Machine Learning (/tags/#Machine Learning)

# Variational Inference

*Posted by Gwan Siu on March 29, 2018*

## 1. What’s variational inference?

Representation, learning, and inference are 3 cores problem of machine learning. For statistic inference, it involves finding the approxiate model and parameters to represent the distribution given observed variables. In other work, given complete data $x$ and $y$ and unknown parameter $\theta$, this is classical parameter estimation problem in ML area. Usually, we adopt maximum likelihood estimation(MLE):

$$\theta^{*}=\arg\max_{\theta} p(x|y;\theta) \tag{1}$$

in many real situations, we are given imcomplete data, i.e., only data $x$, in this case, latent variables $z$ are introduced. For example, in gassian mixture model, we introduce $z_i$ to indicate the underlying gaussian distribution. Thus, the overall formulation is changed

$$p(x;\theta)=\int_{z} p(x,z;\theta)\mathrm{d}z \tag{2}$$

in fact, this integration usually is high-dimensional integration, which is intractable. It means that extract inference is impossible in this case. Therefore, we need to introduce approximate inference techniques in this case. Samling-based algorithms and variation-based algorithms are two kinds of approximate inference algorithms in modern bayesian statistics.

In this article, we mainly focus on variational inference(VI). **The core idea of VI is to posit a family of distribution and then to find the member of that family which is close to the target, where closeness is measured using the Kullback-Leibler divergence.**

In my previous article-EM (http://gwansiu.com/2018/11/21/Expectation-Maximization/), we can see that data likelihood can be decomposed into evidence lower bound and KL divergence:

$$\mathcal{L}(\theta) = \mathcal{F}(q,\theta) + \mathrm{KL}(q(z), p(z|x;\theta)) \tag{3}$$

where $\mathcal{F}(q,\theta)$ is evidence lower bound for marginal likelihood due to $\mathrm{KL}(q(z), p(z|x;\theta))$ is non-negative.

$$\mathcal{L}(\theta) \geq \mathcal{F}(q,\theta) \tag{4}$$

Instead of maximize marginal likelihood directly, EM algorithm and variational inference maximize the lower bound.

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(z)\ln\frac{p(x,z;\theta)}{q(z)}\mathrm{d}z && (5)\\
&= \mathbb{E}_{q(z)}\big[\ln\frac{p(x,z;\theta)}{q(z)}\big] && (6)\\
&= \mathbb{E}_{q(z)}\big[\ln\frac{p(x|z;\theta)p(z;\theta)}{q(z)}\big] && (7)\\
&= \mathbb{E}_{q(z)}\big[\ln p(x|z)\big] - \mathrm{KL}\big(q(z), p(z;\theta)\big) && (8)
\end{aligned}$$

1. The first term is the expectation of the data likelihood and thus $\mathcal{F}(q,\theta)$ encourage distributions put their mass on configurations of latent variables that explain observed data.
2. The second term is the negative KL divergence between the variational distribution and the prior, so the $\mathcal{F}(q,\theta)$ force $q(z)$ to close to the prior $p(z)$.

**Hence, maximize** $\mathcal{F}(q,\theta)$ ** means to balance the likelihood and prior.**

# 2. Expectation-Maximization

In EM framework, we assume $q(z) = p(z|x;\theta^{old})$. The ELBO becomes:

$$\begin{aligned}\mathcal{F}(q,\theta) &= \int q(z)\ln\frac{p(x,z;\theta)}{q(z)}\mathrm{d}z && (9)\\ &= \int q(z)\ln p(x,z;\theta)\mathrm{d}z - \int q(z)\ln q(z) && (10)\\ &= \int p(z|x;\theta^{old})\ln p(x,z;\theta)\mathrm{d}z - \int p(z|x;\theta^{old})\ln p(z|x;\theta^{old}) && (11)\\ &= Q(\theta,\theta^{old}) - H(q) && (12)\end{aligned}$$

where $H(q)$ is the entropy of $z$ given $x$. It is constant w.r.t $\theta$ and thus we will not take it into account when we maximize ELBO. The EM algorithm is sufficient to maximize $Q(\theta,\theta^{old})$

**E-step:** maximize $\mathcal{F}(q,\theta)$ w.r.t distribution over hidden variables given the parameters:

$$\begin{aligned}q^{(t+1)} &= \arg\max_{q(z)}\mathcal{F}(q(z),\theta^{(t)}) && (13)\\ &\to p(z|x;\theta^{old}) && (14)\end{aligned}$$

**M-step:** maximize $\mathcal{F}(q,\theta)$ w.r.t the parameters given the hidden distribution

$$\int p(z|x;\theta^{old})\ln p(x,z;\theta)\mathrm{d}z \qquad (15)$$

# 3. Mean Field Theory

In EM framework, $q(z)=p(z|x;\theta^{old})$ is computed by iterative method. It means that we can find a analytical solution of $p(x|x;\theta^{old})$, this is possible for simple modles but can not be generalized to complex models. Instead, we approximate the posterior distribuiton by a family of simple dsitributions.

$$q(z)=\prod_{j=1}^{m}q(z_j) \qquad (16)$$

we assume the latent variables are mutually independent and each governed by a distinct factor in the variational distribution， i.e. $z_i\perp z_j$, for $i\neq j$. This is called `mean-field theory` .

# 4. Coodinate Ascent Variational Inference(CAVI)

In this part, I will compbined with mean-field theory and talk about how ELBO is maximize. One latent variabe posterior $q(z_i)$ is updated by the rest latent variables $i \neq j$. Here, I will talk about CAVI algorithm. Let $q(z) = \prod_i q(z_i)$. Then, the EBLO becomes:

$$\begin{aligned}
\mathcal{F}(q,\theta) &= \int q(z) \ln \frac{p(x,z;\theta)}{q(z)} \mathrm{d}z \\
&= \int \prod_i q(z_i) \ln p(x,z;\theta) \mathrm{d}z - \int \prod_i q(z_i) \ln q(z) \mathrm{d}z \\
&= \int q(z_j) \int \prod_{i\neq j} q(z_i) \ln p(x,z;\theta) \prod_{i\neq j} \mathrm{d}z_i \mathrm{d}z_j - \sum_{i\neq j} \int q(z_i) \ln q(z_i) \mathrm{d}z_i - \int q(z \\
&= \int q_j \ln(\frac{\exp(\mathbb{E}[\ln p(x,z;\theta)])}{q(z_j)}) \mathrm{d}z_j - \sum_{i\neq j} \int q(z_i) \ln q(z_i) \mathrm{d}z_i - \int q(z_j) \ln q(z \\
&= \text{\dislaystyle} \int q(z_j \ln \frac{\hat{p}(z_{i\neq j})}{q(z_j)}) \mathrm{d}z_j + H(z_{i\neq j}) + c \\
&= -\mathrm{KL}(q(z_j), \hat{p}(z_{i\neq j})) + H(z_{i\neq j}) + c
\end{aligned}$$

Since KL divergence is non-negative, thus, ELBO is maximized when $\mathrm{KL}(q(z_j), \hat{p}(z_{i\neq j})) = 0$, i.e.

$$q(z_j) = \hat{p}(z_{i\neq j}) = \frac{1}{Z} \exp(\mathbb{E}[\ln p(x,z;\theta)]_{i\neq j}) \tag{23}$$

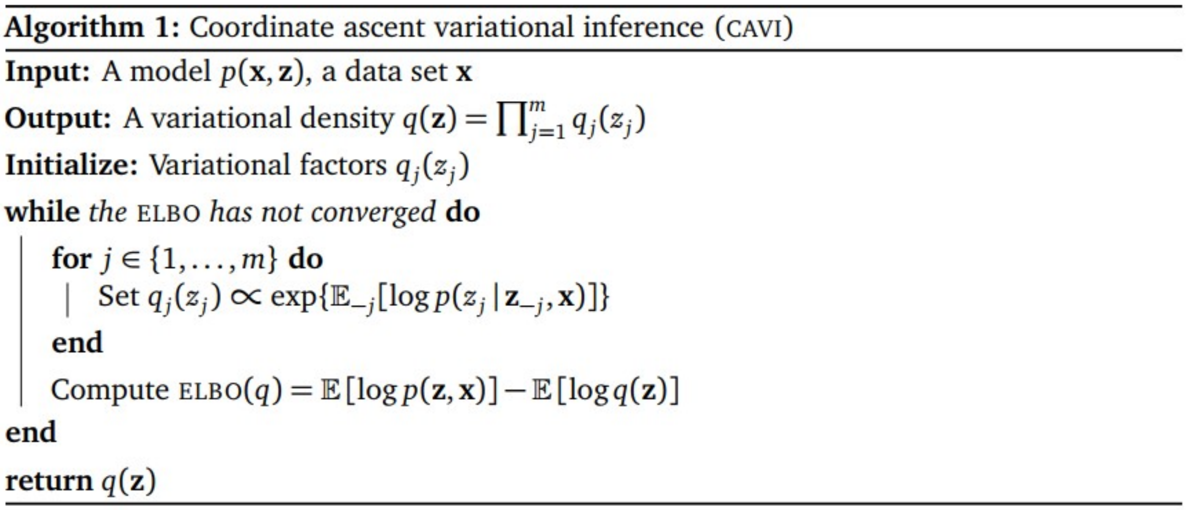
Similarly, in variational EM:

**E-step:** $q^* = \frac{1}{Z} \exp(\mathbb{E}[\ln p(x,z;\theta)]_{i\neq j})$

$$q(z) = \prod_i q(z_i) \tag{24}$$

**M-step:** maximize the $\mathcal{F}(q,\theta)$.

The figure below is the process of CAVI algorithm:



```
Algorithm 1: Coordinate ascent variational inference (CAVI)
Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while the ELBO has not converged do
    for $j \in \{1, \dots, m\}$ do
        Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
    end
    Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$
```

# 4. Variational inference and GMM

In this section, CAVI algorithm is used for Mixture of Gaussians model(GMM). It will be helpful to understand how CAVI works.

## 4.1 Joint distribution computation

Given observed data $X = (x_1, \dots, x_n)$ from $K$ independent gaussian distribution with mean $\mu_k$. One-hot vector $c_i \in \mathbb{R}^k$ indicate the distribution to which each data belong. The hyperparameter $\sigma^2$ is fixed. latent variables are $\mu, c$. The prior is:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) && (25) \\ c_i &\sim Categorical(\frac{1}{K}, \dots, \frac{1}{K}) && (26) \\ x_i &\sim \mathcal{N}(c_i^T, \mu, 1) && (27) \end{aligned}$$

According to bay theroem, we can compute the joint distribution:

$$\begin{aligned} p(\mu, c, x) &= p(\mu)p(c, x|\mu) && (28) \\ &= p(\mu)p(c)p(x|c, \mu) && (29) \\ &= p(\mu)\prod_{i=1}^{n} p(c_i)p(x_i|c_i, \mu) && (30) \end{aligned}$$

once we have joint distribution, we can compute marginal distribution. However, the formulation has no analytical solution, and the computational complexity is $\mathcal{O}(K^n)$.

$$\begin{aligned} p(x) &= \int \sum_{c} p(\mu, c, x) \mathrm{d}\mu && (31) \\ &= \int p(\mu) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i|c_i, \mu) \mathrm{d}\mu && (32) \end{aligned}$$

## 4.2 GMM and CAVI

Now, we should compute variational ditribution $q(z)$, where $m = (m_1, \dots, m_k), s^2 = (s_1^2, \dots, s_K^2), \phi = (\phi_1, \dots, \phi_n)$ are variational parameters, hence the formulation of variational distribution is:

$$\begin{aligned} q(z) &= q(\mu, c) && (33) \\ &= \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \cdot \prod_{i=1}^{n} q(c_i; \phi_i) && (34) \end{aligned}$$

- 1. we can obtain the formulation $\mathrm{ELBO}$, which is a function of $m, s^2, \phi$.

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathrm{ELBO}(m, s^2, \phi) \\ &= \mathbb{E}_{q(z)}[\log p(x|z)] + \mathbb{E}_{q(z)}[\log p(z)] - \mathbb{E}_{q(z)}[\log q(z)] \\ &= \sum_{i=1}^{n} \mathbb{E}[\log p(x_i|c_i, \mu; \phi_i, m, s^2)] + (\sum_{k=1}^{K} \mathbb{E}[\log p(\mu_k; m_k, s_k^2)] + \sum_{i=1}^{n} \mathbb{E}[\log p( \\ &- (\sum_{k=1}^{K} \mathbb{E}[\log q(\mu_k; m_k, s_k^2)] + \sum_{i=1}^{n} \mathbb{E}[\log q(c_i; \phi_i]) \end{aligned}$$

- 1. from section 3, we obtain how CAVI algorithm update latent variables. Now, we applyied it into GMM to compute **cluster indicator** $c$ and update $c$, noted $\mu$ is fixed:

$$\begin{aligned} q_j^* &\propto exp(\mathbb{E}_{\mu}[\log p(c_i, \mu, x_i)]) && (39) \\ &\propto exp(\mathbb{E}[\log p(x_i|c_i, \mu) \cdot \log p(c_i, \mu)]) && (40) \\ &\propto exp(\mathbb{E}_{\mu}[\log p(x_i|c_i, \mu)] + \mathbb{E}_{\mu}[\log p(c_i, \mu)]) && (41) \\ &\propto exp(\mathbb{E}_{\mu}[\log p(c_i|c_i, \mu)] + \log p(c_i)) && (42) \end{aligned}$$

the second term $\log p(c_i)$ is log prior and it is a constant. Hence, we pay our attention to the first term: the distribution of $c_i$ gaussian distribution. In detail, we simplify it due to $c_i = (c_{i1}, \dots, c_{ik})$ is one-hot vector, and we have:

$$\mathbb{E}[\log p(x_i|c_i,\mu)] = \sum_k c_{ik}\mathbb{E}_{\mu_k}[\log p(x_i|\mu_k)] \tag{43}$$

$$= \sum_k c_{ik}\mathbb{E}_{\mu_k}[-\frac{(x-\mu_k)^2}{2}] + const \tag{44}$$

$$= \sum_k c_{ik}(\mathbb{E}_{\mu_k}[\mu_k]x_i + \mathbb{E}_{\mu_k}[\mu_k^2]/2) + const \tag{45}$$

from the formulation above, $\mathbb{E}[\mu_k]$ and $\mathbb{E}[\mu_k^2]$ can be computed. For each data point $i$, parameter $\phi_{ik}$ in the $k$th component of the latent variable $c$. The updated formulation is:

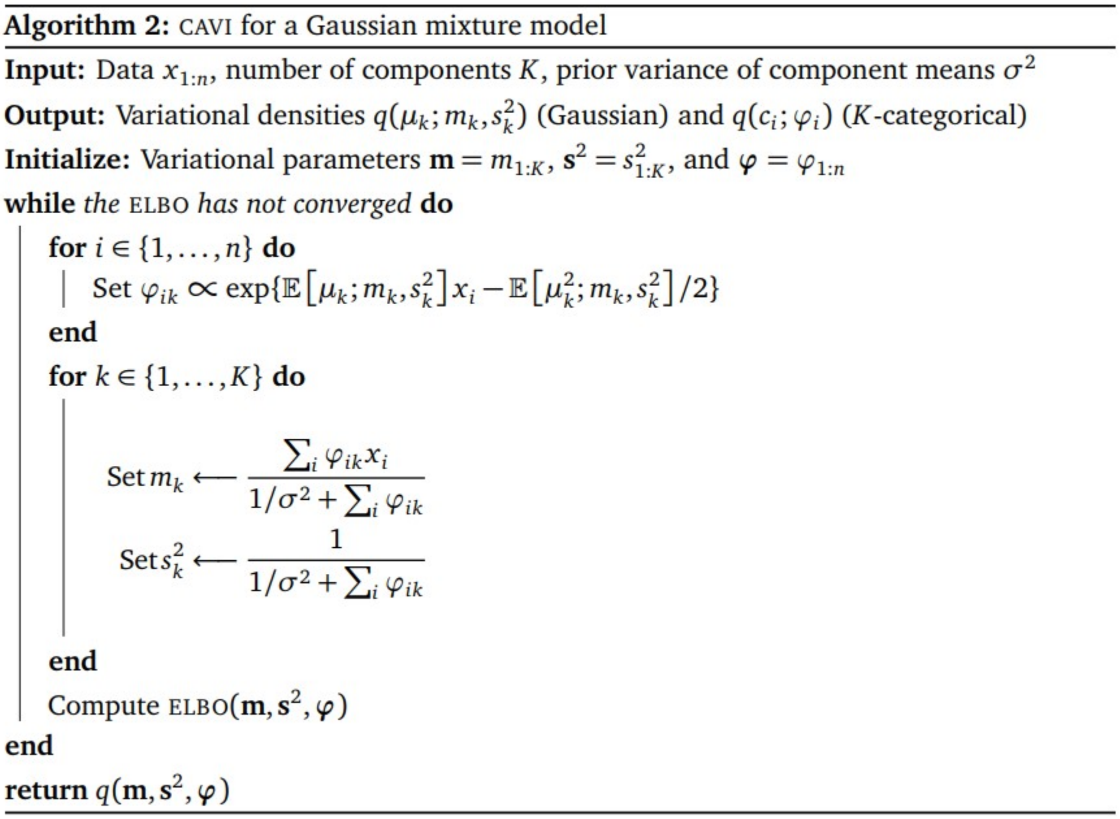
$$\phi_{ik} \propto exp(\mathbb{E}[\mu_k]x_i + \mathbb{E}[\mu_k^2]/2) \tag{46}$$

smiliarly, we can compute latent variable $\mu$ of GMM. Firstly, we should calculate the optimal variational distribution $q(\mu_k)$, and the update the parameter $m_k$, $s_k^2$ of $\mu_k$:

$$m_k = \frac{\sum_i \phi_{ik}\cdot x_i}{1/\sigma^2 + \sum_i \phi_{ik}} \tag{47}$$

$$s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \phi_{ik}} \tag{48}$$

The algorithm of GMM and CAVI is below:



```
Algorithm 2: CAVI for a Gaussian mixture model
Input: Data $x_{1:n}$, number of components $K$, prior variance of component means $\sigma^2$
Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(c_i; \varphi_i)$ ($K$-categorical)
Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s_{1:K}^2$, and $\boldsymbol{\varphi} = \varphi_{1:n}$
while the ELBO has not converged do
    for $i \in \{1, \ldots, n\}$ do
        Set $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$
    end
    for $k \in \{1, \ldots, K\}$ do
        Set $m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$
        Set $s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$
    end
    Compute ELBO($\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}$)
end
return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$
```

# 5. Comparision of MCMC and VI

| MCMC | VI |
|---|---|
| More computationally intensive | Less intensive |
| Gaurantess producing asymptotically exact samples from the target distribution | No such gaurantees |
| Slower | Faster, expecially for large data sets and complex distributions |
| Best for precise inference | Useful to explore many scenarios quickly or large data sets |

Reference

1. AM207-Lecture-Variational-Inference (https://am207.github.io/2017/lectures/lecture24.html)
2. Expectation Maximization and Variational Inference (Part 1) (https://chrischoy.github.io/research/Expectation-Maximization-and-Variational-Inference/)

**PREVIOUS**
SAMPLING METHODS
**(/2018/03/29/SAMPLING/)**

**NEXT**
NOTE(8)-- SUFFICIENCY STATISTICS
**(/2018/04/08/SUFFICIENCY/)**

0条评论 GwanSiu

1 登录

♡ 推荐 推文 f 分享

评分最高

开始讨论…

通过以下方式登录

或注册一个 DISQUS 帐号 ?

姓名

来做第一个留言的人吧!

订阅 在您的网站上使用 Disqus添加 Disqus添加 Disqus 隐私政策隐私政策隐私

FEATURED TAGS (/tags/)

Deep Learning (/tags/#Deep Learning) Image Processing (/tags/#Image Processing)

Machine Learning (/tags/#Machine Learning) Life and Thinking (/tags/#Life and Thinking)

Segmentation and Dectection (/tags/#Segmentation and Dectection) Analysis (/tags/#Analysis)

Statistics and Bayesian Analysis (/tags/#Statistics and Bayesian Analysis)

FRIENDS

K_Augus (https://jiqiujia.github.io/) Liwen (https://liwen.site/)

● (https://www.zhihu.com/people/Gwan-siu)

● (http://weibo.com/u/6056969706)

● (https://github.com/Gwan-Siu)

Theme by Hux (http://huangxuan.me) | Star 4,976