Advanced Probabilistic Machine Learning and Applications

Martina Contisciani and Caterina De Bacco

November 1, 2021

Tutorial 2: Mixture Models and Expectation Maximization

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.
2. Compute the closed-form expression for the E-step, i.e. $Q(\Theta, \Theta^{old})$ where $\Theta = (\boldsymbol{\pi}, \{\boldsymbol{\theta}\}_{k=1}^{K})$.
3. Compute the closed-form equations for the M-step, i.e. the expressions of the MLE for the model parameters $\Theta = (\boldsymbol{\pi}, \{\boldsymbol{\theta}\}_{k=1}^{K})$.
4. Open the jupyter notebook, and play around with the dataset.
5. Implement the EM algorithm.
6. Show the ten most representative words for each topic using a wordcloud, and the ten most relevant documents for each topic.