Machine Learning

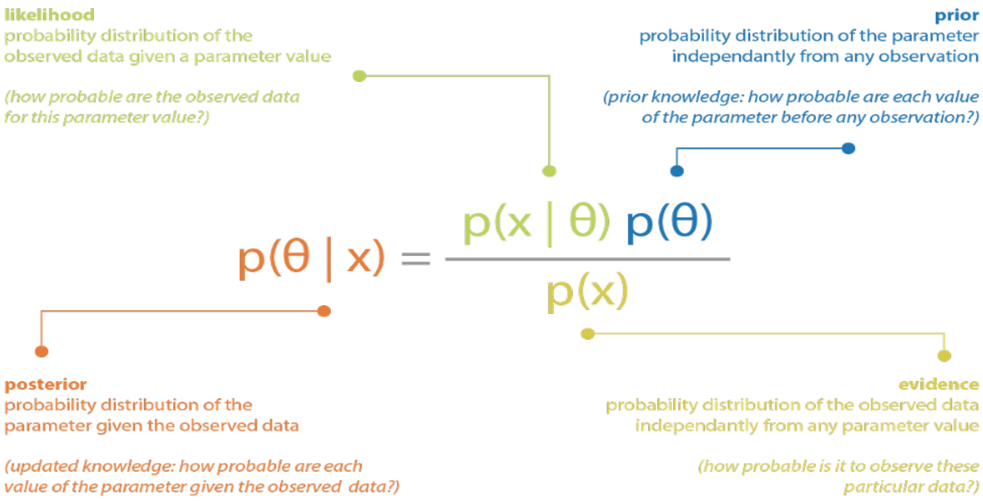
Lecture 11 - Variational Inference

Lecturer: Haim Permuter *Scribe: Daniel Duenias*

I. Introduction [2]

Statistical inference is the process of drawing conclusions such as punctual estimations, confidence intervals or distribution estimations about some latent variables in a population, based on some observed variables.

Bayesian inference is the process of producing statistical inference taking a Bayesian point of view. Bayesian paradigm is embed in the so called Bayes theorem that expresses the relation between the updated knowledge (the "posterior"), the prior knowledge (the "prior") and the knowledge coming from the observation (the "likelihood"). Let's assume a model where data x are generated from a probability distribution depending on an unknown parameter θ and that the parameter θ is distributed $p(\theta)$. Then, when data x are observed, we can update the prior knowledge about this parameter using the Bayes theorem as follows


likelihood
probability distribution of the observed data given a parameter value
(how probable are the observed data for this parameter value?)
prior
probability distribution of the parameter independantly from any observation
(prior knowledge: how probable are each value of the parameter before any observation?)
p(θ | x) = p(x | θ) p(θ) / p(x)
posterior
probability distribution of the parameter given the observed data
(updated knowledge: how probable are each value of the parameter given the observed data?)
evidence
probability distribution of the observed data independantly from any parameter value
(how probable is it to observe these particular data?)

The Bayes theorem tells us that the computation of the posterior requires three terms: a prior, a likelihood and an evidence. The first two can be expressed easily as they are part of the assumed model. However, the third term, requires to be computed such that

$$P(x) = \int_{\theta} P(x|\theta)P(\theta). \tag{1}$$

Although in low dimension this integral can be computed without too much difficulties, it can become intractable in higher dimensions.

II. Notation

consider the following notations

- x^n - an known observation vector of size n with the i'th coordinate x_i.
- z^m - a hidden/latent variable (equivalent to θ mentioned above) - vector of size m with the i'th coordinate z_i.
- z^{-i} - a group of all the vector coordinates except of the i'th one. in general, upper script notation is a vector and lower script means an entry in that vector.

III. Variational Inference

In this lecture we introduce Variational Inference (VI), a method that approximates probability densities through optimization [2]. Throughout the lecture we will use VI on a Bayesian mixture of Gaussians as an example. As mentioned earlier, we are interested in computing the posterior distribution,

$$P(z^m|x^n) = \frac{P(z^m, x^n)}{P(x^n)}. \tag{2}$$

IV. Bayesian mixture of Gaussians

A Bayesian mixture of Gaussians is a model that assumes that the data is distributed as a mixture of k Gaussians with the following parameters [3]:

- the expectation is also a random variable, normally distributed - $\mu_i \sim \mathcal{N}(0, \sigma^2), i = 1, 2...k$, for some known σ.
- given the expectation μ_i, the standard deviation of the Gaussian is 1 - $G_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$.

- $P(c_i)$ - the probability that x_i belongs to some Gaussian, i.e, the assignment of each i'th observation, has a uniform distribution, $c_i \sim Uniform(k)$.

We encode c_i into 'one hot' k sized vector. We then draw that $x_i|c_i, \mu^k \sim \mathcal{N}(c_i^T \mu, 1)$. define $z^m = (\mu^k, c^n), m = k + n$, an m sized vector of hidden variables.

For a sample of size n, the joint density of latent and observed variables is,

$$\begin{aligned} P(x^n, z^m) =& P(x^n, c^n, \mu^k) \\ =& P(\mu^k) \prod_{i=1}^{n} P(c_i) P(x_i|c_i, \mu^k). \end{aligned} \tag{3}$$

Here, the evidence is,

$$P(x^n) = \int P(\mu^k) \prod_{i=1}^{n} \sum_{c_i} P(c_i) P(x_i|c_i, \mu^k). \tag{4}$$

Therefore, eq. (2) is an equation with time complexity of numerically evaluating Kdimensional integral - $O(K^n)$ [1].

V. THE EVIDENCE LOWER BOUND (ELBO) [1]

In variational inference, we specify a family of densities over the latent variables - $q(z^m)$. We then try to approximate the exact conditional distribution $P(z^m|x^n)$ with that densities family, i.e, find the closest $q(z^m)$ to the conditional distribution $P(z^m|x^n)$. That is done by solving the following optimization problem:

$$q^*(z^m) = \underset{q(z^m)}{\operatorname{argmin}}\ D(q(z^m)||P(z^m|x^n)). \tag{5}$$

We know that by definition,

$$D(q(z^m)||P(z^m|x^n)) = E_{q(z^m)}[\log q(Z^m)] - E_{q(z^m)}[\log P(Z^m|x^n)]. \tag{6}$$

Therefor, using Bayes and logarithm rules we get,

$$D(q(z^m)||P(z^m|x^n)) = E_q[\log q(Z^m)] - E_q[\log P(Z^m, x^n)] + E_q[\log P(x^n)]. \tag{7}$$

Expand the conditional knowing that $p(x^n)$ is not a function of the random variable Z^m and therefor $E_q[\log P(x^n)] = \log P(x^n)$,

$$\log P(x^n) = D(q(z^m)||P(z^m|x^n)) + E_q[\log P(Z^m, x^n)] - E_q[\log q(Z^m)]. \tag{8}$$

Noting the term $E_q[\log P(Z^m, x^n)] - E_q[\log q(Z^m)]$ as ELBO and using the non negativity of D we get,

$$\begin{aligned} \log P(x^n) =& D(q(z^m)||P(z^m|x^n)) + ELBO \\ \geq& ELBO. \end{aligned} \tag{9}$$

And thereby its name - The Evidence ($P(x^n)$) Lower Bound. Using eq. (9) and the fact that w.r.t $q(z^m)$, $\log P(x^n) = const$, we can write D as,

$$D(q(z^m)||P(z^m|x^n)) = -ELBO + const. \tag{10}$$

So, by maximizing $ELBO$ we actually minimize $D(q(z^m)||P(z^m|x^n))$, therefore we may solve the optimization problem

$$\begin{aligned} q^*(z^m) =& \underset{q(z^m)}{\text{argmin}}\ D(q(z^m)||P(z^m|x^n)) \\ =& \underset{q(z^m)}{\text{argmax}}\ ELBO, \end{aligned} \tag{11}$$

instead of solving eq. (5).

In addition, using Bayes and logarithms rules , ELBO can be written as,

$$\begin{aligned} ELBO =& E_q[\log P(Z^m)] + E_q[\log P(x^n|Z^m)] - E_q[\log q(Z^m)] \\ =& E_q[\log P(x^n|Z^m)] - D(q(z^m)||P(z^m)). \end{aligned} \tag{12}$$

Knowing eq. (12) and eq. (11), we can present another interpretation of our optimization problem:

$$\begin{aligned} \underset{q(z^m)}{\text{argmax}}\ ELBO =& \underset{q(z^m)}{\text{argmax}}\ E_q[\log P(x^n|z^m)] - D(q(z^m)||P(z^m)) \\ =& \underset{q(z^m)}{\text{argmax}}\ \sum q(z^m) \log P(x^n|z^m) - D(q(z^m)||P(z^m)). \end{aligned} \tag{13}$$

Looking at eq. (13) we can see that there is a trade-off between minimizing D and maximizing the sum. As for minimizing D, we would like $q(z^m)$ to be as close as possible to $P(z^m)$, while as for maximizing the sum, we understand that $q(z^m)$ which gives more weight to z^m that make the term $\log P(x^n|z^m)$ bigger i.e, z^m that contains more information about x^n, will get better results. As we can see, the more samples there are, the more significant the term $\sum q(z^m) \log P(x^n|z^m)$ will be over the divergence.

VI. Coordinate ascent - Alternating maximization procedure

We will use a method called coordinate ascent. This method is a maximization method of a multi-variable functions. In this method we fix all the variables except one, maximize the function as an normal one variable function, then again fix all the variables except the next one and repeat. If the function is concave in all of its variables, the method will get the global maximum, otherwise, a local one [4], [5].

Example 1 (maximizing two variable function using coordinate ascent) Suppose we want to maximize $f(x, y)$:

Algorithm 1: coordinate ascent / Alternating maximization procedure

```
Input: $f(x, y)$.
Output: $x, y$ of Local/Global maximum of $f(x, y)$.
initiate $y_0$ to some value.
solve $x_0 = \max_x f(x, y_0)$
$i = 0$
while $f(x_i, y_i)$ not converged do
    $i = i + 1$
    $y_i = \max_y f(x_{i-1}, y)$
    $x_i = \max_x f(x, y_{i-1})$
    Compute $f(x_i, y_i)$
end
Return $x_i, y_i$
```

VII. Coordinate ascent mean-field variational inference [1]

In a mean-field variational family the latent variables are mutually independent. A generic member of the family is

$$q(z^m) = \prod_{i=1}^{m} q(z_i). \tag{14}$$

We assume that this is the case in our problem. Using Bayes and logarithm rules, we can write $ELBO$ as,

$$\begin{aligned} ELBO &= E_q[\log P(x^n)] + E_q[\log P(Z^m|x^n)] - E_q[\log q(Z^m)] \\ &= \log P(x^n) + E_q[\log P(Z^m|x^n)] - E_q[\log q(Z^m)] \\ &= const + E_q[\log P(Z^m|x^n)] - \sum_{i=1}^{n} E_{q(z_i)}[\log q(Z_i)], \end{aligned} \tag{15}$$

while the second transaction is because $P(x^n)$ is not random in $q(z^m)$ and the third one is by using eq. (14). Applying coordinate ascent and fixing $q(z^{-i})$ (all $q(z_m)$ except of the i'th coordinate) we get,

$$\underset{q(z_i)}{\text{argmax}}\ ELBO = \underset{q(z_i)}{\text{argmax}}\ q(z_i)E_{q(z^{-i})}[\log P(Z_i, z^{-i}|x^n)] - E_{q(z_i)}[\log q(Z_i)] + const, \tag{16}$$

$$\frac{\partial ELBO}{\partial q(z_i)} = E_{q(z^{-i})}[\log P(Z_i|z^{-i}x^n)] - \log q(z_i) + 1 = 0. \tag{17}$$

Which yields,

$$\log q^*(z_i) \propto E_{q(z^{-i})}[\log P(z_i|Z^{-i}x^n)], \tag{18}$$

$$q^*(z_i) \propto \exp(E_{q(z^{-i})}[\log P(z_i|Z^{-i}x^n)]). \tag{19}$$

Therefore, coordinate ascent variational inference (CAVI) - algorithm may be written as:

```
Algorithm 2: CAVI
Input: model $P(x^n, z^m)$, data $x^n$.
Output: variation density $q(z^m) = \prod_{i=1}^{m} q(z_i)$ and ELBO (evidance lower bound
  of $P(x^n)$).
Initialization - initiate $q(z_i)$ for some $i$.
while the ELBO has not converged do
    for i = 1,2...m do
        Set $q(z_i) \propto \exp(E_{q(z^{-i})}[\log P(z_i|Z^{-i}x^n)])$
    end
    Compute $ELBO = E_q[\log P(Z^m, x^n)] - E_q[\log q(Z^m)]$
end
Return $\prod_{i=1}^{m} q(z_i)$, ELBO
```

Note that in order to compute $E_q[\log P(Z^m, x^n)]$ we use,

$$E_q[\log P(Z^m, x^n)] = \sum_{z^m} q(z^m) \log P(x^n, z^m). \tag{20}$$

VIII. CAVI FOR A BAYESIAN MIXTURE OF GAUSSIANS MODEL [1]

As we saw before, we need to find $P(z^m|x^n)$ and the term is hard to compute so we will approximate it with $q(z^m)$. In order to do so we will use the CAVI algorithm modified to our example. We assume now that the mixture of Gaussians is defined by the parameters φ, m^k, s^k as follows:

- the expectation is normally distributed - $\mu_i \sim \mathcal{N}(m_i, s_i^2), i = 1, 2...k$.
- $P(c_i)$ - has a categorical distribution, $c_i \sim \varphi_i^k$ (k sized vector of non-negative number that sums to 1). Therefore, φ is an $n * k$ matrix - the row i is a k sized vector noted φ_i.

That said we define the initialization of the algorithm like the model presented in section IV: $\mu_i \sim \mathcal{N}(0, \sigma^2), i = 1, 2...k$, the expectation of each Gaussian (σ^2 is a known hyper parameter) and, $\varphi_i \sim Uniform(k), i = 1, 2...n$. In each iteration we will update our distributions parameters φ, m^k, s^k.

Let us evaluate the ELBO of the mixture assuming mean field family,

$$\begin{aligned} ELBO(\varphi, m^k, s^k) = &\sum_{i=1}^{k} E[\log P(\mu_i); m_i, s_i^2] \\ &+ \sum_{j=1}^{n} E[\log P(c_j); \varphi_j] + E[\log P(x_j|c_j, \mu^k); \varphi_j, m^k, (s^2)^k] \\ &- \sum_{j=1}^{n} E[\log q(c_j; \varphi_j)] - \sum_{i=1}^{k} E[\log q(\mu_i; m_i, s_i^2)]. \end{aligned} \tag{21}$$

Expanding equation (21) using equation (19) we derive that that the following holds:

$$\varphi_{ji} \propto \exp(E[\mu_i; m_j, s_j^2] x_j - E[\mu_i^2; m_j, s_j^2]/2). \tag{22}$$

$$q(\mu_i) \propto \exp(E[\log p(\mu_i) + \sum_{j=1}^{n} E[\log P(x_j|c_j, \mu^k); \varphi_j, m^{-i}, (s^2)^{-i}]). \tag{23}$$

Continue developing those eqations we eventually get that the update for $q(\mu_i)$ is,

$$m_i = \frac{\sum_{j=1}^{n} \varphi_{ji} x_j}{1/\sigma_2 + \sum_{j=1}^{n} \varphi_{ji}}, s_i^2 = \frac{1}{1/\sigma_2 + \sum_{j=1}^{n} \varphi_{ji}}. \tag{24}$$

Therefore, we can write the algorithm as follows:

```
Algorithm 3: CAVI for mixture of Gussians model
Input: Data $x^n$, number of components K, prior variance of component means
  $\sigma^2$.
Output: Variational densities $q(\mu_i; m_i, s_i^2)$ (Gaussian) and $q(c_i; \varphi_i)$
  (K-categorical).
Initialization as discribed in the beginning of this section.
while the ELBO has not converged do
    for j = 1,2...n do
        Set $\varphi_{ji} \propto \exp(E[\mu_i; m_j, s_j^2] x_j - E[\mu_i^2; m_j, s_j^2]/2)$
    end
    for i = 1,2...k do
        Set $m_i = \frac{\sum_{j=1}^{n} \varphi_{ji} x_j}{1/\sigma_2 + \sum_{j=1}^{n} \varphi_{ji}}$
        Set $s_i^2 = \frac{1}{1/\sigma_2 + \sum_{j=1}^{n} \varphi_{ji}}$
    end
    Compute $ELBO(\varphi, m^k, (s^2)^k)$
end
Return $q(\varphi, m^k, (s^2)^k)$
```

References

[1] D. M. Blei, A. Kucukelbir, J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association* , 112:859-877, 2017.

[2] https://towardsdatascience.com/bayesian-inference-problem-mcmc-and-variational-inference-25a8aa9bce29

[3] https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf

[4] I. Naiss and H. H. Permuter, ”Alternating maximization procedure for finding the global maximum of directed information,” 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010, pp. 000545-000549, doi: 10.1109/EEEI.2010.5662161.

[5] I. Naiss and H. H. Permuter, ”Extension of the Blahut–Arimoto Algorithm for Maximizing Directed Information,” in IEEE Transactions on Information Theory, vol. 59, no. 1, pp. 204-222, Jan. 2013, doi: 10.1109/TIT.2012.2214202.