{I} \theta_{k,m} = 1. \tag{3}$$

The expression for the log li,kelihood of the observed data $\mathcal{D}$ looks li,ke this

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{D} \text{Cat}(x_j|\boldsymbol{\theta}_k). \tag{4}$$

However this function is hard to optimize due to the sum inside the log. In order to solve this we introduce the latent variable $\mathcal{Z}$. The marginal distribution for the hidden variable z_i is defined by the mixing coefficient of the mixture π_k as follows

$$p(z_i = k) = \pi_k. \tag{5}$$

Since $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ is a vector with the probabilities for the mixtures, certain restrictions must be satisfied:

$$0 < \pi_k < 1 \tag{6}$$

$$\text{and} \quad \sum_{k=1}^{K} \pi_k = 1. \tag{7}$$

Now we need to find the expression for the complete data log li,kelihood $p(\mathcal{D}, \mathcal{Z})$. First of all, the marginal of z_i is given by

$$p(z_i|\boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}\{z_i=k\}}, \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The probability density function for x_i given z_i is given by the following expression

$$p(\mathbf{x}_i|z_i, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k)^{\mathbb{I}\{z_i=k\}}. \tag{9}$$

Since we are interested in the joint probability, applying Bayes' rule

$$p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) = p\left(\mathbf{x}_i|z_i, \boldsymbol{\theta}\right) p(z_i|\boldsymbol{\theta}) \tag{10}$$

yields the resulting expression

$$p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) = \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right)^{\mathbb{I}\{z_i=k\}}. \tag{11}$$

2 Complete data log li,kelihood $l_c(\boldsymbol{\theta})$

Assuming that our observed data is independent and identically distributed (i.i.d) and taking natural logarithm we can find expression of the complete data log li,kelihood

$$l_c(\boldsymbol{\theta}) = \log p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) = \log \prod_{n=1}^{N} p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) \tag{12}$$

$$= \sum_{i=1}^{N} \log p\left(\mathbf{x}_i, z_i|\boldsymbol{\theta}\right) \tag{13}$$

$$= \sum_{i=1}^{N} \log \prod_{k=1}^{K} \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right)^{\mathbb{I}\{z_i=k\}} \tag{14}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right). \tag{15}$$

We can simplify the expression for the marginal of $\mathbf{x}_i$ as

$$p(\mathbf{x}_i) = \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}) = \prod_{j=1}^{D}\prod_{m=1}^{I} \theta_m^{\mathbb{I}\{x_{i,j}=m\}} = \prod_{m=1}^{I}\prod_{j=1}^{D} \theta_m^{\mathbb{I}\{x_{i,j}=m\}} \tag{16}$$

$$= \prod_{m=1}^{I} \theta_m^{\sum_{j=1}^{D}\mathbb{I}\{x_{i,j}=m\}} = \prod_{m=1}^{I} \theta_m^{\mu_{i,m}}, \tag{17}$$

where we have defined a new metric $\mu_{i,m}$ that represents the number of times the word associated to the category m appears at document i, *i.e.* ,

$$\mu_{i,m} = \sum_{j=1}^{D} \mathbb{I}(x_{i,j} = m). \tag{18}$$

Therefore, the final expression will be

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \left(\pi_k \prod_{j=1}^{D} \text{Cat}(x_{ij}|\boldsymbol{\theta}_k) \right) \tag{19}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \prod_{m=1}^{I} \theta_{k,m}^{\mu_{i,m}} \tag{20}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(z_i = k) \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m}. \tag{21}$$

3 ML Inference

In order to solve the maximum li,kelihood problem, we define an auxiliary function Q which will be the expected complete data log li,kelihood over the hidden variable $\mathcal{Z}$. This function Q will be evaluated in the E-step and maximized for the model parameters $\boldsymbol{\theta}$ to update the model parameters in each iteration of the algorithm.

3.1 E-step: $Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ for ML inference

Taking the expectation with respect to the latent variables z we get the following expected complete data log-li,kelihood.

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \mathbb{E}_Z\{l_c(\boldsymbol{\theta})|\mathcal{D}, \boldsymbol{\theta}^{t-1}\} \tag{22}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} \tag{23}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m}, \tag{24}$$

where we have defined the metric $r_{i,k} \triangleq p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1})$ [1] as the responsibility that mixture k, *i.e.* topic k, takes at explaining the document $\mathbf{x}_i$. It is defined as follows

$$r_{i,k} = \mathbb{E}_Z\{\mathbb{I}(z_i = k)\} = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \tag{25}$$

$$= \frac{p(z_i = k, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{t-1})} = \frac{p(z_i = k, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})}{\sum_{k\prime}^{K} p(z_i = k\prime, \mathbf{x}_i|\boldsymbol{\theta}^{t-1})} \tag{26}$$

$$= \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)}{\sum_{k\prime}^{K} \pi_{k\prime} p(\mathbf{x}_i|\boldsymbol{\theta}_{k\prime})} = \frac{\pi_k \prod_{j=1}^{D} \mathrm{Cat}(x_{i,j}|\boldsymbol{\theta}_k)}{\sum_{k\prime}^{K} \pi_{k\prime} \prod_{j=1}^{D} \mathrm{Cat}(x_{i,j}|\boldsymbol{\theta}_{k\prime})}. \tag{27}$$

This quantity must satisfy the following constrains, given that z it is a probability

$$0 \leq r_{i,k} \leq 1, \tag{28}$$

$$\text{and} \quad \sum_{k=1}^{K} r_{i,k} = 1. \tag{29}$$

3.1.1 Arithmetic underflow solved by log-sum-exp trick

In our problem we have I categories which correspond to the words from a dictionary. When we perform the product in Equation 25 we obtain really small values that cannot be represented by the computer. This problem is known as **arithmetic underflow**. However, in order to maintain the range in probability if categories $\theta_{k,m}$ that influence the value of $r_{i,k}$ we use the **log-sum-exp trick** [2].
Taking the logarithm from Equation 25 and using Equation 16 for $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$ we obtain the following expression

$$\log r_{i,k} = \log \pi_k \prod_{j=1}^{D} \mathrm{Cat}(x_{i,j}|\boldsymbol{\theta}_k) - \log \sum_{k\prime=1}^{K} \pi_{k\prime} \prod_{j=1}^{D} \mathrm{Cat}(x_{i,j}|\boldsymbol{\theta}_{k\prime}) \tag{30}$$

$$= \log \pi_k + \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} - \log \sum_{k\prime=1}^{K} \pi_{k\prime} \prod_{m=1}^{I} \theta_{k\prime,m}^{\mu_{i,m}}. \tag{31}$$

The log-sum-exp trick will be applied on the last term from the above equation. The trick states that

$$\log \sum_{v=1}^{V} e^{g_v} = a + \log \sum_{v=1}^{V} e^{g_v - a}, \tag{32}$$

where $a = \max_v g_v$ and g_v is defined as

$$g_v = \log \pi_{k\prime} \prod_{m=1}^{I} \theta_{k\prime,m}^{\mu_{i,m}} = \log \pi_{k\prime} + \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k\prime,m} - \tag{33}$$

Once $\log r_{i,k}$ from Equation 30 is known, we can find $r_{i,k}$ with an exponential. With this trick we do not loose any information by forcing values assigned to be 0 by the computer using clipping and we also increase computational speed.

3.2 M-step for ML inference

Now we need to find the closed-form formulas to update the model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k, \boldsymbol{\pi}\}$. Since we must tackle a maximization problem with constraints we will apply Lagrange Multipliers to solve equation

$$\boldsymbol{\theta}^t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1}\right). \tag{34}$$

3.2.1 Maximization of π_k

Using the constraints on π from Equations 6 we propose as Lagrangian the function

$$L\left(Q(\pi_k), \lambda\right) = Q(\pi_k) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right), \tag{35}$$

which will be optimized in this way

$$\min_{\lambda} \max_{\pi_k} \{L\left(Q(\pi_k), \lambda\right)\}. \tag{36}$$

First, we take the derivative of Equation 35 w.r.t π_k and equate it to 0, which yields

$$\frac{\partial L}{\partial \pi_k} = 0 = \sum_{i=1}^{N} \frac{r_{i,k}}{\pi_k} - \lambda, \tag{37}$$

$$\pi_k = \frac{1}{\lambda} \sum_{i=1}^{N} r_{i,k}. \tag{38}$$

Now, we take the derivate w.r.t λ which yields

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0, \tag{39}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{40}$$

If we sum over k at both sides of Equation 38

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{1}{\lambda} \sum_{i=1}^{N} r_{i,k} \tag{41}$$

$$1 = \frac{1}{\lambda} \sum_{i=1}^{N} \sum_{k=1}^{K} r_{i,k}, \tag{42}$$

and using the constraints on $r_{i,k}$ from Equations 28 we get the value of λ

$$1 = \frac{1}{\lambda} \sum_{i=1}^{N} 1. \tag{43}$$

$$\lambda = N. \tag{44}$$

Knowing the value of λ the estimated value of $\hat{\pi}_k$ can be expressed as follows

$$\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} r_{i,k}s = \frac{N_k}{N}, \tag{45}$$

where

$$N_k = \sum_{i=1}^{N} r_{i,k}. \tag{46}$$

This result is actually intuitive since N_k represents the 'weight' of topic k at explaining the documents, and therefore π_k is just the percentage of topic k at explaining the data.

3.2.2 Maximization of $\boldsymbol{\theta}_k$

We follow a similar approach. Now, using the constraints on $\boldsymbol{\theta}_k$ from Equations 2 we propose as Lagrangian the following function

$$L\left(Q(\theta_{k,m}), \lambda\right) = Q(\theta_{k,m}) + \lambda\left(\sum_{m=1}^{I}\theta_{k,m} - 1\right), \tag{47}$$

which will be optimized as a min-max problem

$$\min_{\lambda}\max_{\theta_{km}}\{L\left(Q(\theta_{km}), \lambda\right)\}. \tag{48}$$

Firstly, we take the derivative of Equation 47 w.r.t $\theta_{k,m}$ and equate it to 0, which yields

$$\frac{\partial L}{\partial \theta_{k,m}} = \sum_{i=1}^{N}\frac{r_{i,k}\mu_{i,m}}{\theta_{k,m}} - \lambda = 0, \tag{49}$$

$$\theta_{k,m} = \frac{1}{\lambda}\sum_{i=1}^{N} r_{i,k}\mu_{i,m}. \tag{50}$$

Secondly, we take the derivative w.r.t λ, which yields

$$\frac{\partial L}{\partial \lambda} = \sum_{m=1}^{I}\theta_{k,m} - 1 = 0, \tag{51}$$

$$\sum_{m=1}^{I}\theta_{k,m} = 1. \tag{52}$$

If now we sum over I at both sides of Equation 50, the value of λ is obtained

$$\sum_{m=1}^{I}\theta_{k,m} = \sum_{m=1}^{I}\frac{1}{\lambda}\sum_{i=1}^{N} r_{i,k}\mu_{i,m} \tag{53}$$

$$1 = \frac{1}{\lambda}\sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}, \tag{54}$$

$$\lambda = \sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}. \tag{55}$$

Using the value of λ in Equation 50, $\theta_{k,m}$ is obtained

$$\hat{\theta}_{k,m} = \frac{\sum_{i=1}^{N} r_{i,k}\mu_{i,m}}{\sum_{i=1}^{N}\sum_{m=1}^{I} r_{i,k}\mu_{i,m}}. \tag{56}$$

Again, the result is quite intuitive. $\theta_{k,m}$ is just the average of the category m weighted by the responsibility $r_{i,k}$.

4 MAP inference

Maximum li,kelihood estimations is an estimation that tends to overfit [1]. A solution to such problem is applying maximum a posteriori estimation (MAP). In this case, we do not only consider the li,kelihood, but also some prior information on the model parameters $\boldsymbol{\theta}$. From Bayes' rule we know that

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{Z}) \propto p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{57}$$

If we take logarithms at both sides of the equation we get

$$\log p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{Z}) \propto \log p(\mathcal{D}, \mathcal{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}). \tag{58}$$

Notice that the first term is just the complete data log li,kelihood from Equation 21 and the second term corresponds to the prior information. The goal now is to define some prior over the model parameters $\boldsymbol{\theta}$ and use it on the function Q from Equation 24 to work with the posterior instead of the li,kelihood. Afterwards, we need to reformulate Equation 45 and 56 for π_k and $\theta_{k,m}$ to take into account the priors.

4.1 E-step: $Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1})$ for MAP inference

From [3] and [1] we know it is natural that the prior on the mixture weights π and the category probabilities $\boldsymbol{\theta}_k$ follow a Dirichlet distribution, *i.e.* ,

$$\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\beta}), \quad \text{s.t.} \quad p(\boldsymbol{\pi}|\boldsymbol{\beta}) = \frac{1}{\text{B}(\boldsymbol{\beta})}\prod_{k=1}^{K}\pi_k^{\beta_k - 1}, \tag{59}$$

$$\boldsymbol{\theta}_k \sim \text{Dir}(\boldsymbol{\alpha}), \quad \text{s.t.} \quad p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}) = \frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_{m=1}^{I}\theta_{k,m}^{\alpha_m - 1}, \tag{60}$$

where the function $\text{B}(\cdot)$ stands for the Beta function, and the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the parameters or hyperpriors of the Dirichlet distributions for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k$ respectively.

The expression for the new Q function is as follows

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} + \log p(\boldsymbol{\pi}|\boldsymbol{\beta}) + \sum_{k=1}^{K} \log p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}), \tag{61}$$

where $r_{i,k}$ remains the same. Taking logarithm on the probability of the priors we get

$$\log p(\boldsymbol{\pi}|\boldsymbol{\beta}) = \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\prod_{k=1}^{K}\pi_k^{\beta_k-1}\right) = \log\frac{1}{\mathrm{B}(\boldsymbol{\beta})} + \sum_{k=1}^{K}(\beta_k-1)\log(\pi_k), \tag{62}$$

$$\log p(\boldsymbol{\theta}_k|\boldsymbol{\alpha}) = \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{m=1}^{I}\theta_{k,m}^{\alpha_m-1}\right) = \log\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} + \sum_{m=1}^{I}(\alpha_m-1)\log(\theta_{k,m}). \tag{63}$$

Using these results in Equation 61 we get the expression for $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{t-1})$ using MAP estimation

$$\begin{aligned} Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t-1}) = &\sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \log \pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} r_{i,k} \sum_{m=1}^{I} \mu_{i,m} \log \theta_{k,m} \\ &+ \log\left(\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\right) + \sum_{k=1}^{K}(\beta_k-1)\log(\pi_k) \\ &+ \sum_{k=1}^{K}\log\frac{1}{\mathrm{B}(\boldsymbol{\alpha})} + \sum_{k=1}^{K}\sum_{m=1}^{I}(\alpha_m-1)\log(\theta_{k,m}). \end{aligned} \tag{64}$$

Notice that $\mathrm{B}(\boldsymbol{\alpha})$ can be decomposed as

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},$$

where $\Gamma(\cdot)$ is the Gamma function. The same applies for $\mathrm{B}(\boldsymbol{\beta})$

4.2 M step for MAP inference

We will follow the same procedure as for the ML case; but now we must use function 64 instead to take into account the priors.

4.2.1 MAP estimation of π_k

Again, we have the same maximization problem and hence we propose the same Lagrangian as in Equation 35 and we use the same constraints on π_k

$$L\left(Q(\pi_k), \lambda\right) = Q(\pi_k) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right). \tag{65}$$

Again, we derivative Equation 64 w.r.t π_k and equate it to 0

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{N} \frac{r_{i,k}}{\pi_k} + \frac{\beta_k - 1}{\pi_k} - \lambda = 0, \tag{66}$$

$$\pi_k = \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} + \beta_k - 1 \right). \tag{67}$$

Now we derivate w.r.t λ

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0, \tag{68}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{69}$$

Summing over K at both sides of Equation 67

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} + \beta_k - 1 \right) \tag{70}$$

$$1 = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{N} r_{i,k} + \frac{1}{\lambda} \sum_{k=1}^{K} \beta_k - \frac{1}{\lambda} \sum_{k=1}^{K} 1, \tag{71}$$

we get the value of λ

$$\lambda = N + \sum_{k=1}^{K} \beta_k - K. \tag{72}$$

Once λ is known the estimated value of π_k is as follows

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} r_{i,k} + \beta_k - 1}{N + \sum_{k=1}^{K} \beta_k - K}. \tag{73}$$

4.2.2 MAP estimation of $\theta_{k,m}$

For the estimation of $\theta_{k,m}$ we use the same Lagrangian from Equation 47 and the same restrictions from 2

$$L\left(Q(\theta_{k,m}), \lambda\right) = Q(\theta_{k,m}) + \lambda \left(\sum_{m=1}^{I} \theta_{k,m} - 1 \right). \tag{74}$$

As before, we first derivate w.r.t $\theta_{k,m}$, which yields

$$\frac{\partial L}{\partial \theta_{k,m}} = \sum_{i=1}^{N} \frac{r_{i,k}\mu_{im}}{\theta_{k,m}} + \frac{\alpha_m - 1}{\theta_{km}} - \lambda = 0, \tag{75}$$

$$\theta_{k,m} = \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k}\mu_{i,m} + \alpha_m - 1 \right). \tag{76}$$

And later with respect to λ

$$\frac{\partial L}{\partial \lambda} = 0 = \sum_{m=1}^{I} \theta_{k,m} - 1, \tag{77}$$

$$\sum_{m=1}^{I} \theta_{k,m} = 1. \tag{78}$$

Summing at both sides over I Equation 76 we can obtain the value of λ

$$\sum_{m=1}^{I} \theta_{k,m} = \sum_{m=1}^{I} \frac{1}{\lambda} \left(\sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \alpha_m - 1 \right) \tag{79}$$

$$1 = \frac{1}{\lambda} \sum_{m=1}^{I} \left(\sum_{i=1}^{N} \frac{r_{i,k} \mu_{i,m}}{\theta_{k,m}} + \frac{\alpha_m}{\theta_{k,m}} - 1 \right), \tag{80}$$

$$\lambda = \sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - \sum_{m=1}^{I} 1, \tag{81}$$

$$\lambda = \sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - I. \tag{82}$$

And with that value of λ the estimated value of $\theta_{k,m}$ is

$$\hat{\theta}_{k,m} = \frac{\sum_{i=1}^{N} r_{i,k} \mu_{im} + \alpha_m - 1}{\sum_{m=1}^{I} \sum_{i=1}^{N} r_{i,k} \mu_{i,m} + \sum_{m=1}^{I} \alpha_m - I}. \tag{83}$$

5 Experiments

5.1 ML Experiments

INCLUDE FIGURES

5.2 MAP Experiments

INCLUDE FIGURES

References

[1] K. P. Murphy, *Machine learning: A probabilistic perspective. adaptive computation and machine learning*, 2012.

[2] R. Eisele, *The log-sum-exp trick in machine learning*. [Online]. Available: `https://www.xarg.org/2016/06/the-log-sum-exp-trick-in-machine-learning/`.

[3] A. Artés-Rodríguez, *Notes for advanced signal processing*, 2018.

[4] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of classification*, vol. 24, no. 2, pp. 155–181, 2007.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. [Online]. Available: `http://research.microsoft.com/en-us/um/people/cmbishop/prml/`.