Estimation Statistique Avancée - TP3 LDA

Centrale Lille Institut - G3 SDI

Latent Dirichlet allocation (LDA) is a probabilistic model proposed in Blei et al. (2003) to describe collections of discrete data, in particular text data (but has also seen other application fields, e.g. with genome data). The goal is to discover latent topics that run through a collection of documents.

Let K be the number of latent topics. We assume that we have a corpus of D documents. Each document d is a sequence of L_d words, and each word w is an item of a finite vocabulary of size $V : \{1, ..., V\}$.

We suppose that each document d is represented by a distribution over topics $\boldsymbol{\theta}_d \in [0,1]^K$ with $\sum_k \theta_{dk} = 1$, and that each topic k is represented by a distribution over words $\boldsymbol{\beta}_k \in [0,1]^V$ with $\sum_v \beta_{kv} = 1$. LDA assumes the following generative process.

- For each document d (from 1 to D) :
  - For each word n (from 1 to L_d) :
    - Draw a topic : $\mathbf{z}_{dn} \sim \mathrm{Cat}(\boldsymbol{\theta}_d)$
      $\mathbf{z}_{dn}$ is a indicator vector of size K
    - Draw a word in the topic : $\mathbf{w}_{dn} \sim \mathrm{Cat}(\boldsymbol{\beta}_{\mathbf{z}_{dn}})$
      $\mathbf{w}_{dn}$ is a indicator vector of size V

The lengths L_d are given by the observations and are not modeled. We also assume the following priors :

- $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = \alpha \mathbf{1}_K \in \mathbb{R}^K$.
- $\boldsymbol{\beta}_k \sim \mathrm{Dir}(\boldsymbol{\eta})$, $\boldsymbol{\eta} = \eta \mathbf{1}_V \in \mathbb{R}^V$.

"Dir" denotes the Dirichlet distribution. We assume that α and η are known and fixed.

The posterior $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} | \mathcal{D})$ is intractable, and we resort to variational inference. Note that due to conditional conjugacy, a Gibbs sampler would be feasible as well.