PGM tutorial: Variational Inference

Olga Mikheeva, olgamik@kth.se

Introduction

This tutorial focuses on basic principles of Variational Inference. The detailed example is given on Bayesian Multivariate Gaussian Mixture Model. There are 3 theoretical assignments in the example, they should be solved in order to implement missing parts in the code (assignment 4).

1 Theory

In Bayesian statistics all inference problems about unknown quantities are framed as a calculation involving the posterior density. Variation Inference (VI) is a method to approximate probability through optimization. The main idea of VI is to select a family of distributions for the approximation and then find the member of that family that is close to the target distribution. The closeness is measured by Kullback-Leibler divergence.

1.1 Problem set up

Consider a model with observed variables $\boldsymbol{x} = x_{1:n}$ and latent variables $\boldsymbol{z} = z_{1:m}$. The joint density is

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}).$$

In Bayesian models the inference problem is to compute the posterior over latent variables conditioned on the observed ones. This posterior can be written as follows

$$p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$$

The denominator is the marginal density of the observations (the "evidence"). Computing the evidence requires marginalizing out latent variables from the joint density

$$p(\boldsymbol{x}) = \int p(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{z}.$$

In complex Bayesian models this integral is either unavailable in closed form or requires exponential time to compute. Such models require approximate inference.

1.2 Evidence lower bound (ELBO)

The idea of variational inference is to specify a family $\mathcal{Q}$ of densities over latent variables, and then find the best candidate in terms of KL-divergence to the exact posterior. Inference is therefore framed as the following optimization problem:

$$q^*(\boldsymbol{z}) = argmin_{q(\boldsymbol{z})\in\mathcal{Q}} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

The KL-divergence is still not computable since it includes posterior which requires computation of the evidence (which was the problem in the first place):

$$\begin{aligned} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}|\boldsymbol{x})] \\ &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned} \tag{1}$$

However, the log evidence does not depend on the variational distribution $q(\boldsymbol{z})$ we are trying to find. Therefore we can optimize an alternative objective function which is equivalent to the divergence up to a constant:

$$\begin{aligned} ELBO(q) &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] + E_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - KL(q(\boldsymbol{z})||p(\boldsymbol{z}) \end{aligned} \tag{2}$$

1.3 Mean-field approximation

The mean-field variational family of distributions is one of the most commonly used. A generic member of the mean-field family is

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_i(z_i) \tag{3}$$

where latent variables are mutually independent and each z_j is governed by it's own variational factor $q_j(z_j)$. Each variational factor can take any parametric form appropriate to the corresponding random variable.

1.4 Coordinate ascent mean-field variational inference

Using the ELBO and the mean-field variational family, the problem of approximate inference is cast as an optimization. The most commonly used algorithm for this problem is coordinate ascend variational inference (CAVI), which iteratively optimizes each factor of the variational density. The algorithm converges to a local optimum.

The optimal $q_j^*(z_j)$ given that all other factors are fixed is

$$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})]\} \tag{4}$$

2 Example: Bayesian Multivariate Gaussian Mixture Model

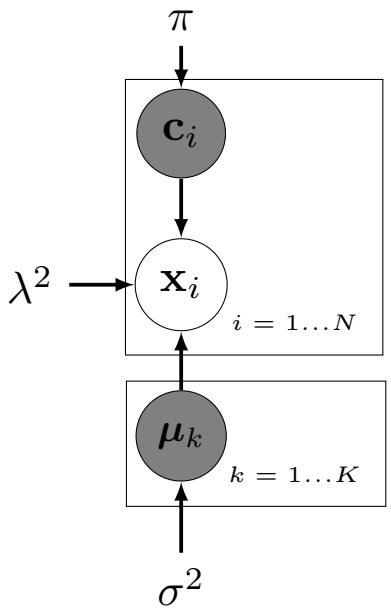

Figure 1: PGM of Bayesian Multivariate Gaussian Mixture Model with K components and N samples.

$$
\begin{aligned}
&\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}) && k = 1, ..., K \\
&c_i \sim Categorical\Big(\frac{1}{K}, ..., \frac{1}{K}\Big) && i = 1, ..., N \\
&\boldsymbol{x}_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(c_i^T \boldsymbol{\mu}, \lambda^2 \boldsymbol{I}) && i = 1, ..., N
\end{aligned} \tag{5}
$$

c_i is an indicator vector. Data are p-dimensional

$$\boldsymbol{\mu}_k \in \mathbb{R}^p, \boldsymbol{x}_i \in \mathbb{R}^p$$

The joint density of observed and latent variables for a sample of size N is

$$p(\boldsymbol{\mu}, \boldsymbol{c}, \boldsymbol{x}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{i=1}^{N} p(c_i) p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu}) \tag{6}$$

Evidence is

$$\begin{aligned} p(\boldsymbol{x}) &= \int \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{i=1}^{N} \sum_{c_i} p(c_i) p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \sum_{\boldsymbol{c}} p(\boldsymbol{c}) \int \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{i=1}^{N} p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu}) d\boldsymbol{\mu} \end{aligned} \tag{7}$$

Each integral is tractable (Gaussian conjugacy), but there are K^N of them (each possible configuration of cluster assignments). Therefore computing the evidence is intractable.

To approximate the posterior over latent variables $\boldsymbol{\mu}, \boldsymbol{c}$ we can use mean-field approximation of the form

$$p(\boldsymbol{\mu}, \boldsymbol{c}) \approx q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(c_i) \tag{8}$$

where posterior over each component's mean parameter is a Gaussian

$$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_k, s_k^2 \boldsymbol{I})$$

and posterior over mixture assignment is a categorical

$$q(c_i) \sim Categorical(\boldsymbol{\phi}_i).$$

$\boldsymbol{m}_k, s_k^2$ and $\boldsymbol{\phi}_i = (\phi_{i,1}, ..., \phi_{i,K})$ are variational parameters that will be optimized.

2.1 ELBO

ELBO is a function of variational parameters

$$\begin{aligned} \mathcal{L}(\boldsymbol{x} | \boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) &= E_q[\log p(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{c})] - E_q[\log q(\boldsymbol{\mu}, \boldsymbol{c})] \\ &= E_q\Big[\sum_{k=1}^{K} \log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} \big(\log p(c_i) + \log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})\big)\Big] \\ &- E_q\Big[\sum_{k=1}^{K} \log q(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} \log q(c_i)\Big] \\ &= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})\Big] \\ &- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big] \end{aligned} \tag{9}$$

Each expectation can be computed in closed form.

Assignment 1
Compute ELBO in closed form (by plugging in all distributions and taking expectations under the approximate distribution). This result will be later used to track the convergence of the model and compare different runs of optimization.

2.2 The variational density for the mixture assignments

Variational update for cluster assignments c_i is derived using equation 4

$$q^*(c_i) \propto \exp\{E_{q(\boldsymbol{c}_{-i},\boldsymbol{\mu})}[\log p(\boldsymbol{x}, c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu})]\} \tag{10}$$

Log joint distribution is

$$\begin{aligned}\log p(\boldsymbol{x}, c_i, \boldsymbol{c}_{-i}, \boldsymbol{\mu}) = \log p(\boldsymbol{\mu}) + \sum_{j\neq i} \big(\log p(c_j) + \log p(\boldsymbol{x}_j|c_j, \boldsymbol{\mu})\big) \\ + \log p(c_i) + \log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\end{aligned} \tag{11}$$

Terms that are not functions of c_i are constants, therefore

$$q^*(c_i) \propto \exp\{\log p(c_i) + E[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})]\} \tag{12}$$

Assignment 2
Show that the variational update for i-th cluster assignment is

$$\phi_{i,k} \propto \exp\left\{\frac{\boldsymbol{x}_i^T E[\boldsymbol{\mu}_k]}{\lambda^2} - \frac{E[\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k]}{2\lambda^2}\right\} \tag{13}$$

Hint. Use the fact that

$$p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\boldsymbol{x}_i|\boldsymbol{\mu}_k)^{c_{i,k}}$$

Notice, that update for each assignment depends only only on the variational parameters of the mixture components, and does not depend on any other cluster assignments.

2.3 The variational density for the mixture-component means

Similarly to the derivation of $q(c_i)$

$$q^*(\boldsymbol{\mu}_k) \propto \exp\{\log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} E_{q(\boldsymbol{\mu}_{-k}, c_i)}[\log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})]\} \tag{14}$$

$q^*(\boldsymbol{\mu}_k)$ is proportional to the product of Gaussians, therefore the resulting distribution is also a Gaussian. Let us look at the expression in the exponent:

$$\begin{aligned}
&\log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} E_{q(\boldsymbol{\mu}_{-k}, c_i)}[\log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})] \\
&= \log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} E_{q(\boldsymbol{\mu}_{-k}, c_i)}\Big[\sum_{j=1}^{K} c_{i,j} \log p(\boldsymbol{x}_i | \boldsymbol{\mu}_j)\Big] \\
&= \{\text{components that do not depend on } \boldsymbol{\mu}_k \text{ are constants}\} \\
&= \log p(\boldsymbol{\mu}_k) + \sum_{i=1}^{N} E_{q(\boldsymbol{\mu}_{-k}, c_i)}\Big[c_{i,k} \log p(\boldsymbol{x}_i | \boldsymbol{\mu}_k)\Big] + const \\
&= -\frac{\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k}{2\sigma^2} + \sum_{i=1}^{N} E_{q(c_i)}\Big[c_{i,k}\Big] \log p(\boldsymbol{x}_i | \boldsymbol{\mu}_k) + const \\
&= -\frac{\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k}{2\sigma^2} + \sum_{i=1}^{N} \phi_{i,k} \log p(\boldsymbol{x}_i | \boldsymbol{\mu}_k) + const \\
&= -\frac{\boldsymbol{\mu}_k^T \boldsymbol{\mu}_k}{2\sigma^2} + \sum_{i=1}^{N} \phi_{i,k}\Big(-\frac{1}{2\lambda^2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\Big) + const
\end{aligned} \tag{15}$$

Assignment 3

Complete the square to find the parameters of the optimal Gaussian $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{m}_k, s_k^2 \boldsymbol{I})$. Those parameters will be used for variational updates of the posterior of the mixture component means.

3 Coordinate ascent variational inference

The CAVI algorithm for Bayesian mixture of Gaussians is as follows:

Assignment 4

Using results of the previous assignments implement missing parts of the algorithm in the provided python code.

Algorithm 1 CAVI for a Gaussian Mixture Model

Input:
Output:
Initialize:
while the ELBO has not converged **do**
    for $i \in \{1, ..., N\}$ **do**
        Update ϕ_i (result from assignment 2)
    for $k \in \{1, ..., K\}$ **do**
        Update $\boldsymbol{m}_k$ (result from assignment 3)
        Update s_k^2 (result from assignment 3)
    Compute $ELBO(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi})$ (result from assignment 1)
return $q(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi})$

References:

1. Jordan, Michael I., et al. "An introduction to variational methods for graphical models." Machine learning 37.2 (1999): 183-233.

2. Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American Statistical Association just-accepted (2017).

3. D. Koller and N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

4. Christopher M. Bishop,"Pattern Recognition and Machine Learning","Springer, 2006.

5. Wainwright, Martin J., and Michael I. Jordan. "Graphical models, exponential families, and variational inference." Foundations and Trends in Machine Learning 1.12 (2008): 1-305.

6. Blei, David M., and Michael I. Jordan. "Variational inference for Dirichlet process mixtures." Bayesian analysis 1.1 (2006): 121-143.

7. Hoffman, Matthew D., et al. "Stochastic variational inference." The Journal of Machine Learning Research 14.1 (2013): 1303-1347.