CAVI

lruczu

1 Coordinate ascent variational inference

These notes are fully based on the great research paper: "Variational inference: A Review for Statisticians" (David M. Blei et al.).

In probabilistic computation we often face the situation in which the posterior distribution must be sampled or approximated. This is very often a posteriori distribution of parameters of the model.

$$p(\mathbf{z}|\mathbf{x}) \tag{1}$$

Here, $\mathbf{z}$ is a vector of latent variables, but as in the Blei's paper, $\mathbf{z}$ covers:

- local latent variables, assigned to individual observations
- global latent variables, parameters of the model

There are two ways we can approach this problem

- by sampling
- by optimizing

The first one is simply applying some algorithm from Markov Chain Monte Carlo family, or MCMC for short. Unfortunately, sampling is slow and very rarely applicable to real problems, with big data and with complex model. The second way is to look for a distribution which is the closest to the real one. The distance is measured by Kullback-Leibler divergence. So mathematically, the optimization problem can written as:

$$q^*(\mathbf{z}) = \arg\min_q KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})), \tag{2}$$

where q is searched through some predefined probability distribution set D.

The evidence lower bound, or elbo for short, is very important concept. It allows us to formulate the above problem of finding the closest density in a way that makes the computation more tractable.

First, let's rewrite the Kullback-Leibler divergence.

$$\begin{aligned} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = E_{q(\mathbf{z})}[log(q(\mathbf{z}))] - E_{q(\mathbf{z})}[log(p(\mathbf{z}|\mathbf{x}))] = \\ E_{q(\mathbf{z})}[log(q(\mathbf{z}))] - E_{q(\mathbf{z})}[log(p(\mathbf{z},\mathbf{x}))] + log(p(\mathbf{x})) = \\ E_{q(\mathbf{z})}[log(q(\mathbf{z}))] - E_{q(\mathbf{z})}[log(p(\mathbf{z},\mathbf{x}))] + const = const - ELBO(q) \end{aligned} \tag{3}$$

So, elbo is

$$ELBO(q) = E_{q(\mathbf{z})}[log(p(\mathbf{z}, \mathbf{x}))] - E_{q(\mathbf{z})}[log(q(\mathbf{z}))] \tag{4}$$

when we maximize elbo, we minimize KL. The elbo has also a nice interpretation. By rewriting

$$\begin{aligned} ELBO(q) = E_{q(\mathbf{z})}[log(p(\mathbf{z}, \mathbf{x}))] - E_{q(\mathbf{z})}[log(q(\mathbf{z}))] = \\ E_{q(\mathbf{z})}[log(p(\mathbf{z}))] + E_{q(\mathbf{z})}[log(p(\mathbf{x}|\mathbf{z}))] - E_{q(\mathbf{z})}[log(q(\mathbf{z}))] = \\ E_{q(\mathbf{z})}[log(p(\mathbf{x}|\mathbf{z}))] - KL(q(\mathbf{z})||p(\mathbf{z})) \end{aligned} \tag{5}$$

By maximizing elbo:

- q tries to explain the data (first term)
- q tries to be close to the prior (second term)

Now let's try to specify the family of distributions D. If it satisfies the following

- each latent variable has its own factor
- latent variables are independent

Then, D is called mean-field variational family. In this case, q can be written as

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j) \tag{6}$$

To derive update rule, let's recall the law of iterated expectation

$$E_x[X] = E_Y[E_{X|Y}[X|Y]] = E_Y[g(Y)], \tag{7}$$

where g is the function of random variable Y. Now, we can move this law to the context of elbo. First, recall the definition of elbo:

$$ELBO(q) = E_{q(\mathbf{z})}[log(p(\mathbf{z}, \mathbf{x}))] - E_{q(\mathbf{z})}[log(q(\mathbf{z}))] \tag{8}$$

The first term can be rewritten as follows:

$$E_{q(\mathbf{z})}[log(p(\mathbf{z}, \mathbf{x}))] = E_{q(\mathbf{z})}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))] = E_j[E_{-j|j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))|z_j]] \tag{9}$$

The second term can be rewritten as follows:

$$\begin{aligned} E_{q(\mathbf{z})}[log(q(\mathbf{z}))] = E_{q(\mathbf{z})}[\sum_{j=1}^{m} log(q(z_j))] = \sum_{j=1}^{m} E_{q_j(z_j)}[log(q(z_j))] = \\ \sum_{j=1}^{m} E_j[log(q(z_j))] \end{aligned} \tag{10}$$

Now, let's assume that apart from q_j, each distribution of latent vararіables (each factor) is known and we maximize elbo only by manipulating q_j. We can now write elbo as:

$$\begin{aligned} ELBO(q) = E_j[E_{-j|j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))|z_j]] - E_j[log(q_j(z_j))] + const = \\ -KL(log(q_j)||E_{-j|j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))|z_j]) + const \end{aligned} \tag{11}$$

In this case, elbo is maximized, when KL is zero. Hence, our update rule for factor q_j is

$$q_j(z_j) = exp(E_{-j|j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))|z_j]) \tag{12}$$

To be thorough, let's simplify the notation by noting the following fact,

$$\begin{aligned} E_{-j|j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))|z_j] = \\ \int log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))q_1(z_1|z_j)..q_{j-1}(z_{j-1}|z_j)q_{j+1}(z_{j+1}|z_j)..q_m(z_m|z_j)d\mathbf{z}_{-j} = \\ \int log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))q_1(z_1)..q_{j-1}(z_{j-1})q_{j+1}(z_{j+1})..q_m(z_m)d\mathbf{z}_{-j} = \\ E_{-j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))] \end{aligned} \tag{13}$$

So, eventually

$$q_j(z_j) = exp(E_{-j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))]) \tag{14}$$

In derivation, one has to be careful when conditioning on z_i assumes random variable and when a single number.

2 Example - Gaussian Mixture Model

Here as an example, a bayesian version of Gaussian Mixture Model will be presented. For simplicity one-dimensional case will be considered. The probabilistic formulation:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma) \\ c_i &\sim Categorical(\frac{1}{k}, .., \frac{1}{k}) \\ x_i|c_i, \mu &\sim \mathcal{N}(c_i^T \mu, 1), \end{aligned} \tag{15}$$

where σ is the hyperparameter. Joint density distribution

$$p(\mu, c, x) = p(\mu) \prod_{i=1}^{n} p(c_i)p(x_i|c_i, \mu) \tag{16}$$

Here, the vector of latent variables is equal to $z = \{\mu, c\}$. Before applying cavi, we need to specify mean-field variational family.

$$q(z) = q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \phi_i) = \prod_{k=1}^{K} \mathcal{N}(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} Categorical(c_i; \phi_i) \tag{17}$$

2.1 Update rule for k-th component

General update rule is:

$$q_j(z_j) = exp(E_{-j}[log(p(z_j, \mathbf{z}_{-j}, \mathbf{x}))]) \tag{18}$$

Now step by step specific update rule will be derived. Everything that doesn't depend on μ_k will be added to "const" part.

$$q(\mu_k) = e^{E_{-j}[p(\mu,c,x)]} = e^{E_{-j}[p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i|c_i,\mu)]} \tag{19}$$

For the ease of notation, let's drop the exponent for the time being

$$
\begin{aligned}
log(q(\mu_k)) = E_{-j}[p(\mu)\prod_{i=1}^{n} p(c_i)p(x_i|c_i,\mu)] = \\
E_{-j}[\sum_{k=1}^{K} log(p(\mu_k)) + \sum_{i=1}^{n} log(p(c_i)) + log(p(x_i|c_i,\mu))] = \\
E_{-j}[log(p(\mu_k)) + \sum_{i=1}^{n} c_{ik}p(x_i|\mu_k)] + const = \\
E_{-j}[-\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} c_{ik}p(x_i|\mu_k)] + const = \\
E_{-j}[-\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} c_{ik}\mathcal{N}(x_i;\mu_k,1])] + const = \\
E_{-j}[-\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} c_{ik}\frac{(x_i-\mu_k)^2}{-2}] + const = \\
E_{-j}[-\frac{\mu_k^2}{2\sigma^2} + \sum_{i=1}^{n} c_{ik}(-\frac{x_i^2}{2} + x_i\mu_k - \frac{\mu_k^2}{2})] + const = \\
-\frac{\mu_k^2}{2\sigma^2} + \mu_k\sum_{i=1}^{n} E_{-j}[c_{ik}]x_i - \frac{\mu_k^2}{2}\sum_{i=1}^{n} E_{-j}[c_{ik}] + const = \\
-\frac{\mu_k^2}{2\sigma^2} + \mu_k\sum_{i=1}^{n} \phi_{ik}x_i - \frac{\mu_k^2}{2}\sum_{i=1}^{n} \phi_{ik} + const = \\
\mu_k^2(-\frac{1}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n} \phi_{ik}) + \mu_k\sum_{i=1}^{n} \phi_{ik}x_i + const
\end{aligned}
\tag{20}
$$

Coming back to the left expectation we have

$$
q(\mu_k) = E^{\mu_k^2(-\frac{1}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n} \phi_{ik}) + \mu_k\sum_{i=1}^{n} \phi_{ik}x_i + const} \tag{21}
$$

We can deduce that this is Gaussian distribution with mean and standard deviation specified below. This is also the update rule for the factor corresponding to μ_k.

$$
m_k = \frac{\sum_{i=1}^{n} \phi_{ik}x_i}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \phi_{ik}} \tag{22}
$$

$$
s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^{n} \phi_{ik}} \tag{23}
$$

2.2 Comments

Here, μ_k is treated as random variable. It might be confusing, but depending on the distribution we consider, some quantities might be different. The notation

very often doesn't allow us to capture it. That's why here I tried to be very clear as far as notation is concerned. So, on the one hand, we have $\mu_k \sim \mathcal{N}(0, \sigma^2)$, and its expecation we would understand as follows:

$$E[\mu_k] = \int \mu_k \mathcal{N}(\mu_k; 0, \sigma^2) d\mu_k = 0 \tag{24}$$

We've just taken expected value of random value distributed according to normal distribution with 0 mean and some standard deviation. On the other hand, if we write $E_{-j}[\mu_k]$ or $E_{q(z)}[\mu_k]$, we mean something completely different. In the context of Gaussian Mixture:

$$\begin{aligned} E_{q(z)}[\mu_k] = \int \mu_k q(\mu, c) d\mu dc = \int \mu_k q(\mu_k; m_k, s_k^2) d\mu_k = \\ \int \mu_k \mathcal{N}(\mu_k; m_k, s_k^2) d\mu_k = m_k \end{aligned} \tag{25}$$

So, we have simply a random variable, that is described by different distributions. This is how, real posterior distribution with posited one is related.

2.3 Update rule for cluster assignments

Here, the derivation is analogous to the process detailed above.

$$\begin{aligned} log(q(c_i)) = E_{-j}[p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu)] = \\ E_{-j}[\sum_{k=1}^{K} log(p(\mu_k)) + \sum_{i=1}^{n} log(p(c_i)) + log(p(x_i | c_i, \mu))] = \\ E_{-j}[log(p(c_i)) + p(x_i | c_i, \mu)] + const = E_{-j}[p(x_i | c_i, \mu)] + const = \\ E_{-j}[\sum_{k=1}^{K} c_{ik} \mathcal{N}(x_i; \mu_k, 1)] + const = \sum_{k=1}^{K} c_{ik} E_{-j}[\frac{(x_i - \mu_k)^2}{-2}] + const = \\ \sum_{k=1}^{K} c_{ik} E_{-j}[-\frac{\mu_k^2}{2} + \mu_k x_i] + const \end{aligned} \tag{26}$$

Above, $log(p(c_i))$ was absorbed by the constant term, because it is equal to $-\frac{1}{K}$. The the update rule is as follows:

$$q(c_i) = e^{\sum_{k=1}^{K} c_{ik} E_{-j}[-\frac{\mu_k^2}{2} + \mu_k x_i]} \tag{27}$$

$$\phi_{ik} = e^{E_{-j}[-\frac{\mu_k^2}{2} + \mu_k x_i]} = e^{-\frac{s_k^2 + m_k^2}{2} + m_k x_i} \tag{28}$$

Of course, we have to remember to normalize $q(c_i)$ after the update.