Advanced Probabilistic Machine Learning and Applications

Martina Contisciani and Caterina De Bacco

November 2, 2021

Contents

1 Tutorial 2: Mixture Models and EM

1.1 Exercise 1: Categorical Mixture Model (CMM)

The dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}^\top$ describes a set of N documents, here tweets generated by U users. Each tweet has been cleaned and pre-processed (lemmatisation, lowerization, and stemming) using a dictionary of words I, and it is represented as $\mathbf{x}_n = (x_{n1}, \dots, x_{nW_n})$, i.e. as a vector of W_n words. Each word $x_{nj} \in \{1, \dots, |I|\}$ is described by its position in the dictionary.
Given the dataset $\mathbf{X}$, we want to cluster tweets into groups with similar content. For this purpose, we introduce a mixture model for categorical data with the following likelihood:

$$p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}_n|\boldsymbol{\theta}_k) \quad \text{where} \quad p(\mathbf{x}_n|\boldsymbol{\theta}_k) = \prod_{j=1}^{W_n} Cat(x_{nj}|\boldsymbol{\theta}_k)$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ are the mixing proportions and satisfy the constraints $\pi_k \geq 0, \forall k = 1, \dots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. The parameters $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{k|I|})$ represent the probabilities of the words in the dictionary for a given topic k, thus θ_{km} is the probability of the word at position m in the topic k. Again, $\sum_{m=1}^{|I|} \theta_{km} = 1$.

1. Derive the expression of the complete-data log-likelihood.

The complete-data log-likelihood is given by:

$$\begin{aligned}
\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\theta}) &= \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}|\boldsymbol{\pi}) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} [z_n = k] (\log p(\mathbf{x}_n|\boldsymbol{\theta}_k) + \log \pi_k) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} [z_n = k] \sum_{j=1}^{W_n} \sum_{m=1}^{|I|} [x_{nj} = m] \log(\theta_{km}) + \sum_{n=1}^{N} \sum_{k=1}^{K} [z_n = k] \log \pi_k. \qquad (1)
\end{aligned}$$

2. Compute the closed-form expression for the E-step, i.e. $Q(\Theta, \Theta^{old})$ where $\Theta = (\boldsymbol{\pi}, \{\boldsymbol{\theta}\}_{k=1}^{K})$.

$$
\begin{aligned}
Q(\Theta, \Theta^{old}) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\pi}^{old}, \boldsymbol{\theta}^{old})} \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\theta}) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}([z_n = k]) \sum_{j=1}^{W_n}\sum_{m=1}^{|I|} [x_{nj} = m] \log(\theta_{km}) + \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}([z_n = k]) \log \pi_k, \qquad (2)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbb{E}([z_n = k]) &= p(z_n = k|\mathbf{x}_n, \boldsymbol{\pi}^{old}, \boldsymbol{\theta}^{old}) = \frac{p(\mathbf{x}_n, z_n = k|\boldsymbol{\pi}^{old}, \boldsymbol{\theta}^{old})}{p(\mathbf{x}_n|\boldsymbol{\pi}^{old}, \boldsymbol{\theta}^{old})} \\
&= \frac{\pi_k^{old} \prod_{j=1}^{W_n} Cat(x_{nj}|\boldsymbol{\theta}_k^{old})}{\sum_{k'} \pi_k'^{old} \prod_{j=1}^{W_n} Cat(x_{nj}|\boldsymbol{\theta}_k'^{old})} := r_{nk}. \qquad (3)
\end{aligned}
$$

By substituting equation (3) into equation (2) we get:

$$
Q(\Theta, \Theta^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \sum_{j=1}^{W_n}\sum_{m=1}^{|I|} [x_{nj} = m] \log(\theta_{km}) + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \log \pi_k, \qquad (4)
$$

which is the closed-form expression for the E-step.

3. Compute the closed-form equations for the M-step, i.e. the expressions of the MLE for the model parameters $\Theta = (\boldsymbol{\pi}, \{\boldsymbol{\theta}\}_{k=1}^{K})$.

We need to compute the derivatives of $Q(\Theta, \Theta^{old})$ w.r.t. π_k and θ_{km}, taking into account the constraints $\sum_k \pi_k = 1$ and $\sum_m \theta_{km} = 1$, for which we use Lagrange multipliers. Let's start with the derivative w.r.t π_k:

$$
\frac{\partial}{\partial \pi_k} Q(\Theta, \Theta^{old}) + \lambda\big(\sum_k \pi_k - 1\big) = \sum_{n=1}^{N} \frac{r_{nk}}{\pi_k} + \lambda = 0 \iff \pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N} r_{nk}. \qquad (5)
$$

In order to get λ we substitute equation (5) into the restriction over $\boldsymbol{\pi}$:

$$
\sum_k \pi_k = -\frac{1}{\lambda}\sum_k\sum_n r_{nk} = 1 \iff \lambda = -\sum_k\sum_n r_{nk} = -N, \qquad (6)
$$

and substituting back to equation (5), we get the MLE of π_k:

$$
\pi_k^{ML} = \frac{1}{N}\sum_{n=1}^{N} r_{nk}. \qquad (7)
$$

With the same procedure, we get the MLE of θ_{km}.

$$
\begin{aligned}
\frac{\partial}{\partial \theta_{km}} Q(\Theta, \Theta^{old}) + \lambda\big(\sum_m \theta_{km} - 1\big) &= \sum_{n=1}^{N}\sum_{j=1}^{W_n} \frac{[x_{nj} = m] r_{nk}}{\theta_{km}} + \lambda = 0 \\
\iff \theta_{km} &= -\frac{1}{\lambda}\sum_{n=1}^{N}\sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}. \qquad (8)
\end{aligned}
$$

In order to get λ we substitute equation (8) into the restriction over $\boldsymbol{\theta}$:

$$\sum_m \theta_{km} = -\frac{1}{\lambda} \sum_n r_{nk} \sum_j \sum_m [x_{nj} = m] = 1 \iff \lambda = -\sum_n r_{nk} \sum_j \sum_m [x_{nj} = m], \tag{9}$$

and substituting back to equation (8), we get the MLE of θ_{km}:

$$\theta_{km}^{ML} = \frac{\sum_{n=1}^{N} \sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}}{\sum_{m=1}^{|I|} \sum_{n=1}^{N} \sum_{j=1}^{W_n} [x_{nj} = m] r_{nk}}. \tag{10}$$

4. Open the jupyter notebook, and play around with the dataset.
5. Implement the EM algorithm.
6. Show the (approximated) log-likelihood, the ten most representative words for each topic using a wordcloud, and the ten most relevant documents for each topic.

Solution in the files *L2_ solution.ipynb* and *categorical_ em_ solution.py*.

2 Q&A

Question 1

Why Q depends on both Θ and Θ^{old}?

Answer 1

The dependence on Θ comes from the $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$, while the dependence on Θ^{old} from the posterior on $\mathbf{Z}$ just calculated at the E-step (see Eq.(10) notes L2). Hence, the dependence on Θ^{old} is *implicit*, via the posterior. You may not see the explicit Θ in the formula of Q, but they are implicitly contained in the parameters of the posterior of $\mathbf{Z}$. For instance, in the GMM, when updating $\boldsymbol{\mu}_k$ you use $\boldsymbol{\gamma}_k$. The latter only implicitly contains the dependence on $\boldsymbol{\mu}_k$ in the previous time step. At the same time, Q contains *explicitly* a dependence on Θ coming from the log-likelihood contribution. Again, in the GMM, to get to the update of $\boldsymbol{\mu}_k$ you started from an expression which contained both $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ (explicitly).

Question 2

Why do we need to separate the E-step from the M-step? Would be possible to implement these steps together, e.g. by substituting the formula of the posterior of $\mathbf{Z}$ in Q and computing directly the derivatives on this?

Answer 2

In principle yes, but this will bring you back to the reason why we introduced the $\mathbf{Z}$ in the first place: calculating $\log p(\mathbf{X}|\boldsymbol{\theta})$ is intractable. If you substitute the expression of the posterior (and see an explicit dependence on $\boldsymbol{\theta}$), you will get an intractable expression, i.e. you won't be able to calculate derivative and extract a closed-form update for the $\boldsymbol{\theta}$. There is likely going to be terms with logarithms of sum of $\boldsymbol{\theta}$ which cannot be easily unpacked. By introducing $\mathbf{Z}$ and fixing its posterior in the M-step expressions simplify significantly. In addition, it may be values in-and-of itself to have an interpretable latent variable, and thus its posterior.