Variational Inference with Coin Toss example

AntiCodeOn
email: anticodeon@gmail.com

November 2017

1 Introduction

Variational inference (VI) is a machine learning method for approximation of the difficult to compute probability densities that appear in Bayesian methods. In general, the goal is to find a surrogate distribution over hidden variables and parameters of the original Bayesian model that is close to the original distribution but where the evaluation is computationally efficient. The closeness is defined in terms of a divergence measure between the approximate distribution and the original distribution.

2 Motivation

Consider classical Bayesian setup

$$p(Z|X) = \frac{p(X,Z)}{p(X)} \tag{1}$$

where $X = x_{1:n}$ denotes the observed data while $Z = z_{1:k}$ denotes all the hidden variables and parameters of the model. Numerator of the fraction is called joint distribution

$$p(X,Z) = p(X|Z)\,p(Z) \tag{2}$$

and the denominator (often called evidence) is a marginal distribution

$$p(X) = \int_Z p(X,Z)\,dZ \tag{3}$$

Computing the posterior probability is often a very hard problem (for example, due to the exponentially large number of hidden states or because required integrations do not have a closed-form analytical solution). This is why we resort to techniques such as approximation, where the objective is to find a joint distribution $q(Z)$ that replaces original posterior $p(Z|X)$ in a way that enables us to find computationally tractable solutions.

2.1 Variational inference setup

Variational inference posits a set of densities $\mathcal{Q}$ over the latent variables Z. We typically try to minimize the Kullback-Leibler (KL) divergence

$$KL(q(Z)\,||\,p(Z|X)) = E_q\left[log\,\frac{q(Z)}{p(Z|X)}\right] \tag{4}$$

of the approximate and the original, true joint distribution

$$q^*(Z) = \underset{q(Z)\in\mathcal{Q}}{\arg\min}\; KL(q(Z)\,||\,p(Z|X)) \tag{5}$$

KL divergence tells us how much information we loose by choosing the approximate distribution instead of true posterior distribution. We can expand the right hand side of (4)

$$KL(q(Z)\,||\,p(Z|X)) = E_q[q(Z)] - E_q[log\,p(Z|X)] \tag{6}$$

Using (1) we get

$$KL(q(Z)\,||\,p(Z|X)) = p(X) - \underbrace{(E_q[log\,p(X,Z)] - E_q[q(Z)])}_{\text{ELBO(q)}} \tag{7}$$

We want to minimize the KL divergence by varying distribution q(Z). $p(X)$ does not depend on q and is always positive (or equal to zero) so we need to maximize the $ELBO(q)$. The only assumption we are making is that $q(Z)$ factorizes in the following way

$$q(Z) = \prod_{k=1}^{K} q_k(z_k) \tag{8}$$

which means that we are breaking possible dependencies between hidden variables in our approximation process.

$$ELBO(q) = \int q(Z)\,log\,p(X,Z)\,dZ - \int q(Z)\,log\,q(Z)\,dZ \tag{9}$$

In our problem we are dealing with the discrete distribution. Fortunatelly, the setup is the same, we only need to changethe integral with the summation terms.

3 Variational inference mixture model

For each observation x_n we have a corresponding latent variable z_n comprising a 1-of-K binary vector with the elements z_{nk} for $k = 1, \ldots, K$. We denote the observed data set by $X = x_1, \ldots, x_N$, and similarly we denote the latent variables by $Z = z_1, \dots, z_N$. We can write down the conditional distribution of Z, given the mixing coefficients π, in the form:

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \tag{10}$$

We introduce priors over the parameter π. The mathematical analysis is considerably simplified if we choose conjugate prior distribution to model the parameter π. We choose a Dirichlet distribution over the mixing coefficients π. The parameter α can be interpreted as the effective prior number of observations associated with each component of the mixture. If the starting α value is small (bellow 1), then the posterior distribution is primarily influence by the data rather then our choise of prior.

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0) \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{11}$$

Next, we write down the conditional distribution of the observed data vectors, given the latent variables and the component parameters. Each of the five experiments X comprises of the ten independent Bernoulli events.

$$p(X|Z,\Theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} \prod_{j=1}^{J} [\Theta_k^{x_n^j}(1-\Theta_k)^{1-x_n^j}]^{z_{nk}} \tag{12}$$

We can model priors for the Θ parameter with the conjugate prior distribution of the Bernoulli distribution. As we are dealing with the coin biasses, which are defined on the interval [0, 1], beta distribution lends itself as a natural choice for our model.

$$p(\Theta|a,b) = \prod_{k=1}^{K} \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{Beta(a,b)} \tag{13}$$

Note. The Beta in the denominator denotes the Beta function, not the Beta probability distribution.

Joint distribution is given by

$$p(X, Z, \Theta, \pi, a, b) = p(X|Z,\Theta)p(Z|\pi)p(\pi)p(\Theta|a,b) \tag{14}$$

with factors on the right side of the equation as defined above. Only variables $X = x_1, \dots, x_N$ were observed. In the next step we consider a variational distribution with the following factorization

$$q(Z, \pi, \Theta) = q(Z)q(\pi, \Theta). \tag{15}$$

We are separating the latent variables and parameters and this is actually the only assumption we are making. In order to proceed with the factorization, we use the general result given by Bishop[2006].

$$ln\, q^*(Z) = E_{\pi,\Theta}[ln\, p(X, Z, \Theta, \pi, a, b)] + const \tag{16}$$

For the factor of **Z** we are only interested in those components of the decomposition 14 that depend on **Z**. All other terms are absorbed into the additive normalization constant, giving:

$$ln\, q^*(Z) = E_\pi[ln\, p(Z|\pi)] + E_\Theta[ln\, p(X|Z,\Theta)] + const \tag{17}$$

where

$$E_\pi[ln\, p(Z|\pi)] = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} E_\pi[ln\, \pi_k] \tag{18}$$

and

$$E_\Theta[ln\, p(X|Z,\Theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} E_\Theta[\sum_{j=1}^{J} ln\, \Theta_k^{x_n^j}(1-\Theta_k)^{1-x_n^j}] \tag{19}$$

Hence

$$ln\, q^*(Z) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(E_\pi[ln\, \pi_k] + E_\Theta[ln\, \Theta_k^{n_H}(1-\Theta_k)^{n_T}]\right) + const \tag{20}$$

where n_H is a number of heads and n_T number of tails in the n-th experiment. Proceeding similarly for π and Θ we get the following expression

$$ln\, q^*(\pi,\Theta) = E_Z[ln\,(\pi) + ln\, p(Z|\pi)] + E_Z[ln\, p(X|Z,\Theta) + ln\, p(\Theta)] + const \tag{21}$$

Since there are no expressions that depend on both parameters at the same time we may proceed with the factorization of the parameters independently

$$q(\pi,\Theta) = q(\pi)\, q(\Theta) \tag{22}$$

where

$$ln\, q^*(\pi) = ln\, p(\pi) + E_Z[ln\, p(Z|\pi)] + const \tag{23}$$

$$ln\, q^*(\pi) = \sum_{k=1}^{K}(\alpha_k - 1)\ln\, \pi_k + \sum_{k=1}^{K} r_k\, ln\, \pi_k + const \tag{24}$$

where $r_k = \sum_{n=1}^{N} E[z_{nk}]$ We have already derived the expectations over Z in EM document. Taking the exponential of equation 24 we get that the

$$q(\pi) = Dir(\pi|\boldsymbol{\alpha}) \tag{25}$$

where $\boldsymbol{\alpha}$ is vector with K components $\boldsymbol{\alpha}_k = \alpha_k + r_k$

$$ln\, q^*(\Theta) = ln\, p(\Theta) + E_Z[ln\, p(X|Z,\Theta)] + const \tag{26}$$

$$ln\, q^*(\Theta) = \sum_{k=1}^{K}(ln\,[\Theta_k^{a_k-1}(1-\Theta_k)^{b_k-1}] + \sum_{n=1}^{N} r_{nk} \sum_{j=1}^{J} ln[\Theta_k^{x_n^j}(1-\Theta_k)^{1-x_n^j}]_{+} const \tag{27}$$

$$ln\, q^*(\Theta) = \sum_{k=1}^{K} ln\, [\Theta_k^{a_k - 1 + \sum_{n=1}^{N} n_H r_{nk}} (1-\Theta_k)^{b_k - 1 + \sum_{n=1}^{N} n_T r_{nk}}] + const \tag{28}$$

where n_H equals number of the outcomes of $x_n^j = 1$ (number of heads), and n_T number of the outcomes of $x_n^j = 0$ (number of tails). We see that the factor for Θ is Beta distributed variable:

$$q^*(\Theta) = \prod_{k=1}^{K} Beta(\Theta|A, B) \tag{29}$$

where A is a vector with parameter components $a = a_k + \sum_{n=1}^{N} n_H r_{nk}$ and B with components $b = b_k + \sum_{n=1}^{N} n_T r_{nk}$.

4 Variational lower bound

Variational lower bound L is given by

$$L = E[ln\, p(X, Z, \Theta, \pi, a, b)] - E[ln\, q(Z, \Theta, \pi, a, b)] \tag{30}$$

Using results from 14, 15 and 22 this simplifies to

$$\begin{aligned} L =& E[ln\, p(X|Z, \Theta)] + E[ln\, p(Z|\pi)] + E[ln\, p(\pi)] + E[ln\, p(\Theta|a, b)] \\ &- E[ln\, q(Z)] - E[ln\, q(\pi)] - E[ln\, q(\Theta)] \end{aligned} \tag{31}$$

We write down each of the terms separately:

$$E[ln\, p(\pi)] = ln\, C(\alpha_0) + \sum_{k=1}^{K} (\alpha_k - 1)(\psi(\alpha_k) - \psi(\sum_{k=1}^{K} \alpha_k)) \tag{32}$$

$$E[lnp(\Theta_k|a, b)] = \sum_{k=1}^{K} -lnBeta(a_k, b_k) + (a_k - 1)\psi(a_k) - (b_k - 1)\psi(b_k) + (a_k + b_k - 2)\psi(a_k + b_k) \tag{33}$$

$$E[ln\, p(Z|\pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}(\psi(\alpha_k) - \psi(\sum_{k=1}^{K} \alpha_k)) \tag{34}$$

$$E[ln\, p(X|Z, \Theta)] = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}[n_H(\psi(a_k) - \psi(a_k + b_k)) + n_T(\psi(b_k) - \psi(a_k + b_k))] \tag{35}$$

5 Algorithm

The following table gives the algorithmic description of the problem.

Algorithm 1 Coordinate ascent algorithm

```
Require: Data $x_{1:n}$, number of components K, model
Ensure: Variational densities q(z), q(Θ), q(π)
  Initialize: Variational parameters $a_{1:K}$, $b_{1:K}$, $\alpha_{1:K}$
  repeat
    Calculate responsibilities $r_{nk} = E[z_{nk}]$
    for $k = 1$ to $K$ do
      $\alpha_k^{new} = \alpha_k^{old} + r_k$
      $a_k^{new} = a_k^{old} + r_k$
      $b_k^{new} = b_k^{old} + r_k$
    end for
    Compute $ELBO(q) = E[log\, p(z, x)] + E[log\, q(z)]$
  until ELBO has not converged
```

6 Results

EM. Original paper where the problem of determining coin toss biases appeared uses Expectation Maximization algorithm in order to find the solution. However, EM algorithm gives us only the point estimate of the bias values as shown in Figure 1.

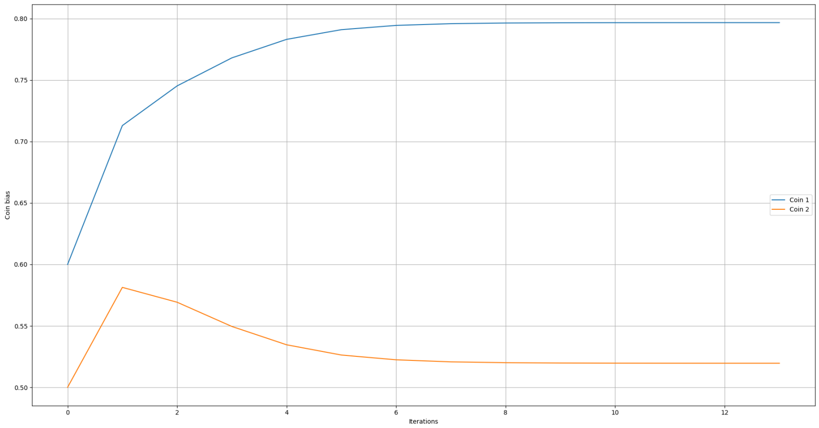

Figure 1: Coin bias convergence curve using Expectation Maximization algorithm

VI. In contrast, variational inference (Bayes in general) uses the underlaying prior distribution of the parameters. These can be further modeled with the additional distributions, or as we have done in our solution, initialized using hyperparameters. The result of the Variational Inference algorithm contains not only the most probable solution but also the information about confidence

Iteration 0		Iteration 10		Iteration 19	
Coin1	Coin2	Coin1	Coin2	Coin1	Coin2
0.3234	0.6765	0.3857	0.6142	0.4249	0.5750
0.0103	0.9896	0.0096	0.9903	0.0107	0.9892
0.0264	0.9735	0.0268	0.9731	0.0303	0.9696
0.5541	0.4458	0.6403	0.3596	0.6794	0.3205
0.0660	0.9339	0.0724	0.9275	0.0823	0.9176

Table 1: Coin responsibilities

in the calculated values. Figure 2 shows how confidence in the values grows as we increase number of iterations.

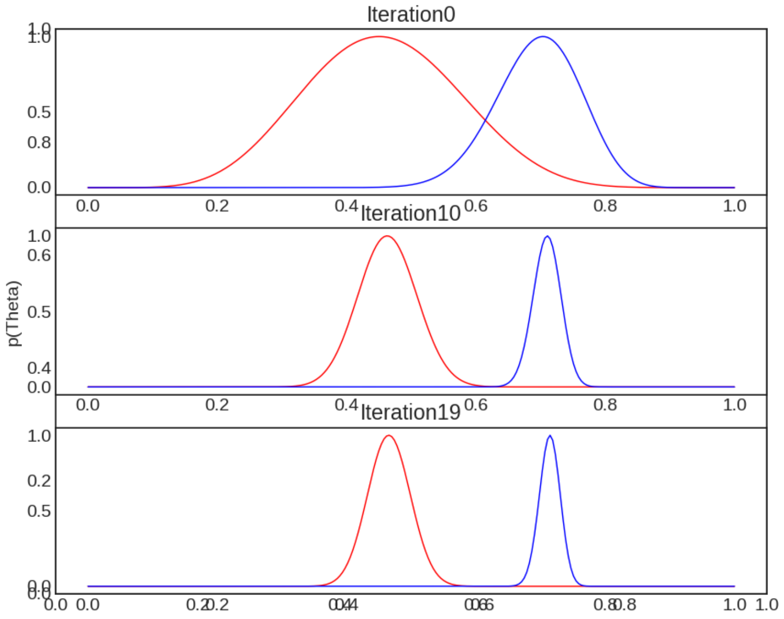

Figure 2: Coin bias convergence curve using Variational Inference algorithm

Responsibilities. Also, we have modeled the distribution of the K components, therefore we can track how expected values of the classes converge to their real values. This is show using the Table 1 for the same iteration steps as in Figure 2. We call this responsibilities, because it shows us the probability that the coin k (columns) is responsible for the event n (rows) in the iteration i of the algorithm.

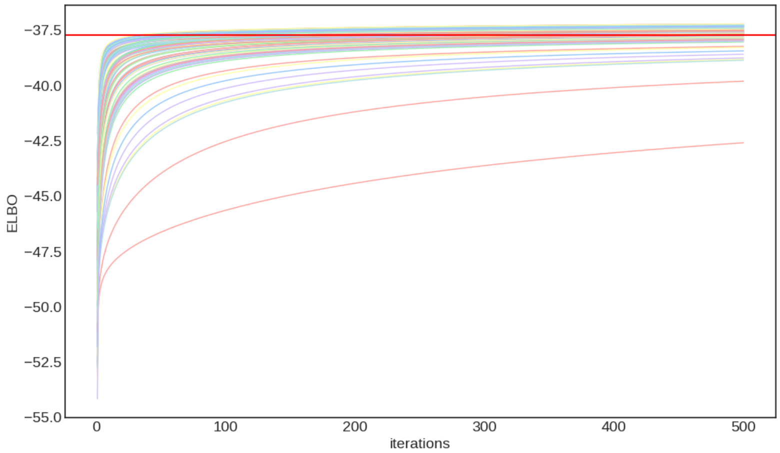

Figure 3: Different initializations converge to different local optima of the ELBO

Initialization. The ELBO is generally a non-convex objective function. Convergence to the local optimum is guaranteed, however, it can be sensitive with regard to the initialization. Figure 3 shows the ELBO convergence for 100 random initializations of the model parameters. Red line denotes the average of the convergence asymptote for all runs. **Convergence.** We define the treshold for the change of ELBO value between subsequent algorithm iterations. Once this difference has fallen under the treshold the procedure stops. **Calculation**. During the calculation of the coefficient (which may grow arbitrarily large) we have encountered the problems with the calculation of the normalization coefficient for Dirichlet distribution. Replacing the hand-crafted version with the one from *scipy* package resolved the problem.

7 Conclusion

We described variational inference algorithm with mean-field approximation. Update equations for a simple mixture models were derived. Evidence lower bound optimization procedure was used as a criteria for algorithm convergence. We have choosen probability distribution from exponential family with conjugate priors in order to simplify the mathematical model construction.