

Advanced Probabilistic Machine Learning and Applications

Solution Tutorial 2: Categorical Mixture Model (CMM) + EM

Exercise 1: Compute analytical forms of the the E and M steps for the EM-Algorithm

a. *Complete data log-likelihood:* we describe the expression for the complete data log-likelihood for the mixture of multinomials models $l_c(\boldsymbol{\theta})$. Since all datapoints are considered i.i.d. we can write the log-likelihood as a sum

$$\begin{aligned} l_c(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}) &= \log p(\{\mathbf{x}_n\}_{n=1}^N, \mathbf{z} | \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}) = \sum_{n=1}^N \log(p(\mathbf{x}_n | z_n, \{\boldsymbol{\theta}_k\}) p(z_n | \boldsymbol{\pi})) = \sum_{n=1}^N \log \left(\prod_{k=1}^K (p(z_n = k) p(\mathbf{x}_n | \boldsymbol{\theta}_k))^{[z_n=k]} \right) \\ &= \sum_{n=1}^N \log \left(\prod_{k=1}^K (\pi_k p(\mathbf{x}_n | \boldsymbol{\theta}_k))^{[z_n=k]} \right) = \sum_{n=1}^N \sum_{k=1}^K [z_n = k] \log(\pi_k p(\mathbf{x}_n | \boldsymbol{\theta}_k)) \end{aligned}$$

Then, if we write the expression for $p(\mathbf{x}_n | \boldsymbol{\theta}_k) = \prod_{j=1}^{W_n} \text{Cat}(x_{nj} | \boldsymbol{\theta}_k)$ we can rewrite the previous expression

$$\begin{aligned} l_c(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}) &= \sum_{n=1}^N \sum_{k=1}^K [z_n = k] \log \left(\pi_k \prod_{j=1}^{W_n} \text{Cat}(x_{nj} | \boldsymbol{\theta}_k) \right) = \sum_{n=1}^N \sum_{k=1}^K [z_n = k] \log \left(\pi_k \prod_{j=1}^{W_n} \prod_{m=1}^I \theta_{km}^{[x_{nj}=m]} \right) \\ &= \sum_{n=1}^N \sum_{k=1}^K [z_n = k] \log(\pi_k) + \sum_{n=1}^N \sum_{k=1}^K \sum_{j=1}^{W_n} \sum_{m=1}^I [z_n = k] \log(\theta_{km}^{[x_{nj}=m]}) \end{aligned}$$

So finally, the expresion of the complete data log-likelihood is

$$l_c(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}) = \sum_{n=1}^N \sum_{k=1}^K [z_n = k] \log(\pi_k) + \sum_{n=1}^N \sum_{k=1}^K \sum_{j=1}^{W_n} \sum_{m=1}^I [z_n = k] [x_{nj} = m] \log(\theta_{km}) \quad (1)$$

b. *E step:* We derive the expressions to perform the estimation of the parameters using the Expectation-Maximization algorithm. Firstly, we compute the E-step of the EM Algorithm, that is, we find an expression for $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = E_z [l_c(\boldsymbol{\theta}) | D, \boldsymbol{\theta}^{\text{old}}]$. To improve the readability of the equations, z will be omitted from the expectations. Starting from Equation 1

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K E[[z_n = k]] \log(\pi_k) + \sum_{n=1}^N \sum_{k=1}^K \sum_{j=1}^{W_n} \sum_{m=1}^I E[[z_n = k]] [x_{nj} = m] \log(\theta_{km})$$

where $E[[z_n = k]] = p(z_n = k | \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}})$. Given $\boldsymbol{\theta}^{\text{old}}$, the conditional distribution of z_n is computed using Bayes Theorem.

$$p(z_n = k | \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \frac{p(z_n = k, \mathbf{x}_n | \boldsymbol{\theta}^{\text{old}})}{p(\mathbf{x}_n | \boldsymbol{\theta}^{\text{old}})} = \frac{\pi_k \prod_{j=1}^{W_n} \text{Cat}(x_{nj} | \boldsymbol{\theta}_k)}{\sum_{k'=1}^K \pi_{k'} \prod_{j=1}^{W_n} \text{Cat}(x_{nj} | \boldsymbol{\theta}_{k'})} = r_{nk}$$

Thus, the expected complete data log-likelihood has the following expression:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \log(\pi_k) + \sum_{n=1}^N \sum_{k=1}^K \sum_{j=1}^{W_n} \sum_{m=1}^I r_{nk} [x_{nj} = m] \log(\theta_{km})$$

This can be rewritten as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \log(\pi_k) + \sum_{n=1}^N \sum_{k=1}^K r_{nk} \sum_{j=1}^{W_n} \sum_{m=1}^I [x_{nj} = m] \log(\theta_{km})$$

c. *M step:* we present the ML estimation of the new set of parameters $\boldsymbol{\theta}$.

Regarding $\boldsymbol{\pi}$, we want to maximise $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to each π_k taking into consideration the restriction $\sum_{k=1}^K \pi_k = 1$. Thus, the method of Lagrange multipliers will be used to perform the optimisation.

$$L(\pi_k, \lambda) = Q(\pi_k) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) \quad \rightarrow \quad \min_{\lambda} \max_{\pi_k} \{L(\pi_k, \lambda)\}$$

First, we differentiate $L(\pi_k, \lambda)$ and set it equal to 0.

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^N \frac{r_{nk}}{\pi_k} + \lambda = 0 \rightarrow \pi_k = \frac{1}{-\lambda} \sum_{n=1}^N r_{nk}$$

Now we calculate the value of lambda

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^K \pi_k - 1 = 0 \rightarrow - \sum_{k=1}^K \frac{1}{\lambda} \sum_{n=1}^N r_{nk} - 1 = 0 \rightarrow \frac{1}{\lambda} \sum_{n=1}^N \sum_{k=1}^K r_{nk} = -1 \rightarrow \lambda = - \sum_{n=1}^N 1 = -N$$

Thus, the ML estimate of π_k is

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^N r_{nk}$$

Regarding $\boldsymbol{\theta}_k$, want to maximise $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to each θ_{km} taking into consideration the restriction $\sum_{m=1}^I \theta_{km} = 1$. Thus, the method of Lagrange multipliers will be used to perform the optimisation.

$$L(\theta_{km}, \lambda) = Q(\theta_{km}) + \lambda \left(\sum_{m=1}^I \theta_{km} - 1 \right) \quad \rightarrow \quad \min_{\lambda} \max_{\pi_k} \{L(\theta_{km}, \lambda)\}$$

First, we differentiate $L(\theta_{km}, \lambda)$ and set equal to 0.

$$\frac{\partial L}{\partial \theta_{km}} = \sum_{n=1}^N \sum_{j=1}^{W_n} r_{nk} [x_{nj} = m] \frac{1}{\theta_{km}} + \lambda = 0 \rightarrow \theta_{km} = \frac{1}{-\lambda} \sum_{n=1}^N \sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}$$

Now we calculate the value of lambda

$$\frac{\partial L}{\partial \lambda} = \sum_{m=1}^I \theta_{km} - 1 = 0 \rightarrow - \sum_{m=1}^I \frac{1}{\lambda} \sum_{n=1}^N \sum_{j=1}^{D_i} [x_{nj} = m] r_{nk} = 1 \rightarrow \lambda = - \sum_{m=1}^I \sum_{n=1}^N \sum_{j=1}^{D_i} [x_{nj} = m] r_{nk}$$

Thus, the ML estimate of θ_{km} is

$$\hat{\theta}_{km} = \frac{\sum_{n=1}^N \sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}}{\sum_{m=1}^I \sum_{n=1}^N \sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}}$$