Problem Set 6 Solutions

May 7, 2018

1. Associate each node $i \in V$ with a random variable X_i, such that $X_i = 1$ if $i \in A$, otherwise $X_i = 0$. Define singleton potential $\phi(x_i) = e^{\theta_i x_i}$, edge potential $\psi(x_i, x_j) = \mathbb{1}\{x_i + x_j > 0\}$, then the MRF over the graph is:

$$P(x) = \frac{1}{Z} \prod_{i \in V} \phi(x_i) \prod_{(i,j) \in E} \psi(x_i, x_j) = \frac{1}{Z} \prod_{i \in V} e^{\theta_i x_i} \prod_{(i,j) \in E} \mathbb{1}\{x_i + x_j > 0\}$$

It's easy to see that for any valid vertex cover A, we have $P(A) \propto w(A)$.

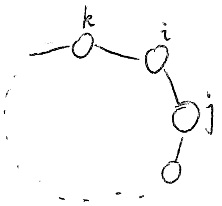
2. $$\arg\min_{A \subset V} P(A) = \arg\min_{X : P(X) > 0} P(X) = \arg\min_{x : P(x) > 0} \prod_{i \in V} e^{\theta_i x_i} = \arg\min_{x : P(x) > 0} \sum_{i \in V} \theta_i x_i = \arg\max_{x : P(x) > 0} \sum (-\theta_i) x_i$$

So we can ~~redefine~~ negate the weights ($\theta_i := -\theta_i$) and simply run max-product. Alternatively, we can directly apply min-product on the original model, where the MIN operation is used (instead of MAX or SUM).

3. The sum-product algorithm attempts to find a fixed point m^* of the update equations:

$$m_{ij}(x_j) \propto \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in Nb(i) \setminus j} m_{ki}(x_i)$$

Consider a K-cycle graph and the message from i to j, with k being the other neighbor of i:



$$m_{ij}(x_j) \propto \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) m_{ki}(x_i)$$

Plugging in $\phi_i(x_i) = e^{x_i}$, $\psi_{ij}(x_i, x_j) = \mathbb{1}\{x_i + x_j > 0\}$ gives

$$m_{ij}(1) \propto m_{ki}(0) + e\, m_{ki}(1), \quad m_{ij}(0) \propto e\, m_{ki}(1)$$

so $$m_{ij}(1) = \frac{m_{ki}(0) + e\, m_{ki}(1)}{2e\, m_{ki}(1) + m_{ki}(0)}, \quad m_{ij}(0) = \frac{e\, m_{ki}(1)}{2e\, m_{ki}(1) + m_{ki}(0)}$$

At convergence, a fixed point m^* should satisfy $m^*_{ij} = m^*_{ki}$ due to symmetry in the graph, thus

$$\begin{cases} m^*_{ki}(1) = \dfrac{m^*_{ki}(0) + e\, m^*_{ki}(1)}{2e\, m^*_{ki}(1) + m^*_{ki}(0)} \\ m^*_{ki}(0) = \dfrac{e\, m^*_{ki}(1)}{2e\, m^*_{ki}(1) + m^*_{ki}(0)} \\ m^*_{ki}(1) \geq 0 \\ m^*_{ki}(0) \geq 0 \end{cases}$$

Solving the system of equations gives the converged messages (identical for all neighboring nodes $k, i \in V$):

$$m^*_{ki}(0) = \frac{3e - \sqrt{e(4+e)}}{4e - 2} \simeq 0.437$$

$$m^*_{ki}(1) = \frac{e - 2 + \sqrt{e(4+e)}}{4e - 2} \simeq 0.563$$

The answer is independent of K, the size of the cycle.

Once we have m^*, we can also compute the converged beliefs easily:

$$b^*_i(x_i) = \gamma_i \phi_i(x_i) \prod_{k \in Nb(i)} m^*_{ki}(x_i)$$

$$b^*_{ij}(x_i, x_j) = \gamma_{ij} \psi_{ij}(x_i, x_j) \phi_i(x_i) \phi_j(x_j) \prod_{k \in N(i) \setminus j} m^*_{ki}(x_i) \prod_{k' \in N(j) \setminus i} m^*_{k'j}(x_j)$$

4. In Gibbs sampling, we sample x_i' from $P(X_i \mid x_{\neg i})$ (where $x_{\neg i}$ is the configuration of all the other variables sampled in previous iterations) with the following probability: (see Koller PGM text section 12.3.3 for more details)

$$P(x_i' \mid x_{\neg i}) \underset{\text{by local Markov property}}{=} P(x_i' \mid x_{Nb(i)}) = \frac{\phi_i(x_i') \prod_{j \in Nb(i)} \psi_{ij}(x_i', x_j)}{\sum_{x_i''} \phi_i(x_i'') \prod_{j \in Nb(i)} \psi_{ij}(x_i'', x_j)}$$

where $\phi_i(x_i) = e^{\theta_i x_i}$. $\psi_{ij}(x_i, x_j) = \mathbb{1}\{x_i + x_j > 0\}$. To sample from $P(x)$, we simply iteratively sample each $x_i, i \in V$ from $P(x_i \mid x_{\neg i})$ while holding the other variables fixed.

5. $$\ell(x^{(1)}, \dots, x^{(M)} \mid \theta) = \frac{1}{M} \sum_m \log P_\theta(x^{(m)}) = \frac{1}{M} \sum_m \Big(\sum_{i \in V} \theta_i x_i^{(m)} + \sum_{(i,j) \in E} \log \mathbb{1}\{x_i^{(m)} + x_j^{(m)} > 0\} \Big) - \log Z$$

$$\frac{\partial \ell}{\partial \theta_i} = \frac{1}{M} \sum_m x_i^{(m)} - \underbrace{\cancel{E_\theta[X_i]}}_{P_\theta(x_i = 1)} = E_{\hat{P}}[X_i] - E_\theta[X_i]$$ (see Koller text section 20.3.1)

6. $$\ell_{PL}(x^{(1)}, \dots, x^{(M)} \mid \theta) = \frac{1}{M} \sum_m \sum_i \log P(x_i^{(m)} \mid x_{\neg i}^{(m)}, \theta)$$

We have derived each conditional probability $P(x_i^{(m)} \mid x_{\neg i}^{(m)})$ in problem 4.

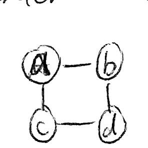
The general form of the derivative is given in section 20.6.1 of Koller text, equation (20.23).

$$\frac{\partial \ell_{PL}}{\partial \theta_i} = \sum_{j: X_j \in Scope[f_i]} \Big(\frac{1}{M} \sum_m f_i(x^{(m)}) - E_{x_j' \sim P_\theta(X_j \mid x_{\neg j}^{(m)})} [f_i(x_j', x_{\neg j}^{(m)})] \Big)$$

In our specific problem, $f_i(x_i) = x_i$, $scope[f_i] = \{x_i\}$, (i.e., the factor f_i associated with θ_i involves x_i only), so

$$\frac{\partial \ell_{PL}}{\partial \theta_i} = \frac{1}{M} \sum_m f_i(x_i^{(m)}) - E_{x_i' \sim P_\theta(x_i \mid x_{\neg i}^{(m)})} [f_i(x_i')] = \frac{1}{M} \sum_m x_i^{(m)} - P_\theta(x_i = 1 \mid x_{\neg i}^{(m)})$$

7. ① Clearly, any graph structure (particularly, the edge set) inconsistent with the data yields a log likelihood of $-\infty$, so we only need to consider graphs consistent with data. The observations $\{a,d\}$, $\{b,c\}$ imply $(b,c) \notin E$ and $(a,d) \notin E$, respectively. To achieve maximum likelihood, we therefore pick the most complex model under these constraints, i.e., $E = \{(a,b), (b,d), (a,c), (c,d)\}$ (see Koller text Section 20.7.3.1 for justification).



The set of valid configurations under this model are $\{a,d\}, \{b,c\}, \{a,b,c\}, \{b,c,d\}, \{a,b,d\}, \{a,c,d\}, \{a,b,c,d\}$

So $$Z = \sum_x \prod_{i \in V} \phi(x_i) \prod_{(i,j) \in E} \mathbb{1}\{x_i + x_j > 0\} = e^{\theta_a + \theta_d} + e^{\theta_b + \theta_c} + e^{\theta_a + \theta_b + \theta_c} + e^{\theta_b + \theta_c + \theta_d} + e^{\theta_a + \theta_b + \theta_d} + e^{\theta_a + \theta_c + \theta_d} + e^{\theta_a + \theta_b + \theta_c + \theta_d}$$

Given $M = 3$ samples $\{a,d\}$, $\{a,d\}$, $\{b,c\}$, with ℓ_2 regularizer $-\frac{\lambda}{2}\|\theta\|_2^2$, $\lambda = 100$

$$\frac{\partial \ell}{\partial \theta_a} = \frac{1}{M} \sum_m x_a^{(m)} - \frac{\partial}{\partial \theta_a} \log Z - \lambda \theta_a = \frac{2}{3} - \frac{1}{Z}\left(e^{\theta_a + \theta_d} + e^{\theta_a + \theta_b + \theta_c} + e^{\theta_a + \theta_b + \theta_d} + e^{\theta_a + \theta_c + \theta_d} + e^{\theta_a + \theta_b + \theta_c + \theta_d}\right) - 100\,\theta_a$$

Similarly the derivatives w.r.t. θ_b, θ_c, and θ_d can be obtained. Running gradient ascent gives $\theta^* = [-4.8 \times 10^{-4}, -3.8 \times 10^{-3}, -3.8 \times 10^{-3}, -4.8 \times 10^{-4}]$ with optimal log likelihood $\ell^* = -1.95$

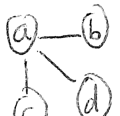
② Given a single sample $\{a\}$, the log-likelihood is $\ell = \log P(x_a = 1, x_b = x_c = x_d = 0)$

~~i) if the edge set is empty, then the MRF distribution factorizes over each variable, with~~ $\cancel{P(x_a) = e^{\theta_a x_a} / (e^{\theta_a} + 1)}$.

$\cancel{\frac{\partial \ell}{\partial \theta_a} = 1 - E_\theta[X_a] = 1 - P_\theta(X_a = 1) = 1 - e^{\theta_a}/(e^{\theta_a}+1)}$ ~~Gradient ascent yields~~ $\cancel{\theta_a \to \infty}$

~~ii) if the edge set is not empty~~ Like before, to maximize the likelihood, we choose the largest edge set compatible with the sample $\{a\}$, i.e., $E = \{(a,b), (a,c), (a,d)\}$

7. (2) cont'd The valid configurations are $\{a\}, \{a,b\}, \{a,c\}, \{a,d\}, \{a,b,c\}, \{a,b,d\}, \{a,c,d\}, \{a,b,c,d\}, \{b,c,d\}$



So $$P_\theta(X_a=1) = \frac{\partial}{\partial\theta_a}\log Z = \frac{\frac{\partial}{\partial\theta_a}Z}{Z} = \frac{e^{\theta_a}+e^{\theta_a+\theta_b}+e^{\theta_a+\theta_c}+e^{\theta_a+\theta_d}+e^{\theta_a+\theta_b+\theta_c}+e^{\theta_a+\theta_c+\theta_d}+e^{\theta_a+\theta_b+\theta_c+\theta_d}}{e^{\theta_a}+e^{\theta_a+\theta_b}+e^{\theta_a+\theta_c}+e^{\theta_a+\theta_d}+e^{\theta_a+\theta_b+\theta_c}+e^{\theta_a+\theta_c+\theta_d}+e^{\theta_a+\theta_b+\theta_c+\theta_d}+e^{\theta_b+\theta_c+\theta_d}}$$

To find the MLE θ_a^*, the moment matching condition

$$\frac{\partial \ell}{\partial\theta_a} = 1 - \mathbb{E}_\theta[X_a] = 1 - P_\theta(X_a=1) = 0$$ would require $\theta_a^* \to \infty$