Variational Inference for Dirichlet Process Mixtures and Beyond

A. Ciceri, M. Caputo, A. Curti, A. Dakouri, E. Di Liberatore, G. Frigerio
Tutored by Mario Beraha

23/24 November 2023

Table of contents

Variational Inference

Variational inference - introduction

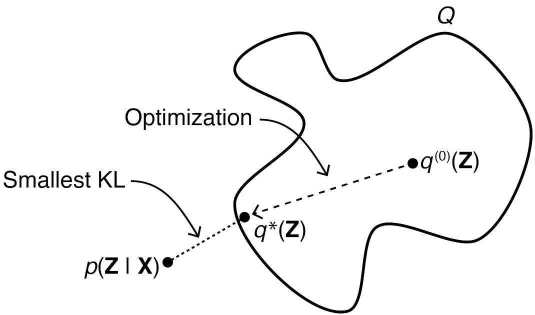

Figure 1: Visualization of VI optimization

Minimization of the Kullback-Leibler (KL) divergence:

$$q^*(z) = \underset{q(z) \in \mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(q(z) \| p(z|x)\right)$$

We choose $\mathcal{Q}$ to be flexible enough to capture a density close to $p(z|x)$, but simple enough for efficient optimization. [2]

Variational Inference - introduction

Since we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant,

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log p(z, x)] - \mathbb{E}_q[\log q(z)]$$

This function is called the evidence lower bound (ELBO). Maximizing the ELBO is equivalent to minimizing the KL divergence.
Finally, we approximate the posterior with the optimized member of the family $q^*(.)$.

Mean-field

We work in a mean-field variational framework, which means that latent variables z are mutually independent and each governed by a distinct factor in the variational density.

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$$

Where m is the dimension of the latent variables.

Comparing variational inference and MCMC

MCMC

- Produce exact solution asymptotically
- Computationally intensive
- Suited for:
  1. small datasets
  2. when more precise samples are needed

VARIATIONAL INFERENCE

- Does not provide an exact solution
- Fast convergence
- Suited for:
  1. large datasets
  2. when we want to quickly explore many models.

Algorithm: Mixture of Gaussians

Bayesian Multivariate Gaussian Mixture Model using VI

Hierarchical model:

$$x_i \mid c_i, \mu \overset{\text{ind}}{\sim} \mathcal{N}_d\left(c_i^T \mu,\ \mathbb{I}_d\right) \qquad i = 1, ..., n$$

$$\mu_k \overset{\text{iid}}{\sim} \mathcal{N}_d\left(0,\ \sigma^2\ \mathbb{I}_d\right) \qquad k = 1, ..., K$$

$$c_i \overset{\text{iid}}{\sim} Cat(\frac{1}{K}, ..., \frac{1}{K}) \qquad i = 1, ..., n$$

i. K is the number of clusters μ
ii. c_i is the cluster assignment, it indicates which latent cluster x_i comes from
iii. n is the size of the sample
iv. d is the data dimension
v. σ^2 is a fixed hyperparameter

Variational densities

We want to approximate the posterior of μ_k, which is still a Gaussian, so we consider the following variational densities q:

$$q(\mu_k) \overset{\text{iid}}{\sim} \mathcal{N}_d\left(m_k,\ s_k^2\ \mathbb{I}_d\right) \qquad\qquad k = 1, ..., K$$

While to approximate the cluster allocation parameters we have:

$$P(c_i = j) = \varphi_{ij} \qquad\qquad k = 1, ..., K; \quad i = 1, ..., n$$

To define our variational densities we have to compute the following parameters:

- m: K vectors of means
- s^2: vector of variances
- φ: NxK matrix of cluster assignments probabilities

Updates

Variational update for the cluster assignment c_i:

$$\varphi_{ik} \propto exp[\ \mathbb{E}[\mu_k;\ m_k,\ s_k^2]x_i - \mathbb{E}[\mu_k^2;\ m_k,\ s_k^2]/2\ ]$$

Updates

Variational update for the cluster assignment c_i:

$$\varphi_{ik} \propto exp[\ \mathbb{E}[\mu_k;\ m_k,\ s_k^2]x_i - \mathbb{E}[\mu_k^2;\ m_k,\ s_k^2]/2\ ]$$

Variational density of the kth mixture component expressed in terms of the variational mean and variance:

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}} \qquad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}$$

Optimization

Since VI can be considered an optimization problem, here we have the function to be optimized. In particular, the goal is to maximize the ELBO.

$$ELBO(m, s^2, \varphi) = \sum_{k=1}^{K} \mathbb{E}[\log p(\mu_k);\; m_k,\; s_k^2] +$$

$$\sum_{i=1}^{n} (\; \mathbb{E}[\log p(c_i);\; \varphi_i] + \mathbb{E}[\log p(x_i | c_i, \mu);\; \varphi_i,\; \hat{m},\; s^2]\;) +$$

$$- \sum_{i=1}^{n} \mathbb{E}[\log q(c_i;\; \varphi_i)] - \sum_{k=1}^{K} \mathbb{E}[\log q(\mu_k;\; m_k,\; s_k^2)]$$

Each expectation can be computed in closed form. [3]

Results

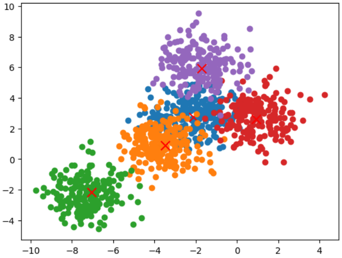

Figure 2: Our generated data

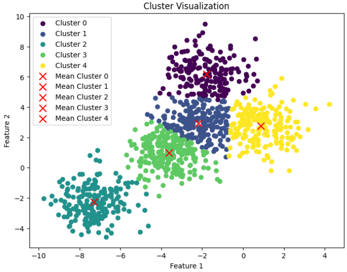

Figure 3: VI clustering

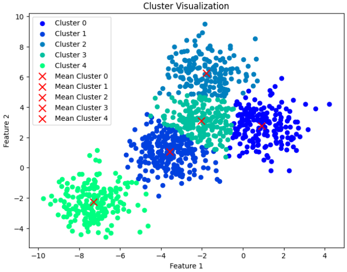

Figure 4: MCMC clustering

Comparison with MCMC (K=5, d=2)

TIME	N = 100	N = 1000	N = 10000
VI	508 ms ± 6.78 ms	2.31 s ± 304 ms	1min 16s ± 778 ms
MCMC	19 s ± 2 s	1min 47 s ± 7 s	21 min ± 1 min

CLUSTER ERROR	N = 100	N = 1000	N = 10000
VI	0.292078797	0.240227441	0.195773545
MCMC	0.326697788	0.246452642	0.197457232

DISTR. ERROR	N = 100	N = 1000	N = 10000
VI	33.17399	34.65825	64.832405
MCMC	0.24917	4.21288	33.59156

What's next?

Our goal

- We will focus on implementing an algorithm for Dirichlet process mixture models using variational inference. The algorithm is described in the paper from Blei and Jordan (2006). [1]
- Moreoever we'll compare the results of the algorithm implemented with VI with the results obtained from MCMC implementation.

To do so we'll move to a nonparametric setting.

Dirichlet process

The Dirichlet process is a stochastic process used in Bayesian nonparametric. It is a distribution over distributions, which means that each draw from a Dirichlet process is itself a distribution.

Distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters, thus the classification as a nonparametric model.

Dirichlet process Mixture Model

Hierarchical model:

$$
\begin{aligned}
G \mid \alpha, G_0 &\sim DP(\alpha,\ G_0) & \\
\eta_n \mid G &\overset{\text{ind}}{\sim} G & n = 1, ..., N \\
X_n \mid \eta_n &\overset{\text{ind}}{\sim} p(x_n,\ \eta_n) & n = 1, ..., N
\end{aligned}
$$

i. α is a positive scaling parameter

ii. G_0 is a non-atomic probability distribution

iii. η_n follows an Pòlya urn distribution

iv. N is the total number of drawn η_n

VI adaptation

We consider this factorized family of variational distributions for meanfield variational inference:

$$q(v, \hat{\eta}, z) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\hat{\eta}_t) \prod_{n=1}^{N} q_{\phi_n}(z_n)$$

Definition of the ELBO, the function to maximize:

$$ELBO(V, \hat{\eta}, Z) = \mathbb{E}_q[\log p(V \mid \alpha)] \; + \mathbb{E}_q[\log p(\hat{\eta} \mid \lambda)] +$$

$$\sum_{n=1}^{N}(\mathbb{E}_q[\log p(Z_n \mid V)] \; + \mathbb{E}_q[\log p(x_n \mid Z_n)]) - \mathbb{E}_q[\log q(V, \; \hat{\eta}, \; Z)]$$

Mean-field coordinate ascent algorithm

We can define the updates to be implemented:

$$\gamma_{t,1} = 1 + \sum_{n=1}^{N} \phi_{n,t}$$

$$\gamma_{t,2} = \alpha + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi_{n,j}\phi_{n,t}$$

$$\tau_{t,1} = \lambda_1 + \sum_{n=1}^{N} \phi_{n,t} x_n$$

$$\tau_{t,2} = \lambda_2 + \sum_{n=1}^{N} \phi_{n,t}$$

$$\phi_{n,t} \propto exp(S_t)$$

Where $t \in \{1, ..., T\}$ and $n \in \{1, ..., N\}$, where

$$S_t = \mathbb{E}_q[\log V_t] + \sum_{i=1}^{t-1} \mathbb{E}_q[\log(1 - V_t)] + \mathbb{E}_q[\hat{\eta}_t]^T X_n - \mathbb{E}_q[a(\hat{\eta})]$$

Iterating these updates optimizes the ELBO with respect to the variational parameters.

References

[1] David M. Blei and Michael I. Jordan. **"Variational inference for Dirichlet process mixtures".** In: *Bayesian Analysis* 1.1 (2006), pp. 121–143. DOI: 10.1214/06-BA104. URL: https://doi.org/10.1214/06-BA104.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. **"Variational Inference: A Review for Statisticians".** In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080%2F01621459.2017.1285773.

[3] Haoliang. **"Variational Inference in Bayesian Multivariate Gaussian Mixture Model".** In: *Bayesian Analysis* (2020). DOI: Hao-2020.