STA561: Probabilistic machine learning

Variational Inference (11/04/13)

Lecturer: Barbara Engelhardt *Scribes: Matt Dickenson, Alireza Samany, Tracy Schifeling*

1 Introduction

In this lecture we will further discuss Variational Inference and then proceed with Loopy Belief Propagation, also known as LBP. In the last lecture we discussed Mean Field Variational Inference and we investigated the matter in the context of univariate Gaussians. In this lecture, we will talk further about Variational Inference in the context of the Ising Model, which is a Markov random field, and then we proceed with Mean Field and Loopy Belief Propagation.

Let us begin by reviewing Variational Inference. In Variational Inference, we have our data set as $\mathcal{D} = \{X_1, X_2, \ldots, X_n\}$ and we describe both our latent variables, $Z_{1:m}$, and the set of our model parameters, θ, by a single parameter Z such that $Z = \{Z_{1:m}, \theta\}$.

We are interested at estimating the posterior probability of latent variables given the data, which is $p(Z|X)$. By definition of the conditional distribution, we know that the posterior probability of latent variables given data is equal to the joint distribution of latent variables and data divided by the marginal probability of the data(observed variables). Therefore, we have:

$$p(Z|X) = \frac{p(Z,X)}{\int_Z p(Z,X)dZ}$$

Variational Inference is another method to compute an approximation to the posterior distribution. In general, computing the posterior distribution for several distributions, such as truncated Gaussians or Gaussians mixture models, can be very hard. This issue obliges us to resort to computing approximations for the posterior distribution. For example, consider the case of Gaussian Mixture Models. Let us assume the standard mixture model where each class specific mean μ_k is distributed according to $\mu_k \sim \mathcal{N}(0, \tau^2)$, our latent variables Z_i have the distribution $Z_i \sim Mult(\pi)$ for a fixed hyperparameter π, and the data X_i is distributed as $X_i \sim \mathcal{N}(\mu_k, \Sigma_k)$. So our model will be:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \tau^2) \\ Z_i &\sim Mult(\pi) \\ X_i &\sim \mathcal{N}(\mu_k, \Sigma_k) \end{aligned}$$

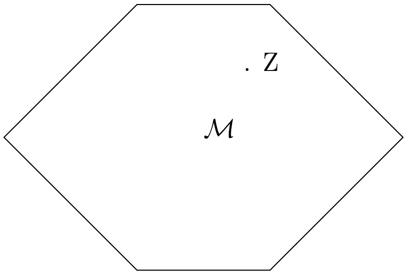
which leads to the following expression for the posterior distribution using the chain rule:

$$p(\mu, Z|X) = \frac{\prod_{k=1}^K p(\mu_k) x \prod_{i=1}^n p(Z_i|\pi) x p(Z_i|X_i, \mu_{1:k})}{\int_{\mu_{1:K}} \Sigma_{Z_{1:n}} \prod_{k=1}^K p(\mu_k) x \prod_{i=1}^n p(Z_i|\pi) x p(Z_i|X_i, \mu_{1:k}) d\mu_{1:K}}$$

The numerator is simple to compute. However, computing the denumerator is exponentially difficult. In addition, the integral in the denumerator is generally not straightforward. These problems serve as our motivation for finding approximate posterior distributions. We will need to tackle the same problem again once we discuss LDA models, which also lead to exponentially hard denumerators.

Let us return to our previous discussion on the space of possible parametrizations of our original posterior distributions named $\mathcal{M}$. It turns out that if we consider all the values of Z which are consistent with the posterior distribution derived above, we have a theorem that for most models such as Gaussian models and many discrete models, for the given data the Z parameters lie in a convex polytope.

The immediate advantage of this fact is that we can do optimization over a convex space to find the parameters Z. However, the density on the space M might not have a form convenient to optimize as we have previously for the Gaussian Mixture Model.

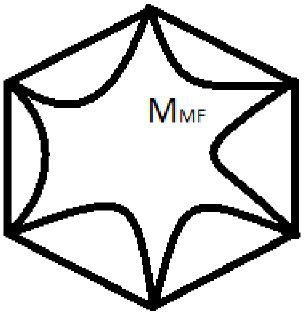


Convex polytope $\mathcal{M}$ conatining the possible parameter values Z

Let us review Mean Field in order to fully understand our goal in Variational Inference. Our goal in the Mean Field is to find a fully factorized posterior distribution. We assume an approximation to the posterior in the form of $q(Z) = \prod_{i=1}^{m} q(Z_i)$ and attempt to obtain the following:

$$\min_{Z} KL(q(Z)||p(Z|X))$$

We should note that each variable is assumed to be independent of the others here. Remember that based on the reverse KL divergence, we need the support of $q(Z)$ to lie entirely in the support of $p(Z|X)$. The marginal independence assumption we imposed here leads us to a parameterization space which is no longer convex as shown in the figure below. However, this new space, called $\mathcal{M}_{MF}$, turns out to be straightforward to search over when we use computable expressions for $q(Z_i)$. We should note that it is very possible that the desired Z^* does not lie in $\mathcal{M}_M F$.



Here we do the optimization above using an unnormalized $p(Z|X)$, but let us look at the optimzation above from an entirely different perspective by using a concept called the Evidence Lower Bound, also known as ELBO.

1.1 Evidence Lower Bound

Let us recall Jensen's Inequality first. According to jensen's Inequality, we know that for a given convex function $f(\cdot)$ we have $E[f(x)] \leq f(E[x])$. This can be derived directly using the definition of expectation and applying the convexity property of $f(\cdot)$. As $\log(x)$ is a convex function, we can apply Jensen's Inequality to it. Let us consider $\log(p(X))$. We have:

$$\begin{aligned} log(p(X)) &= \log \int_Z p(X,Z) dZ \\ &= \log \int_Z p(X,Z) x \frac{q(Z)}{q(Z)} dZ \\ &= \log E_q[\frac{p(X,Z)}{q(Z)}] \end{aligned}$$

Using Jensen's Inequality, we have:

$$\log(p(X)) = \log E_q[\frac{p(X,Z)}{q(Z)}] \geq E_q[\log(\frac{p(X,Z)}{q(Z)})] = E_q[\log(p(X,Z))] - E_q[\log(q(Z))]$$

Let us recall a result we derived in the last lecture:

$$KL(q||\tilde{p}) = -E_q[\log(p(Z,X)] + E_q[\log(q(Z)] + \log(p(X)) \Rightarrow KL(q||\tilde{p}) = -ELBO + \log(p(X))$$

Therefore, ELBO is KL divergence with unnormalized posterior. Since $p(X)$ is independent of $q(Z)$, minimizing the KL divergence is equivalent to minimizing the ELBO. Therefore there are two steps to minimize the KL divergence:

1. Choosing a proposition, q, such that the expectations are computable.
2. Maximize $q(Z)$ with respect to the ELBO in order to obtain the tightest possible approximation to $p(X, Z)$

2 Ising Model

Before proceeding with variational inference, it is helpful to review the Ising model. The main idea behind the Ising model is a lattice of unobserved variables $(x_1, ...x_n)$, each with its own (noisy) observation $(y_1, ..., y_n)$.

For example, suppose our goal is to reconstruct a denoised image given noisy observations of the pixels. We can think of the lattice as the pixels in a black and white image $(x_i \in \{-1, 1\})$, with a noisy grayscale observation of the pixels $(y_i \in R)$. More generally, we wish to draw inferences about the unobserved lattice X from the observed values Y. Figure 1 illustrates an Ising model for $n = 9$, with the latent nodes colored white and the observed nodes shaded grey.

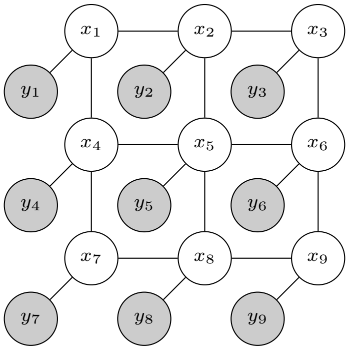


Figure 1: Schematic of an Ising Model

We now define the potential functions of the Ising model:

$$\begin{aligned} \psi_s(x_s) &= p(y_i|x_i) \equiv L_i(x_i) \\ \psi_{st}(x_s x_t) &= W_{st} x_s x_t \end{aligned}$$

Continuing with our image example above, we could set $W_{st} = 1$. In general, we set W to positive values if we want neighbors to agree, and negative values if we want them to differ.

Let $N(i)$ be a function that returns the first-degree neighbors of node i. For example, in Figure 1, calling $N(x_1)$ would return nodes x_2 and x_4.

Now we can specify functions for our prior:

$$p(x) = \frac{1}{z_0} \exp\{-\sum_{i=1}^{n} \sum_{j \in N(i)} x_i x_j\}$$

$$p(y|x) \quad = \quad \prod_{i=1}^{n} \exp\{-L_i(x_i)\}$$

From this, we have the posterior:

$$p(x|y) \quad = \quad \frac{1}{z}\exp\{-\sum_{i=1}^{n}\sum_{j\in N(i)} x_i x_j - \sum_{i=1}^{n} L_i(x_i)\}$$

2.1 Mean Field Version of the Ising Model

Having seen an example of a basic Ising model, we now turn our attention to how we can analyze the mean field version of such a model. We do this by "breaking" the edges between the latent variables. We add a mean value (or variational parameter) μ to each x, such that $\mu_i = \mathbb{E}[x_i]$. The new structure is illustrated in Figure 2, with the same color coding as above.

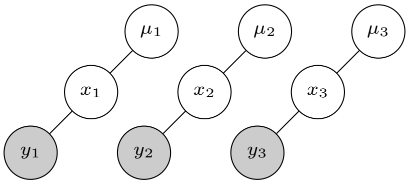


Figure 2: Mean Field Version of an Ising Model

$$\begin{aligned} q(x) &= \prod_{i=1}^{n} q_i(x_i) \\ \log(q_i(x_i)) &= \mathbb{E}_{\neg q}[\log \tilde{p}(x)] \end{aligned}$$

We can maximize $\log(q_i(x_i))$ with a coordinate ascent method, as discussed previously in class.

Now, we rewrite the ELBO in terms of the Ising model:

$$\begin{aligned} \log(q_i(x_i)) &= \mathbb{E}_{\neg q_i}[x_i \sum_{j\in N(i)} x_j + L_i(x_i) + c] \\ q_i(x_i) &\propto \exp\{x_i \sum_{j\in N(i)} \mu_j + L_i(x_i)\} \end{aligned}$$

Thus,

$$q_i(x_i = +1) \quad = \quad \frac{\exp\{\sum_{j\in N(i)} \mu_j + L_i(+1)\}}{\sum_{x_i'\in\{+1,-1\}} \exp\{\sum_{j\in N(i)} \mu_j + L_i(x_i')\}}$$

Note the resemblance of this function to a sigmoid function $\frac{1}{1+q_i(x_i=-1)}$. Effectively, this q_i is an approximation of the marginalized posterior, or basically a Gibbs step.

For the mean field, we now iteratively update:

$$\begin{aligned}\mu_i &= +1(q_i(x_i = +1)) + -1(q_i(x_i = -1)) \\ z_i &\propto \exp[\mathbb{E}_{\neg q}[\log p(z|x)]]\end{aligned}$$

In the model, we have three parameters of interest: μ_1, μ_2, and μ_{12}, where $\mu_{12} = \mathbb{E}[\psi_{12}(x_1 x_2)]$. We have the constraint $0 \leq \mu_{12} \leq \mu_1, \mu_2$. We limit acceptable values to those within the portion of the simplex that satisfies this constraint. We can visualize this in Figure 3, where the acceptable values occupy the space under the shaded face.

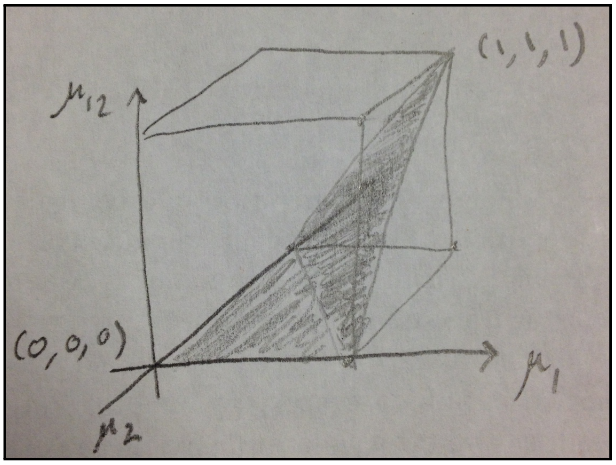


Figure 3: Visualizing Constraints on μ

If we take a slice of this convex polytope in Figure 3 at $\mu_1 = \mu_2$, it is a triangle like the one in Figure 4. The x-axis in the figure is μ_1. The quadratic curve indicates where $\mu_1^2 = \mu_{1,2}$. This quadratic function is due to the marginal independence of μ_1 and μ_2.

3 Loopy Belief Propagation (LBP)

Now q (our approximation to the true distribution) is not necessarily a valid joint probability. Instead, q has what we call "local consistency" or "pairwise consistency," meaning q satisfies the following conditions:

- $\sum_{x_i} q_i(x_i) = 1$
- $\sum_{x_i} q_{ij}(x_i, x_j) = q_j(x_j)$ for all pairs i, j

But we do not promise that $\sum_{x_i, x_j, x_k} q_{ijk}(x_i, x_j, x_k) = 1$.

Recall that belief propagation in a tree is exact; we get our final answer after an inside-outside pass. This is not true in a graphical model with loops. We apply belief propagation to a loopy graph as follows:

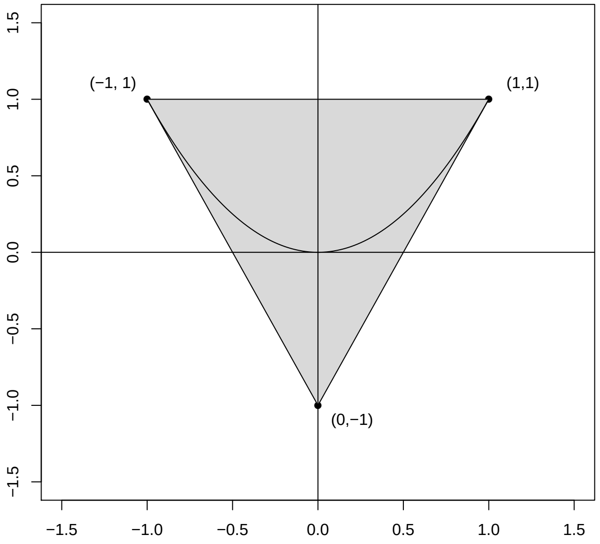


Figure 4: Visualizing μ_{12} when $\mu_1 = \mu_2$

1. Initialize the messages $m_{s\to t}(x_t) = 1$. Initialize the beliefs $\mu_s = 1$.
2. Send messages; $m_{s\to t}(x_t) = \displaystyle\sum_{x_s} L_s(x_s) x_s x_t \prod_{u\in N(s)\setminus t} m_{u\to s}(x_s)$. Update beliefs of each node $\mu_s \propto L_s(x_s)\prod_{t\in N(s)} m_{t\to s}(x_s)$.
3. Repeat until convergence (i.e., until the beliefs do not change a lot).

LBP does not always converge. There are some methods that help with this problem, including:

- Dampening to avoid oscillation.
- Asynchronous updates.
- Scheduling (update the beliefs in an order that makes sense) using a tree-based reparameterization.

The local consistency requirement means there are fewer constraints on the space $\mathcal{M}$, and so $\mathcal{M}_{LBP}$ is an outer-polytope that contains the polytope of $\mathcal{M}$. See Figure 22.7 (c) in the textbook.

Comparing mean field and LBP, note that

- Mean field is not exact in trees, and LBP is exact for trees.
- Mean field optimizes over node marginals, and LBP optimizes over node and edge marginals.
- Mean field as more local optima than the LBP.