Learning Theory - Homework 3

Alexandru Mocanu

May 17, 2019

1 Exercise 1

1) By simple computation we get:

$$q(\underline{x}_{t+1}|\underline{x}_t)p(\underline{x}_t) = \tilde{q}(\underline{x}_{t+1}|\underline{x}_t)p(\underline{x}_t)A(\underline{x}_{t+1},\underline{x}_t) = \min\{\tilde{q}(\underline{x}_{t+1}|\underline{x}_t)p(\underline{x}_t), \tilde{q}(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1})\} \tag{1}$$

$$q(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1}) = \tilde{q}(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1})A(\underline{x}_t,\underline{x}_{t+1}) = \min\{\tilde{q}(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1}), \tilde{q}(\underline{x}_{t+1}|\underline{x}_t)p(\underline{x}_t)\} \tag{2}$$

Therefore, $q(\underline{x}_{t+1}|\underline{x}_t)p(\underline{x}_t) = q(\underline{x}_t|\underline{x}_{t+1})p(\underline{x}_{t+1})$ i.e. detailed balance holds.

2) If $\underline{x}'$ and $\underline{x}$ differ at more than the chosen coordinate i, then $\frac{\tilde{q}(\underline{x}|\underline{x}')}{\tilde{q}(\underline{x}'|\underline{x})}$ is not defined, so $A(\underline{x}',\underline{x}) = 1$.

If they differ at at position i, $\tilde{q}(\underline{x}'|\underline{x}) = p(x'_i|\{x_j\}_{j\neq i})$ and $\tilde{q}(\underline{x}|\underline{x}') = p(x_i|\{x\}_{j\neq i})$, so $\frac{\tilde{q}(\underline{x}|\underline{x}')p(\underline{x}')}{\tilde{q}(\underline{x}'|\underline{x})p(\underline{x})} = \frac{p(x_i|\{x_j\}_{j\neq i})p(\underline{x}')}{p(x'_i|\{x_j\}_{j\neq i})p(\underline{x})} = \frac{p(x_i|\{x_j\}_{j\neq i})p(x'_i|\{x_j\}_{j\neq i})p(\{x_j\}_{j\neq i})}{p(x'_i|\{x_j\}_{j\neq i})p(x_i|\{x_j\}_{j\neq i})p(\{x_j\}_{j\neq i})} = 1$.

In conclusion, Gibbs sampling yields an acceptance probability $A(\underline{x}',\underline{x}) = 1$.

3) For this distribution , given states $\underline{s}$ and $\underline{s}'$ which differ at most at i, we have that we transition from $\underline{s}$ to $\underline{s}'$ with probability $\tilde{q}(\underline{s}'|\underline{s}) = p(s'_i|\{s_j\}_{j\neq i}) = \frac{p(\underline{s}')}{\sum\limits_{s_i} p(\{s_j\})}$. But,

$$\sum_{s_i} p(\{s_j\}) = \frac{1}{Z}[\exp\{\sum_{\substack{(k,l)\in E\\ k,l\neq i}} J_{kl}s_k s_l + \sum_{(k,i)\in E} J_{ki}s_k + \sum_{(i,l)\in E} J_{il}s_l + \sum_{\substack{k\in V\\ k\neq i}} h_k s_k + h_i\} + \exp\{\sum_{\substack{(k,l)\in E\\ k,l\neq i}} J_{kl}s_k s_l - \sum_{(k,i)\in E} J_{ki}s_k - \sum_{(i,l)\in E} J_{il}s_l + \sum_{\substack{k\in V\\ k\neq i}} h_k s_k - h_i\}] \tag{3}$$

Therefore, $\tilde{q}(\underline{s}'|\underline{s}) = \frac{\exp\{\sum\limits_{(k,i)\in E} J_{ki}s_k s'_i + \sum\limits_{(i,l)\in E} J_{il}s'_i s_l + h_i s'_i\}}{\exp\{\sum\limits_{(k,i)\in E} J_{ki}s_k + \sum\limits_{(i,l)\in E} J_{il}s_l + h_i\} + \exp\{-\sum\limits_{(k,i)\in E} J_{ki}s_k - \sum\limits_{(i,l)\in E} J_{il}s_l - h_i\}}$ and this is just $\frac{1}{Z}[1 \pm tanh(\sum\limits_{(k,i)\in E} J_{ki}s_k + \sum\limits_{(i,l)\in E} J_{il}s_l + h_i)]$ for $s'_i = \pm 1$. The term $\sum\limits_{(k,i)\in E} J_{ki}s_k + \sum\limits_{(i,l)\in E} J_{il}s_l = \sum\limits_{(i,j)\in E} J_{ij}s_j$, as $(k,i)\in E$ and $(i,l)\in E$ form the Markov blanket of i through vertices k and l.

2 Exercise 2

The KL-divergence between p and q is:

$$KL(p||q) = \mathbb{E}_p[\log p] - \mathbb{E}_p[\log q] = \mathbb{E}_p[x^T W x - logZ_p(W)] - \mathbb{E}_p[x^T U x - logZ_q(U)] \Rightarrow$$
$$\Rightarrow \arg\min_U KL(p||q) = \arg\max_U \{\mathbb{E}_p[x^T U x] - \log Z_q(U)\} \quad (4)$$

We have that:

$$x^T U x = \begin{bmatrix} x_1 & x_2 & \dots & x_D \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} & \dots & U_{1D} \\ U_{21} & U_{22} & \dots & U_{2D} \\ \vdots & \vdots & \dots & \vdots \\ U_{n1} & U_{n2} & \dots & U_{DD} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \sum_{i=1}^{D} \sum_{j=1}^{D} U_{ij} x_i x_j \quad (5)$$

$$Tr(UC) = \sum_{i=1}^{D} U_{1i} \mathbb{E}_p[x_i x_1] + \dots + \sum_{i=1}^{D} U_{Di} \mathbb{E}_p[x_i x_D] = \sum_{i=1}^{D} \sum_{j=1}^{D} U_{ij} \mathbb{E}_p[x_i x_j] \quad (6)$$

From this we get that $\mathbb{E}_p[x^T U x] = Tr(UC)$.
Therefore $KL(p||q) = \arg\max_U \{Tr(UC) - \log Z_q(U)\}$.

3 Exercise 3

Note: When using the Naive Bayes Classifier, we should actually compare $p(\underline{x}^*, \text{class} = 0)$ and $p(\underline{x}^*, \text{class} = 1)$ to establish the class that the sample was extracted from. This is because $p(\underline{x}^* | \text{class} = 0)$ and $p(\underline{x}^* | \text{class} = 1)$ would only consider how well the sample would fit within the class, without accounting for how frequent the class itself is.

The inequality $p(\underline{x}^*, \text{class} = 0) > p(\underline{x}^*, \text{class} = 1)$ is equivalent to $\log p(\underline{x}^*, \text{class} = 0) > \log p(\underline{x}^*, \text{class} = 1)$. Expanding the probabilities, we get:

$$\log p_1 + \sum_{i=1}^{K} \log p(x_i^* | \text{class} = 1) > \log p_0 + \sum_{i=1}^{K} \log p(x_i^* | \text{class} = 0) \quad (7)$$

Using the notations in the statement, we have that:

$$\log p(x_i^* | \text{class} = 1) = x_i^* \log \theta_i^1 + (1 - x_i^*) \log(1 - \theta_i^1) = x_i^* \log \frac{\theta_i^1}{1 - \theta_i^1} + \log(1 - \theta_i^1)$$
$$\log p(x_i^* | \text{class} = 0) = x_i^* \log \theta_i^0 + (1 - x_i^*) \log(1 - \theta_i^0) = x_i^* \log \frac{\theta_i^0}{1 - \theta_i^0} + \log(1 - \theta_i^0) \quad (8)$$

Therefore, we get:

$$\log p_1+\sum_{i=1}^{K}\log(1-\theta_i^1)+\sum_{i=1}^{K}x_i^*\log\frac{\theta_i^1}{1-\theta_i^1} > \log p_0+\sum_{i=1}^{K}\log(1-\theta_i^0)+\sum_{i=1}^{K}x_i^*\log\frac{\theta_i^0}{1-\theta_i^0} \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{K}x_i^*\log\frac{\theta_i^1(1-\theta_i^0)}{\theta_i^0(1-\theta_i^1)}+\log\frac{p_1}{p_0}+\sum_{i=1}^{K}\log\frac{1-\theta_i^1}{1-\theta_i^0} > 0 \quad (9)$$

We therefore see that we can write the condition of classifying $\underline{x}^*$ in class 1 as $w^T\underline{x}^* + b > 0$, where:

$$w = \left[\log\tfrac{\theta_1^1(1-\theta_1^0)}{\theta_1^0(1-\theta_1^1)} \quad \dots \quad \log\tfrac{\theta_K^1(1-\theta_K^0)}{\theta_K^0(1-\theta_K^1)}\right],\ b = \log\frac{p_1}{p_0}+\sum_{i=1}^{K}\log\frac{1-\theta_i^1}{1-\theta_i^0} \qquad (10)$$

4 Exercise 4

1) The Bayesian network looks as follows:

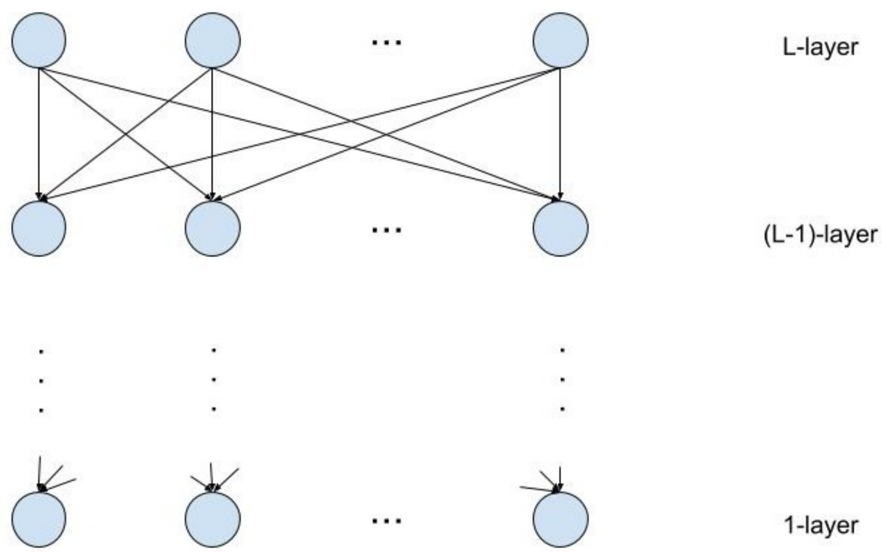

Figure 1: Sigmoid Belief Network

2) Computing $p(\mathbf{x}^0)$ implies computing $\sum_{\mathbf{x}^1}...\sum_{\mathbf{x}^L}p(\mathbf{x}^0,\mathbf{x}^1,...,\mathbf{x}^L)$ for a fixed value of $\mathbf{x}^0$. The complexity of computing a probability $p(x_i^{l-1}|\mathbf{x}^l)$ is $O(w)$ and we have w such probabilities to compute per layer. This gives a complexity of $O(w^2)$ per layer of computing $p(\mathbf{x}^{l-1}|\mathbf{x}^l)$. Due to the structure of the Bayesian

Network, we can decompose our computation into:

$$\sum_{\mathbf{x}^1} ... \sum_{\mathbf{x}^L} p(\mathbf{x}^0, \mathbf{x}^1, ..., \mathbf{x}^L) = \sum_{\mathbf{x}^1} p(\mathbf{x}^0|\mathbf{x}^1) \sum_{\mathbf{x}^2} p(\mathbf{x}^1|\mathbf{x}^2) ... \sum_{\mathbf{x}^L} p(\mathbf{x}^{L-1}|\mathbf{x}^L) p(\mathbf{x}^L) \tag{11}$$

Each $\mathbf{x}^i$ takes 2^w values. To compute $f_{L-1}(\mathbf{x}^{L-1}) = \sum_{\mathbf{x}^L} p(\mathbf{x}^{L-1}|\mathbf{x}^L)p(\mathbf{x}^L)$ for all values of $\mathbf{x}^{L-1}$ requires a complexity of $O(w^2 2^{2w})$. The same applies then for computing $f_{L-2} = \sum_{\mathbf{x}^{L-1}} p(\mathbf{x}^{L-2}|\mathbf{x}^{L-1}) f_{L-1}(\mathbf{x}^{L-1})$ and so on.
The overall complexity is therefore $O(Lw^2 2^{2w})$.
3) The energy term for the Variational EM procedure is:

$$\mathbb{E}_q[\log p(\mathbf{x}^1, ..., \mathbf{x}^L, \mathbf{x}^0)] = \sum_{\mathbf{x}^1,...,\mathbf{x}^L} \prod_{l=1}^{L} \prod_{i=1}^{w} q(x_i^l) \log p(\mathbf{x}^1, ..., \mathbf{x}^L, \mathbf{x}^0) \tag{12}$$

In this case, even if splitting $\log p(\mathbf{x}^1, ..., \mathbf{x}^L, \mathbf{x}^0)$ into its constituents, we can not find a smart factorization, so we need to compute every term in the sum separately. There are 2^{Lw} terms. $\log p(\mathbf{x}^1, ..., \mathbf{x}^L, \mathbf{x}^0)$ takes $O(Lw^2)$ operations to compute and $\prod_{l=1}^{L} \prod_{i=1}^{w} q(x_i^l)$ takes $O(Lw)$. Therefore, the total complexity is $O(Lw^2 2^{Lw})$.

5 Exercise 5

Our objective is to maximize $\sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}) = \sum_{n=1}^{N} \log \sum_{k=1}^{H} p(\mathbf{x}^{(n)}|\mathbf{m}_k, \sigma_k^2)p(k)$, where we have considered the training set $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$.
Given the probability distributions $\{\{q_{nk}^{(t)}\}_{k=1}^{H}\}_{n=1}^{N}$ at step t in the EM algorithm, by convexity we have $\sum_{n=1}^{N} \log \sum_{k=1}^{H} p(\mathbf{x}^{(n)}|\mathbf{m}_k, \sigma_k^2)p(k) \geq \sum_{n=1}^{N} \sum_{k=1}^{H} q_{nk}^{(t)} \log \frac{p(k)p(\mathbf{x}^{(n)}|\mathbf{m}_k, \sigma_k^2)}{q_{nk}^{(t)}}$ and we get equality for $q_{nk}^{(t)} = \frac{p(k)p(\mathbf{x}^{(n)}|\mathbf{m}_k^{(t)}, (\sigma_k^2)^{(t)})}{\sum_{k=1}^{H} p(k)p(\mathbf{x}^{(n)}|\mathbf{m}_k^{(t)}, (\sigma_k^2)^{(t)})}$. This is the expectation step in the EM algorithm.
We now fix $q_{nk}^{(t)}$ and optimize with respect to $p(k)$, $\mathbf{m}_k$ and σ_k^2, which means that we optimize $\sum_{n=1}^{N} \sum_{k=1}^{H} q_{nk}^{(t)} [\log p(k) - \log q_{nk}^{(t)} + \log p(\mathbf{x}^{(t)}|\mathbf{m}_k, \sigma_k^2)]$. Optimizing with respect to $\mathbf{m}_k$ and σ_k^2 implies maximizing $\sum_{n=1}^{N} \sum_{k=1}^{H} q_{nk}^{(t)} \log p(\mathbf{x}^{(n)}|\mathbf{m}_k, \sigma_k^2) = \sum_{n=1}^{N} \sum_{k=1}^{H} q_{nk}^{(t)} [-\frac{D}{2} \log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}(\mathbf{x}^{(n)} - \mathbf{m}_k)^T(\mathbf{x}^{(n)} - \mathbf{m}_k)]$.

Differentiating with respect to $\mathbf{m}_k$ gives:

$$\sum_{n=1}^{N} q_{nk}^{(t)} \frac{1}{\sigma_k^2}(\mathbf{x}^{(n)} - \mathbf{m}_k) = 0 \Rightarrow \mathbf{m}_k^{(t+1)} = \frac{\sum\limits_{n=1}^{N} q_{nk}^{(t)} \mathbf{x}^{(n)}}{\sum\limits_{n=1}^{N} q_{nk}^{(t)}} \tag{13}$$

Differentiating with respect to σ_k^2 yields:

$$\sum_{n=1}^{N} q_{nk}^{(t)} [-\frac{D}{2\sigma_k^2} + \frac{1}{2\sigma_k^2}(\mathbf{x}^{(n)} - \mathbf{m}_k)^T(\mathbf{x}^{(n)} - \mathbf{m}_k)] \Rightarrow (\sigma_k^2)^{(t+1)} = \frac{\sum\limits_{n=1}^{N} q_{nk}^{(t)} (\mathbf{x}^{(n)} - \mathbf{m}_k^{(t)})^T(\mathbf{x}^{(n)} - \mathbf{m}_k^{(t)})}{D \sum\limits_{n=1}^{N} q_{nk}^{(t)}} \tag{14}$$

Putting it all together, the update rule of $\mathbf{m}_k$ and σ_k^2 in the M-step of the EM algorithm is:

$$\mathbf{m}_k^{(t+1)} = \frac{\sum\limits_{n=1}^{N} q_{nk}^{(t)} \mathbf{x}^{(n)}}{\sum\limits_{n=1}^{N} q_{nk}^{(t)}}; \; (\sigma_k^2)^{(t+1)} = \frac{\sum\limits_{n=1}^{N} q_{nk}^{(t)} (\mathbf{x}^{(n)} - \mathbf{m}_k^{(t)})^T(\mathbf{x}^{(n)} - \mathbf{m}_k^{(t)})}{D \sum\limits_{n=1}^{N} q_{nk}^{(t)}} \tag{15}$$

6 Exercise 6

1) The log-likelihood is:

$$L(W) = \log p(\underline{v}^{(1)}, \underline{v}^{(2)}, ..., \underline{v}^{(N)}|W) = \sum_{n=1}^{N} \log p(\underline{v}^{(n)}|W) = \sum_{n=1}^{N} \log \sum_{\underline{h}} p(\underline{h}, \underline{v}^{(n)}|W) \tag{16}$$

We have that $p(\underline{h}, \underline{v}|W) = \frac{1}{Z} \exp\{\sum\limits_{i=1}^{K} \sum\limits_{j=1}^{M} W_{ij} v_i h_j\}$, with $Z = \sum\limits_{\underline{h},\underline{v}} W_{ij} v_i h_j$, so:

$$\frac{\partial}{\partial W_{ij}} L(W) = \sum_{n=1}^{N} \left(\frac{\partial}{\partial W_{ij}} \log \sum_{\underline{h}} \exp\{\sum_{i=1}^{K} \sum_{j=1}^{M} W_{ij} v_i^{(n)} h_j\} - \frac{\partial}{\partial W_{ij}} \log Z \right) \tag{17}$$

We now compute each of the partial derivative terms:

$$\frac{\partial}{\partial W_{ij}} \log Z = \frac{1}{Z} \sum_{\underline{h},\underline{v}} v_i h_j \exp\{\sum_{i=1}^{K} \sum_{j=1}^{M} W_{ij} v_i h_j\} = \sum_{\underline{h},\underline{v}} v_i h_j p(\underline{h}, \underline{v}|W) = \langle v_i h_j \rangle \tag{18}$$

$$\frac{\partial}{\partial W_{ij}} \log \sum_{\underline{h}} \exp\{\sum_{k=1}^{K}\sum_{l=1}^{M} W_{kl} v_k^{(n)} h_l\} = \frac{1}{\sum\limits_{\underline{h}} \exp\{\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{M} W_{kl} v_k^{(n)} h_l\}} \sum_{\underline{h}} v_i^{(n)} h_j \exp\{\sum_{k=1}^{K}\sum_{l=1}^{M} W_{kl} v_k^{(n)} h_l\} =$$

$$= \frac{1}{\sum\limits_{h_j} \exp\{\sum\limits_{k=1}^{K} W_{kj} v_k^{(n)} h_j\} \sum\limits_{\substack{h_l \\ l\neq j}} \exp\{\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{M} W_{kl} v_k^{(n)} h_l\}} \sum_{h_j} v_i^{(n)} h_j \exp\{\sum_{k=1}^{K} W_{kj} v_k^{(n)} h_j\} \sum_{\substack{h_l \\ l\neq j}} \exp\{\sum_{k=1}^{K}\sum_{l=1}^{M} W_{kl} v_k^{(n)} h_l\}$$

$$= \sum_{h_j} v_i^{(n)} h_j p(h_j|\underline{v}^{(n)}, W) = \mathbb{E}_{p(h_j|\underline{v}^{(n)},W)}[v_i^{(n)} h_j] \quad (19)$$

In conclusion,

$$\frac{\partial}{\partial W_{ij}} L(W) = \sum_{n=1}^{N} \mathbb{E}_{p(h_j|\underline{v}^{(n)},W)}[v_i^{(n)} h_j] - \langle v_i h_j \rangle \quad (20)$$

2) We have that:

$$p(h_j, \underline{v}, W) = \frac{1}{Z} \sum_{\substack{h_k \\ k\neq j}} \exp\{\sum_{i=1}^{K} W_{ij} v_i h_j + \sum_{\substack{k=1 \\ k\neq j}}^{M} \sum_{i=1}^{K} W_{ik} v_i h_k\} \Rightarrow$$

$$\Rightarrow p(h_j|\underline{v}, W) = \frac{p(h_j, \underline{v}|W)}{\sum\limits_{\underline{h}} p(h_j, \underline{v}|W)} = \frac{\exp\{\sum\limits_{i=1}^{K} W_{ij} v_i h_j\}}{\sum\limits_{h_k \in \{-1,1\}} \exp\{\sum\limits_{i=1}^{K} W_{ik} v_i h_k\}} \quad (21)$$

Therefore, we get:

$$\mathbb{E}_{p(h_j|v_i^{(n)},W)}[v_i^{(n)} h_j] = v_i^{(n)}[p(h_j = 1|v_i^{(n)}, W) - p(h_j = -1|v_i^{(n)}, W)] =$$

$$= v_i^{(n)} \frac{\exp\{\sum\limits_{k=1}^{K} W_{kj} v_k^{(n)}\} - \exp\{\sum\limits_{k=1}^{K} -W_{kj} v_k^{(n)}\}}{\exp\{\sum\limits_{k=1}^{K} W_{kj} v_k^{(n)}\} + \exp\{\sum\limits_{k=1}^{K} -W_{kj} v_k^{(n)}\}} = v_i^{(n)} \tanh(\sum_{i=1}^{K} W_{kj} v_k^{(n)}) \quad (22)$$

Finally, using this, we get the desired result:

$$\frac{\partial}{\partial W_{ij}} L(W) = \sum_{n=1}^{N} \left(v_i^{(n)} \tanh(\sum_{k=1}^{K} W_{kj} v_k^{(n)}) - \langle v_i h_j \rangle \right) \quad (23)$$