Problem Set 5 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Gaussian Graphical Model

1.1 Derive $\phi_{ij}(x_i, x_j)$ and $\phi_i(x_i)$

$$P(X \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \Sigma}} exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)) \tag{1}$$

$$P(X \mid \mu, \Omega) \propto exp(-\frac{1}{2}X^T \Omega X + (\Omega\mu)^T X) \tag{2}$$

$$= \prod_{i \in V} exp(-\frac{1}{2}x_i^T \Omega_{ii} x_i + (\Omega\mu)_i^T x_i) \prod_{(i,j) \in E} exp(-\frac{1}{2}x_i^T \Omega_{ij} x_j) \tag{3}$$

$$\psi_{ij}(x_i, x_j) = exp(-\frac{1}{2}x_i^T \Omega_{ij} x_j) \tag{4}$$

$$\psi_i(x_i) = exp(-\frac{1}{2}x_i^T \Omega_{ii} x_i + (\Omega\mu)_i^T x_i) \tag{5}$$

1.2 Prove $(i,j) \notin E \iff X_i \perp X_j | X_{V \setminus \{i,j\}}$

If $(i,j) \notin E$, there is no edge between node i and node j, then we can get:

$$\Omega_{ij} = 0, \qquad \psi_{ij}(x_i, x_j) = 1 \tag{6}$$

$$P(x_i, x_j \mid V\setminus\{i,j\}, \mu, \Omega) = \psi'_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j) \cdot \psi_i(x_i)^{\frac{1}{n(i)}} \cdot \psi_j(x_j)^{\frac{1}{n(j)}} \tag{7}$$

$$= \psi_i(x_i)^{\frac{1}{n(i)}} \cdot \psi_j(x_j)^{\frac{1}{n(j)}} \tag{8}$$

$$= P(x_i \mid V\setminus\{i\}, \mu, \Omega) \cdot P(x_j \mid V\setminus\{j\}, \mu, \Omega) \tag{9}$$

$$= P(x_i \mid V\setminus\{i,j\}, \mu, \Omega) \cdot P(x_j \mid V\setminus\{i,j\}, \mu, \Omega) \tag{10}$$

We can notice the above process could be inversed, which means if we know that the two points are independent, we could get that there is no edge between these two nodes. Therefore, we could prove the conclusion.

2 Sampling

2.1 Inverse Sampling

2.1.1 Prove Inverse Sampling

First we need to prove for y' $\in$ R, then for $0 < z' < 1$, we have:

$$P(h^{-1}(z) <= y') = P(infy : h(y) = z < y') \tag{11}$$

$$= P(z <= h(y')) = h(y') \tag{12}$$

$$P(h(y) < z') = P(y < h^{-1}(z')) = z' \tag{13}$$

Therefor, we could prove that inverse sampling is reasonable.

The drawback is in real application, it would require a closed form expression of F(x), which could be untractable in some cases.

2.1.2 Find the Cauchy distribution transfermation

Given the density function, we could derive:

$$h(y) = \int_{-\infty}^{y} p(y')dy' = \frac{1}{2} + \frac{1}{\pi}arctan(y) \tag{14}$$

$$y = g(z) = h^{-1}(z) = tan(\pi(z - \frac{1}{2})) \tag{15}$$

2.2 Rejection Sampling

As described, we would see that the accpeted probability is the area of the $\tilde{p}(z)$ curve:

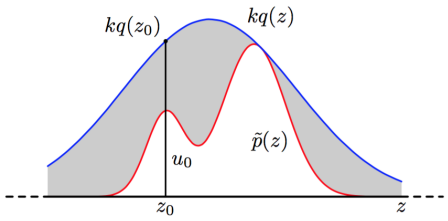

Figure 1: The representation of rejection sampling

$$P(u_0) = P(accepted) = \int \frac{\tilde{p}(z)}{kq(z)} q(z)dz = \frac{1}{k} \int \tilde{p}(z)dz \tag{16}$$

Based on the figure representation, we could have an intuitive prove on the rejection sampling algorithm. Basically, we know that the k is the smallest value to keep $kq(z) > \tilde{p}(z)$. The probability of the points that are rejected by this method depends on the ratio of area under the unnormailized distribution $\tilde{p}(z)$ to the area under the cureve of $kq(z)$. If the points fall into the grey area, then it would be rejected, as it is not generated by the $\tilde{p}(z)$.

The drawback of this rejection sampling method could be that during the sampling iterations, the algorithms would generate too many rejected points if the real distribution is "small", which would make it slow to converge.

2.3 Markov Chain Monte Carlo

2.3.1 MH Algorithms

MH is a MCMC method for obtaining a sequence of random samples. For a Markov chain, the condition for a chain to have stationary distribution is to satisfy the perporty of detailed balance:

$$p(z^{t+1}) = \sum_{z^t} T(z^{t+1}, z^t) p(z^t) \tag{17}$$

$$p(z^{t+1}) T(z^t, z^{t+1}) = p(z^t) T(z^{t+1}, z^t) \tag{18}$$

Here $T(z, z')$ is the transfer matrix or transfer condition. For the MH algorithm, we have:

$$p(z^t) q(z^*|z^t) A(z^*, z^t) = min(p(z^t) q(z^*|z^t), p(z^*) q(z^t|z^*)) \tag{19}$$

$$= min(p(z^*) q(z^t|z^*), p(z^t) q(z^*|z^t)) \tag{20}$$

$$= p(z^*) q(z^t|z^*) A(z^t, z^*) \tag{21}$$

Therefore, we could see that MH algorithm would satisfy the perporty of detailed balance.

The drawback of MH is that this algorithm would also have "rejection". When we choose 1 in as the transfer value in a iteration, the new sample is actually "rejected". The state of the sample point does not change in this case.

2.3.2 & 2.3.3 Gibbs Sampling and Comparison with MH

Bascially, to prove that Gibbs Sampling would achieve the p (x) stationary distribution, we only need to prove that the Gibbs Sampling is a special case of MH. If the essence of Gibbs Sampling is the similar to the MH, then we could derive that it satisfy the property of detailed balance, with which we could prove that it would converge to achieve the stationary distribution.

We could obtain Gibbs Sampling as a particular instance of MH. For a iteration, we sample the z_j while fixing the remaining $z_{\setminus j}$, so we would have:

$$q_j(z^*|z) = p(z_j^*|z_{\setminus j}), \quad z_{\setminus j}^* = z_{\setminus j} \tag{22}$$

$$A(z^*, z) = \frac{p(z^*) q_j(z|z^*)}{p(z) q_k(z^*|z)} \tag{23}$$

$$= \frac{p(z^*|z_{\setminus j}^*) p(z_{\setminus j}^*) p(z_k|z_{\setminus j}^*)}{p(z|z_{\setminus j}) p(z_{\setminus j}) p(z_k^*|z_{\setminus j})} = 1 \tag{24}$$

Therefore, we would know that the Gibbs Sampler is the special case of MH, with the transfer value always 1 and accepted. Intuively, we could see that in Gibbs Sampling, we would only change one "dimension" of the data state in one iteration.

In this case, we would know that the Gibbs Sampler would not suffer from the problem as the MH algorithm. The Gibbs Sampler would not be "rejected" in each iteration, since the accepted value is always 1. However, in each iteration, the sampling would only change one "dimension" of the data state.

3 Expectation Maximization and Variational Inference

3.1 EM

3.1.1 Prove Equation

$$lnp(x|\theta) = lnp(x, z|\theta) - lnp(z|x, \theta) \tag{25}$$

$$= lnp(x, z|\theta) - lnq(z) - (lnp(z|x, \theta) - lnq(z)) \tag{26}$$

$$= \int_z q(z) ln \frac{p(x, z|\theta)}{q(z)} - \int_z q(z) ln \frac{p(z|x, \theta)}{q(z)} \tag{27}$$

$$= L(q, \theta) + KL(q||p) \tag{28}$$

3.1.2 E-Step Maximization

In the E-step, we can find if $q(z) = p(z|x, \theta)$, then $KL(q||p) = 0$. Therefore, in this E step, the $L(q, \theta)$ is maximized.

3.1.3 M-Step Maximization

In the M-step, if we fix $q(z)$, the $KL(q||p)$ would not change, since p and q are fixed. We know $lnp(x|\theta) = L(q, \theta) + KL(q||p)$, so if KL would not change, and L was maximized, thus, the $lnp(x|\theta)$ was maximized.

3.2 VI

3.2.1 Prove the Equation

$$L(q) = \int \prod_k q_k(z_k) \{lnp(x, z) - \sum_k q_k(z_k)\} dz_k \tag{29}$$

$$= \int q_k \{ \int lnp(x, z) \prod_{k \neq j} q_j dz_j \} dz_k - \int q_k lnq_k dz_k + const \tag{30}$$

$$= \int q_k(z_k) E_{k \neq j}[lnp(x, z)] dz_k - \int q_k(z_k) lnq_k(z_k) dz_k + const \tag{31}$$

3.2.2 Prove the Optimal

First we could see the first iterm $E_{k \neq j}[lnp(x, z)]dz_k + const$ as a new distribution $m(x, z)$, then we wcould treat the L (q) as the the KL divergence between $m(x, z)$ and $\{q_j\}_{j \neq k}$. Then if we fix $\{q_j\}_{j \neq k}$, maximizing L is the same to minimize the KL divergence.

To minimize the KL divergence, we would know that when the two distributions are the same, the KL divergence are the smallest, therefore, we have:

$$E_{k \neq j}[lnp(x, z)]dz_k + const = lnq_k^*(z_k) \tag{32}$$

4 HMM

4.1 Comparison

1. We should place $<$ here, the first item is smaller than the second item

To prove it, we could view the left as $P(A, B)$, the right as $P(A|B)$. The $P(B)$ is smaller than 1. So the left item is smaller than the right item.

2. We should place $<$ here, the first item is smaller than the second item

To prove it, we could calculate the different part of the two probability. For the first item, p = 0.3 * 0.7 * 0.1 + 0.7 * 0.3 * 0.2. For the second item, p' = 0.7 * 0.1. So we could see the first item is smaller than the second.

3. We should place $=$ here, the first item is the same with the second item

To prove it, we could view the left as $P(A, B)$, the right as $P(A|B)$. The $P(B)$ is the same with 1. So the left item is the same with the right item.

4. We should place $<$ here, the first item is smaller than the second item

To prove it, we could calculate the different part of the two probability. For the first item, p = (0.2 + 0.1) * 0.1. For the second item, p' = 0.4 * 0.6. So we could see the first item is smaller than the second.

4.2 Prove the Equation

$$p(x_1, x_2, ..., x_i, z_i) = p(x_i|x_1, x_2, ..., x_{i-1}, z_i) \cdot p(x_1, x_2, ..., x_{i-1}, z_i) \tag{33}$$

Because of the Markov property, we would know x_i is only related to z_i. So we would have $p(x_i|x_1, x_2, .., z_i) = p(x_i|z_i)$. Therefore:

$$p(x_1, x_2, ..., x_i, z_i) = p(x_i|z_i) \cdot \sum_{z_{i-1}} p(x_1, x_2, ..., x_{i-1}, z_{i-1}) p(z_i|z_{i-1}) \tag{34}$$

5 Bayesian Networks

I found related research task stating about the DAG number of a bayes net, the link is:

http://bnt.googlecode.com/svn/trunk/docs/usage.html#structure_learning

Based on its statement and prove, we would know the number $G(n)$ is:

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \begin{pmatrix} n \\ k \end{pmatrix} 2^{k(n-k)} G(n-k) \tag{35}$$

6 Collaboration

I discussed with Zheng Chen with Problem 4 and 5.