Estimation avancée

G3 SDIA - Centrale Lille Institut
Année universitaire 2023/2024

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 : $\{1, ..., V\}$.

LDA assumes the following generative process :

- Draw K distributions over words $\boldsymbol{\beta}_k \sim \mathrm{Dir}(\boldsymbol{\eta})$
  This represents the distributions over words in each of the K topics
  $\boldsymbol{\beta}_k \in [0,1]^V$ with $\sum_v \beta_{kv} = 1,\ \boldsymbol{\eta} \in \mathbb{R}^V$
- For each document d (from 1 to D) :
  - Draw a distribution over topics $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$
    $\boldsymbol{\theta}_d \in [0,1]^K$ with $\sum_k \theta_{dk} = 1,\ \boldsymbol{\alpha} \in \mathbb{R}^K$
  - 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

We are going to assume that the hyperparameters $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ are fixed. Moreover we assume $\boldsymbol{\eta} = \eta \mathbf{1}_V$ and $\boldsymbol{\alpha} = \alpha \mathbf{1}_K$ (symmetric Dirichlet distributions). The lengths L_d are given by the observations.

The posterior $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} | \mathcal{D})$ is intractable. We are going to resort to variational inference. This model is conditionally conjugate (therefore, a Gibbs sampler would be feasible as well).