Advanced Probabilistic Machine Learning and Applications

Abdullahi Adinoyi Ibrahim and Caterina De Bacco

May 7, 2021

1 Tutorial 3: Bayesian Mixture Model (BMM)+ Gibbs sampling

In this tutorial we will continue working with the CMM and Twitter data-set presented in Tutorial 2. We will use different versions of the Gibbs sampling algorithm to find the posterior distribution of the cluster assignments $\{z_n\}_{n=1}^N$ and model parameters $(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K)$.

Introduction

Notation: Through this document we will use the following notation:

- K: number of mixture components, i.e., we interpret them as topics/clusters.
- N: number of documents, i.e., tweets.
- I: dictionary
- $|I|$: number of words in I.
- $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$: set of likelihood parameters.
- $\mathbf{x}_n \in R^{W_n}$: n-th document with length (i.e., number of words) W_n.
- $\boldsymbol{X} = \{\mathbf{x}_n\}_{n=1}^N$: set of all documents.
- $\boldsymbol{X}_{-n} = \{\mathbf{x}_i | i \neq n\}_{i=1}^N$: set of all documents except for $\mathbf{x}_n$.
- z_n:component assignment variable of document $\mathbf{x}_n$.
- $\boldsymbol{Z} = \{z_n\}_{n=1}^N$: set of all component assignment variables.
- $\boldsymbol{Z}_{-n} = \{\mathbf{z}_i | i \neq n\}_{i=1}^N$: set of all component assignment variables except for z_n.

Summary of Generative Model: We will work with the following Bayesian Mixture Model

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\Theta}) = p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{\Theta}|\gamma) \prod_{n=1}^{N} [p(z_n|\boldsymbol{\pi}) p(\mathbf{x}_n|z_n, \boldsymbol{\Theta})]$$

The conjugate prior for the categorical distribution is the Dirichlet distribution. Therefore, we define the prior distribution for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k$ for all k as Dirichlet distributions with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively. Notice the prior distributions for each $\boldsymbol{\theta}_k$ share the same set of parameters.

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \qquad p(\boldsymbol{\Theta}|\gamma) = \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\theta}_k|\gamma)$$

$$p(z_n|\boldsymbol{\pi}) = \mathrm{Cat}(z_n|\boldsymbol{\pi}) \qquad p(\mathbf{x}_n|z_n, \boldsymbol{\Theta}) = \prod_{j=1}^{W_n} \mathrm{Cat}(x_{nj}|\boldsymbol{\theta}_{z_n})$$

Submission: Copy the Jupyter notebook for Tutorial 3 available in the course webpage `https://github.com/APMLA-2021/APMLA-2021_material/tree/main/L3` and complete the exercises proposed below.

Exercise 1: Derive the Gibbs sampling Algorithms for the CMM

Given the dataset and the probabilistic model described in the previous section, complete the following tasks:

1. **Algorithm 1:** Use the Gibbs sampling algorithm to approximate (using samples) the posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X})$. Derive the conditional distributions you will need to sample in steps (1), (2) and (3) of the following Gibbs sampler:

   Algorithm 1: Gibbs sampling algorithm

   Initialize cluster assignments $\boldsymbol{Z}$ and the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$;
   while *not converged* **do**
     Sample $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}) = p(\boldsymbol{\pi}|\boldsymbol{Z})$; (1)
     for $k = 1, \ldots, K$ **do**
       Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}) = p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z})$; (2)
     end
     for $n = 1, \ldots, N$ **do**
       Sample $z_n \sim p(z_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}, \boldsymbol{\pi}, \boldsymbol{\Theta}) = p(z_n|\mathrm{x}_n, \boldsymbol{\pi}, \boldsymbol{\Theta})$; (3)
     end
   end

2. **Algorithm 2:** Use the (collapsed) Gibbs sampling algorithm to approximate (using samples) the posterior distribution $p(\boldsymbol{Z}, \boldsymbol{\Theta}|\boldsymbol{X})$. Derive the conditional distributions you will need to sample in steps (1) and (2) of the following Gibbs sampler:

   Algorithm 2: $\boldsymbol{\pi}$ collapsed Gibbs sampling algorithm

   Initialize cluster assignments $\boldsymbol{Z}$ and the model parameters $\boldsymbol{\Theta}$;
   while *not converged* **do**
     for $k = 1, \ldots, K$ **do**
       Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z})$; (1)
     end
     for $n = 1, \ldots, N$ **do**
       Sample $z_n \sim p(z_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) = p(z_n|\mathrm{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta})$; (2)
     end
   end

3. **Algorithm 3:** Use the (collapsed) Gibbs sampling algorithm to approximate (using samples) the posterior distribution $p(\boldsymbol{Z}|\boldsymbol{X})$. Derive the conditional distributions you will need to sample in steps (1) the following Gibbs sampler:

   Algorithm 3: $\boldsymbol{\pi}, \boldsymbol{\Theta}$ collapsed Gibbs sampling algorithm

   Initialize cluster assignments $\boldsymbol{Z}$;
   while *not converged* **do**
     for $n = 1, \ldots, N$ **do**
       Sample $z_n \sim p(z_n|\boldsymbol{X}, \boldsymbol{Z}_{-n})$; (1)
     end
   end

Exercise 2: Gibbs Sampling Algorithms implementation & Comparison

For this task, fix the number of clusters to be 5, i.e. $K = 5$. Also, let us consider the log-likelihood as the measure of convergence. Then, complete the following exercises.

1. Implement the log-likelihood, i.e., $\log p(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{Z})$.
2. Implement the functions needed for Algorithm 1, i.e., the three posterior distributions in 1.
3. Implement Algorithm 1, i.e., the fit_no_collapsed method. Then, train the algorithm for 80 iterations with a burn in period $\tau_{\text{burn-in}} = 20$ iterations

- Show the evolution of the log-likelihood per iteration . Do you think the Gibbs sampler has converged (i.e the samples are from the target posterior distribution)?
- Retrieve a sample from the hidden variables at the end of the training.
- Obtain the MAP estimate of the cluster assignments computed after $\tau_{\text{burn-in}}$.
- Show the 10 most representative words for each topic using a cloud of words (Optional)

4. Using your implementation of Algorithm 1 and the implementations of Algorithm 2 & 3 provided in the jupyer notebook, explain the differences in the convergence speed of the algorithms in terms of number of iterations and time. What is the reason behind the differences?