CS7140: Advanced Machine Learning **Spring 2020**

Lecture 5: Variational Inference

Lecturer: Rose Yu *Scribes: Stephen Schmidt*

5.1 Recap

5.1.1 Importance Sampling

- Method for solving **Problem 1 (**Estimating expectations of functions)
- $\Phi = E_{\mathbf{x}\sim\mathbf{P(x)}}[\phi(x)] = \int \mathbf{P}(\mathbf{x})\phi(\mathbf{x})d(\mathbf{x})$
- Sample from an easier distribution Q(x) rather than the difficult P(x) to compute $\hat{\Phi}$

Importance:

$$w_r = \frac{P^*(x^{(r)})}{Q^*(x^{(r)})}$$

Estimate Expectation

$$\hat{\Phi} = \frac{\sum_r w_r \phi(x^{(r)})}{\sum_r w_r}$$

5.1.2 Rejection Sampling

- Method for solving **Problem 2 (**generating samples from a difficult distribution P(x)
- choose a Q(x) such that it envelopes all of P(x): $cQ^*(x) > P^*(x)$

1. Draw a sample x from Q(x)
2. Draw a point u uniform randomly selected from [0,$cQ^*(x)$]
3. Reject x if u > $P^*(x)$, else accept

5.1.3 Metropolis-Hastings Method

- Method for solving **Problem 2 (**generating samples from a difficult distribution P(x)
- Example of MCMC

1. Draw a sample x from Q($x; x^t$)
2. evaluate $a = \frac{p^*(x)Q(x^{(t)};x)}{P^*(x^{(t)})Q(x;x^{(t)})}$
3. If a 1,accept x, set $x^{(t+1)} = x$; else reject, set $x^{(t+1)} = x^{(t)}$

5.2 Approximate Inference

5.2.1 Background

KL Divergence

$$D_{KL}(Q||P) = \sum_x Q(x)\log(\frac{Q(x)}{P(x)})$$

- Equal to 1 only when P(x) = Q(x)
- Always greater than 0
- Non-Symmetric: Does not qualify as a true distance function
  - $D_{KL}(Q||P) \neq D_{KL}(P||Q)$

When probabilistic graphical model is decomposed into a product of factors, the distribution P(x) is represented as:

$$P(x) = \frac{1}{Z}P^*(x) = \frac{1}{Z}\prod_{m=1}^{M}\phi(x_m)$$

Applying this into the KL Divergence equation:

$$D_{KL}(Q||P) = log(Z) - \sum_M E_Q[log(\phi)] - H_Q \geq 0$$

Where H_Q is the entropy of out proposal distribution.

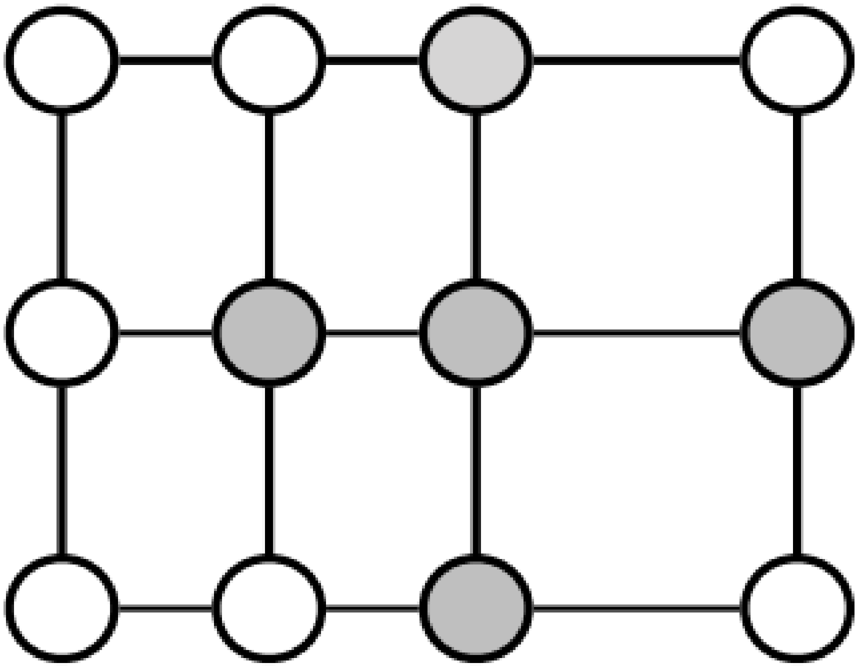
We can describe second half of this equation as the **Variational Free Energy**

$$F[P^*, Q] = -\sum_m E_Q[log(\phi)] - H_Q$$

In addition, the partition function is a lower bound on the Variational Free Energy

$$log(Z) \geq -F[P^*, Q]$$

5.2.2 Ising Model Example



- Binary distribution, node can have one of two values: +1, -1
- A simple markov random field

$$p(x|\beta, J) = \frac{1}{Z} \exp[-\beta E(x; J)]$$

Where β is an inverse temperature constant

Where J is a 2D matrix represting the interaction between node i and j at J_{ij}

$E(x; J)$ represents the energy function

$$E(x; J) = -\frac{1}{2} \sum_{mn} J_{mn} x_m x_n - \sum_n h_n x_n$$

The normalizing function for this equation is

$$Z(\beta, \mathbf{J}) = \sum_x exp[-\beta E(x; \mathbf{J})]$$

Evaluating the normalization constant, as well as describing the properites of the probabliity distribution are hard. So we should use an objective function Q(x,θ) to approximate. Using the variational free energy to measure the quality of our approximation results in

$$F(P^*, Q) = \beta E_Q[E(x; J)] - H_Q$$

Where $\beta E_Q[E(x; J)]$ is the mean-energy

and $H_Q$is the entropy of our proposal distribution

We will approximate with the following seperable distribution:

$$Q(x; a) = \frac{1}{Z_Q} exp(\sum_n a_n x_n)$$

The entropy of this distribution is given by:

$$H_Q = \sum_x Q(x; a) ln(\frac{1}{Q(x; a)}$$

and the mean energy is given by:

$$\langle E(x; J) \rangle_Q = \sum_x Q(x; a) E(x; J)$$

Since the distribution is seperable, the entropy of the approximating distribution is the sum of the entropies of each node, where the probability that each node n is equal to 1 is given by

$$q_n = \frac{e^{a_n}}{e^{a_n} + e^{-a_n}} = \frac{1}{1 + exp(-2a_n)}$$

and the entropy of an individual node is given by

$$H(Q) = q ln(\frac{1}{q}) + (1 - q) ln(\frac{1}{(1 - q)})$$

Since the mean energy is a sum of a product of two independent variables, we can simplify using the mean value for a single node as

$$E_Q(x_n) = \frac{e^{a_n} - e^{-a_n}}{e^{a_n} + e^{-a_n}} = 2q_n - 1$$

With this information we can now find the mean energy:

$$E_Q[E(x; J)] = \sum_x Q(x, a)[-\frac{1}{2} \sum_{m,n} J_{mn} x_m x_n - \sum_n h_n x_n]$$

$$= -\frac{1}{2} \sum_{m,n} J_{mn} \bar{x}_m \bar{x}_n - \sum_n h_n \bar{x}_m$$

with both the mean energy and the entropy of our Q(x) we now can form the full variational free energy

$$F(P^*, Q) = \beta(-\frac{1}{2}\sum_{m,n} J_{mn}\bar{x_m}\bar{x_n} - \sum_n h_n\bar{x_m}) - \sum_n H(q_n)$$

We are trying to find our a values, which are our variational parameters. So to optimize, we take the derivative with respect to a_m. We can start with the entropy

$$\frac{\partial}{\partial a_m} H(q_n) = ln(\frac{1-q}{q}) = -2a$$

Now we have

$$\frac{\partial}{\partial a_m}\beta\widetilde{F}(a) = \beta[-\sum_n J_{mn}\bar{x_n} - h_m](2\frac{\partial q_m}{\partial a_m}) - ln(\frac{1-q_m}{q_m})(\frac{\partial q_m}{\partial a_m})$$

$$= 2(\frac{\partial q_m}{\partial a_m})[-\beta(\sum_n J_{mn}\bar{x_n} + h_M) + a_m]$$

Solving for a_m results in

$$a_m = \beta(\sum_n J_{mn}\bar{x_n} + h_m)$$

References

[1] David J.C. MacKay “Information Theory, Inference, and Learning Algorithms,”

For further reading, consult:

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American Statistical Association 112.518 (2017): 859-877.