CS 6501 Natural Language Processing

Latent Variable Models

Yangfeng Ji

November 5, 2018

Department of Computer Science
University of Virginia

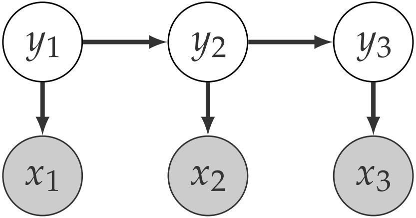
Overview

1. Latent Variable Models
2. Variational Inference
3. Example: Latent Dirichlet Allocation

Latent Variable Models

Latent Variables Models

Hidden Markov Models



Gaussian Mixture Models

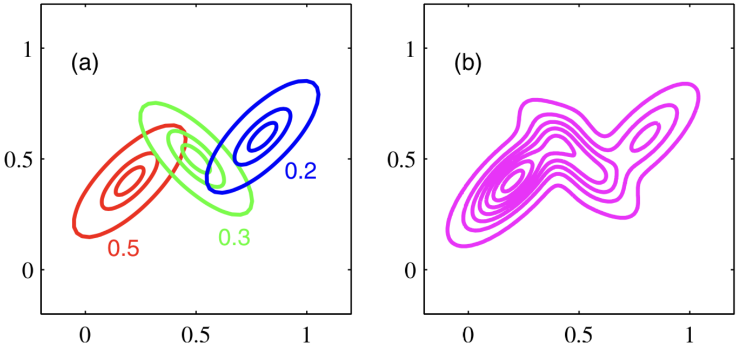
a Gaussian mixture model with K components and each component is a Gaussian distribution

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (1)$$

Parameters

- $\boldsymbol{\mu}_k$: mean of the k-th component
- $\boldsymbol{\Sigma}_k$: variance of the k-th component
- π_k: weight of the k-th component with $\sum_k \pi_k = 1$

Gaussian Mixture Models: Example



[Bishop, 2006]

GMM as a Latent Variable Model

Define a K-dimensional binary random vector $\boldsymbol{z}$ to indicate which mixture component a data point comes from

- only one component of $\boldsymbol{z}$ is 1 and all the rest are 0.
- the probability of z_k is defined as

$$p(z_k = 1) = \pi_k \tag{2}$$

$p(z$

For each z_k

$$p(z_k = 1) = \pi_k \tag{3}$$

Overall,

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{4}$$

is a categorical distribution with parameters $\{\pi_k\}$

$p(\boldsymbol{x} \mid \boldsymbol{z})$

Using $\boldsymbol{z}$ as an indicator vector, we can redefine $p(\boldsymbol{x} \mid \boldsymbol{z})$

- $p(\boldsymbol{x} \mid z_k = 1)$

$$p(\boldsymbol{x} \mid z_k = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

$p(\boldsymbol{x} \mid \boldsymbol{z})$

Using $\boldsymbol{z}$ as an indicator vector, we can redefine $p(\boldsymbol{x} \mid \boldsymbol{z})$

- $p(\boldsymbol{x} \mid z_k = 1)$

$$p(\boldsymbol{x} \mid z_k = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5}$$

- $p(\boldsymbol{x} \mid \boldsymbol{z})$

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{k=1}^{K} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{6}$$

Joint Probability

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}) &= p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) \\
&= \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}
\end{aligned} \tag{7}
$$

Marginal

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) \\
&= \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned} \tag{8}
$$

Graphical Representation

With N data points from the GMM

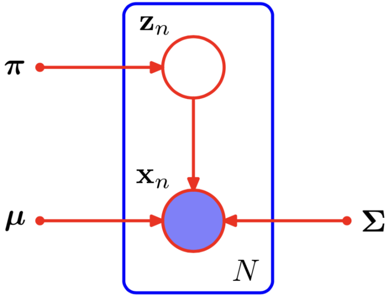
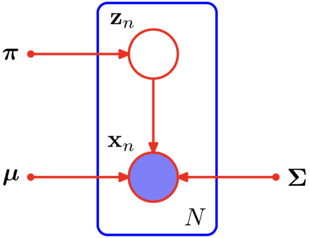


Figure: Graphical representation of GMM [Bishop, 2006].

Generative Story



The generative story of a GMM can be formulated as

1. Randomly pick a mixture component k, with $p(z_k = 1) = \pi_k$
2. Randomly generate a data point from the k component, $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

The procedure can be repeated multiple times

Parameter Estimation

Given N data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, parameter estimation on a GMM is an iteration between the following two steps

1. Estimate $p(\boldsymbol{z}_n)$ for every $\boldsymbol{x}_n$
2. Estimate $\{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ based on $\{p(\boldsymbol{z}_n)\}$ and $\{\boldsymbol{x}_n\}$

Go back to step 1, until convergence

Example

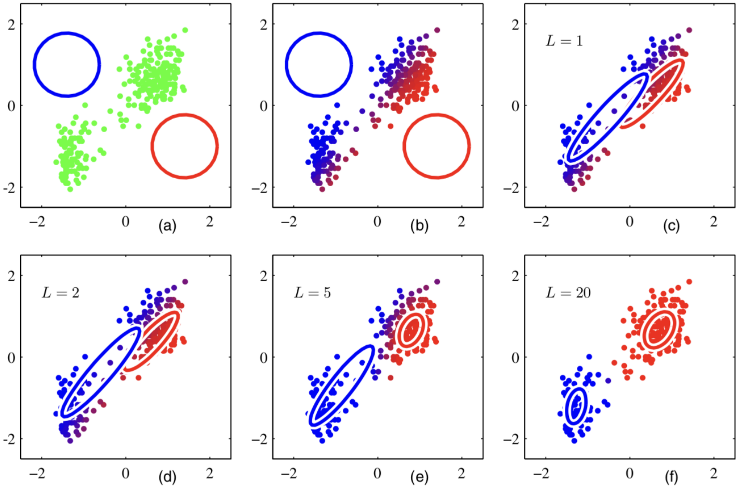
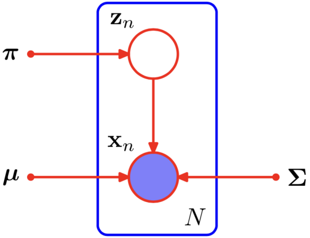


Figure: Illustration of the EM algorithm for GMM parameter estimation.

Comments



A few things about latent variable formulation of GMMs

- z_n is defined on each data point
- the model can be interpreted as a generative story
- marginalizing over z in $p(x, z)$ is tractable

Variational Inference

Ideal Cases

We can solve a latent variable model by compute the following posterior distribution

$$p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})} \tag{9}$$

For example, MLE

$$\arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}) \tag{10}$$

Evidence

However, the challenge comes from

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{x}, z) \tag{11}$$

- Integral may be intractable, when z is continuous
- The space of z could be exponentially large, when z is discrete

BTW, in Bayesian statistics, $p(\boldsymbol{x})$ is called evidence, it measure how good it is for a given model (parameter)

Variational Inference

Instead of computing $p(\boldsymbol{z} \mid \boldsymbol{x})$, we define a family of distribution $\mathbb{Q}$, and compute the following optimization problem

$$\tilde{q}(\boldsymbol{z}) = \arg \min_{q(\boldsymbol{z}) \in \mathbb{Q}} \mathrm{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) \tag{12}$$

where KL divergence is defined as

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{13}$$

More about KL Divergence

The Kullback–Leibler divergence measure the difference between two distributions

$$
\begin{aligned}
\mathrm{KL}(q(x)\|p(x)) &= \sum_q q(x)\log\frac{q(x)}{p(x)} && (14)\\
&= E_q[\log q(x)] - E_q[\log p(x)] && (15)
\end{aligned}
$$

- $\mathrm{KL}(q\|p) = 0$, if $q = p$
- $\mathrm{KL}(q\|p) \geq 0$

ELBO

Does not solve the problem

$$\mathrm{KL}(q \| p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{16}$$

ELBO

Does not solve the problem

$$\mathrm{KL}(q\|p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{16}$$

One more step we need

$$\mathrm{KL}(q\|p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \tag{17}$$

ELBO

Does not solve the problem

$$\mathrm{KL}(q\|p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z} \mid \boldsymbol{x})] \tag{16}$$

One more step we need

$$\mathrm{KL}(q\|p) = E_q[\log q(\boldsymbol{z})] - E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] + \log p(\boldsymbol{x}) \tag{17}$$

Evidence lower bound

$$\mathrm{ELBo} = E_q[\log p(\boldsymbol{z}, \boldsymbol{x})] - E_q[\log q(\boldsymbol{z})] \tag{18}$$

Justification

Put parameters back and consider the log-likelihood

$$\log p(\boldsymbol{x}; \boldsymbol{\theta}) = \log \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{\theta}) \tag{19}$$

Justification

Put parameters back and consider the log-likelihood

$$
\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z};\boldsymbol{\psi})}{q(\boldsymbol{z};\boldsymbol{\psi})}
\end{aligned}
\tag{19}
$$

Justification

Put parameters back and consider the log-likelihood

$$\begin{aligned} \log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\ &= \log \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z};\boldsymbol{\psi})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\ &\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \end{aligned} \tag{19}$$

Justification

Put parameters back and consider the log-likelihood

$$\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{\boldsymbol{z}} \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z};\boldsymbol{\psi})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&= \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \sum_{\boldsymbol{z}} q(\boldsymbol{z};\boldsymbol{\psi}) \log q(\boldsymbol{z};\boldsymbol{\psi})
\end{aligned} \tag{19}$$

Justification

Put parameters back and consider the log-likelihood

$$\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{z} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{z} \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z};\boldsymbol{\psi})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&= \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log q(\boldsymbol{z};\boldsymbol{\psi}) \\
&= E_q[\log p(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\theta})] - E_q[\log q(\boldsymbol{z};\boldsymbol{\psi})]
\end{aligned} \tag{19}$$

Justification

Put parameters back and consider the log-likelihood

$$
\begin{aligned}
\log p(\boldsymbol{x};\boldsymbol{\theta}) &= \log \sum_{z} p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) \\
&= \log \sum_{z} \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})q(\boldsymbol{z};\boldsymbol{\psi})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&\geq \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log \frac{p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z};\boldsymbol{\psi})} \\
&= \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log p(\boldsymbol{x},\boldsymbol{z};\boldsymbol{\theta}) - \sum_{z} q(\boldsymbol{z};\boldsymbol{\psi}) \log q(\boldsymbol{z};\boldsymbol{\psi}) \\
&= E_q[\log p(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\theta})] - E_q[\log q(\boldsymbol{z};\boldsymbol{\psi})] \\
&= E_q[\log p(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\theta})] + H(q) \qquad (19)
\end{aligned}
$$

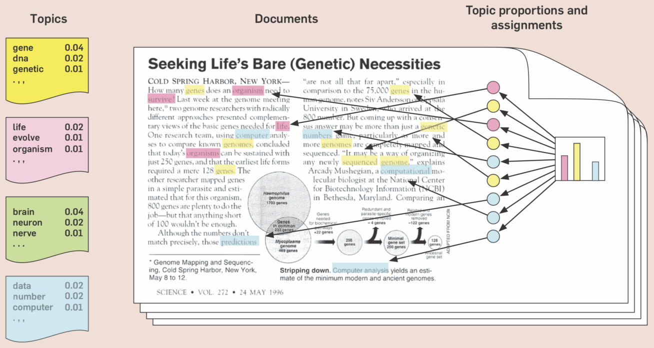
Mean-field Approximation

One special case of variational inference is called **mean field approximation**, which specifies a family of variational distribution, in which different latent variables z_i are independent with each other.

$$p(z; \psi) = \prod_i p(z_i; \psi_i) \tag{20}$$

Example: Latent Dirichlet Allocation

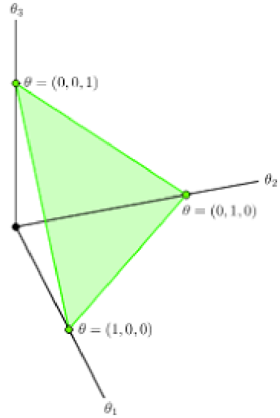
Illustration



The basic idea is that a document is represented as a random *mixture* over latent topics, where each topic is characterized by a distribution over words. [Blei, 2012]

Dirichlet Distribution

$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \tag{21}$$



Generative Story

1. Choose $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. For each word $\boldsymbol{w}_n$
   2.1 Choose a topic $\boldsymbol{z}_n \sim \text{Categorical}(\boldsymbol{\theta})$
   2.2 Choose a word $\boldsymbol{w}_n \sim p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic $\boldsymbol{z}_n$.

where

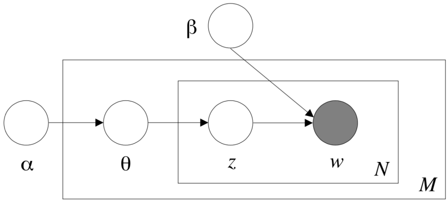
- $\boldsymbol{\theta} \in \mathbb{R}^K$ is a K-dimensional random vector from the Dirichlet distribution with parameter $\boldsymbol{\alpha}$
- $\boldsymbol{\beta} \in \mathbb{R}^{K \times V}$ is a matrix with $\beta_{ij} = p(w_j = 1 \mid z_i = 1)$

Joint Probability

For one document

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}; \boldsymbol{\alpha}) \prod_{n=1}^{N} \left\{ p(z_n; \boldsymbol{\theta}) p(w_n \mid z_n; \boldsymbol{\beta}) \right\} \quad (22)$$

M documents in a corpus



Inference

The key inference problem is to compute the posterior distribution of the hidden variable given a document

$$p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(\boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{\beta})} \tag{23}$$

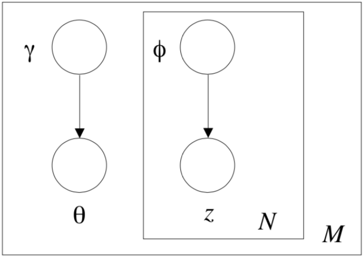
Recall that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the original model. $\boldsymbol{\theta}$ and $\boldsymbol{z}$ are latent variables.

Variational Distribution

For one document

$$q(\theta, z; \gamma, \phi) = q(\theta; \gamma) \prod_n q(z_n; \phi) \tag{24}$$

M documents in a corpus

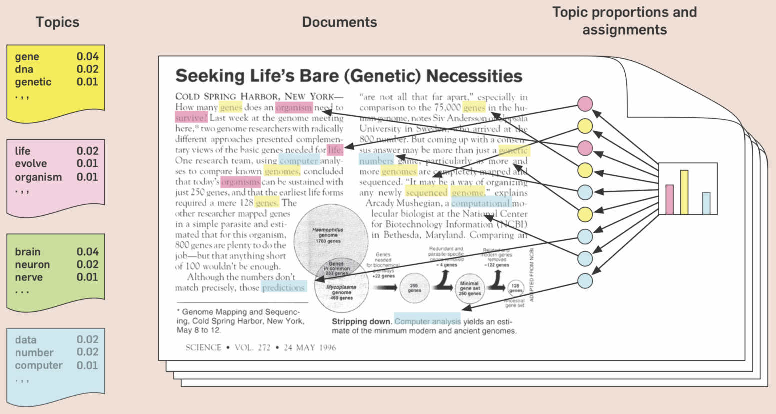


ELBo

$$
\begin{aligned}
\mathrm{ELBo}_{\mathrm{LDA}} =& E_q[\log p(\boldsymbol{\theta}; \boldsymbol{\alpha}] + E_q[\log p(\boldsymbol{z}; \boldsymbol{\theta}) \\
&+ E_q[\log p(\boldsymbol{w} \mid \boldsymbol{z}; \boldsymbol{\beta})] \\
&- E_q[\log q(\boldsymbol{\theta}; \boldsymbol{\gamma})] - E_q[\log q(\boldsymbol{z}; \boldsymbol{\phi})]
\end{aligned}
\tag{25}
$$

As shown in [Blei et al., 2003], every item in Eq. 25 has an analytic form, therefore we can have a closed form solution.

θ and β



[Blei, 2012]

Summary

1. Latent Variable Models
2. Variational Inference
3. Example: Latent Dirichlet Allocation

Reference

Bishop, C. M. (2006).
Pattern Recognition and Machine Learning.
Springer-Verlag.

Blei, D. M. (2012).
Probabilistic topic models.
Communications of the ACM, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022.