CS 726: Practice Questions on Learning Potentials

1. Consider an undirected graphical model G used to model $\Pr(x_1, \ldots, x_n)$ with only a single potential over each edge $(i,j) \in G$ as $\psi(x_i, x_j) = \sigma$ if $x_i = x_j$, $\psi(x_i, x_j) = 1$ otherwise. Thus, $\Pr(x_1, \ldots, x_n|\sigma) = \frac{1}{Z}\prod_{(i,j)\in G}\psi(x_i, x_j)$
Assume each x_j takes values from $1 \ldots m$. Let the training data consist of a single fully labeled graph, that is, $D = \{\mathbf{x}^1\}$.
    (a) Assume, $\mathbf{x}^1 = [0\ 0\ 1\ 0]$ and G a chain graph x_1—x_2—x_3—x_4, and $m = 2$. Write the value of $\Pr(\mathbf{x}^1|\sigma)$ purely in terms of σ, that is, even Z should be written in terms of σ. ..2
    $\Pr(\mathbf{x} = x_1, \ldots, x_n) = \frac{\sigma^{n_s(\mathbf{x})}}{Z(\sigma)}$ where $n_s(\mathbf{x})$ is the number of adjacent vertices in $\mathbf{x}$ that have the same label.
    $\Pr(\mathbf{x}^1|\sigma) = \frac{\sigma^1}{Z}$
    For Z, we go over all 16 possible ways of labeling $\mathbf{x}$ and count for each value of n_s, the count $c(n_s)$ of labelings $\mathbf{x}$ that will have that many adjacent variables with same labels.
    This comes to $\sum_{n_s=0}^{3}\sigma^{n_s}c(n_s) = 2 + \sigma * 6 + \sigma^2 * 6 + \sigma^3 * 2$
    Thus, we have $\Pr(\mathbf{x}^1|\sigma) = \frac{\sigma^1}{2+\sigma*6+\sigma^2*6+\sigma^3*2}$
    (b) Write the gradient of the training objective wrt σ in as simplified a form as possible. [The gradient should be for general graphs, and not just for the example graph in part (a) above.] ..2
    The loglikelihood of the training data
    $LL(D|\sigma) = n_s(\mathbf{x}^1)\log\sigma - \log Z(\sigma)$
    Its gradient wrt σ is
    $\frac{n_s(\mathbf{x}^1)}{\sigma} - \sum_{(i,j)\in E}(\sum_\ell \Pr(x_i = \ell, x_j = \ell))$
    (c) Solve for σ in closed form in terms of properties of D for the case when G is a tree? ..5 In this case since we have no node potentials and only the given edge potential, the message that any node i sends to a node j is uniform. In a tree, the marginal probability of any edge is equal to $\psi_{ij}(x_i, x_j)m_{i\to j}(x_i)m_{j\to i}(x_j)$. This implies that: $\sum_\ell \Pr(x_i = \ell, x_j = \ell) = m\frac{\sigma}{m\sigma+(m-1)m}$. Thus, we solve for $n_s/\sigma - E\frac{\sigma}{\sigma+(m-1)} = 0$ to get the value of σ.
    (d) Now assume that we have a training dataset D with partially observed set of variables with $n = 3, m = 2$, and G a complete graph (a triangle since $n = 3$.). Let $D = \{(x_1^1,\ x_2^1) = (1,\ 1), (x_2^2,\ x_3^2) = (0,\ 1)\}$, that is, the first instance has variable x_3 hidden and second instance has x_1 hidden. We will use the EM algorithm to solve for σ. Assume at some time t, $\sigma_t = 2$. For the next iteration, work out the E and M steps. Solve for the optimal value of σ in the M step.
        i. E-step. ..3 $\Pr(x_3^1 = 1|(x_1^1,\ x_2^1) = (1,\ 1), \sigma_t) = \frac{\sigma_t^2}{\sigma_t^2+1} = 4/(4+1) = 4/5$.
        $\Pr(x_1^2 = 1|(x_2^2,\ x_3^2) = (0,\ 1), \sigma_t) = \frac{\sigma_t}{\sigma_t+\sigma_t} = 1/2$.
        ii. M-step. ..3 Z for this problem is $6\sigma + 2\sigma^3$.
        The M step becomes:
        $\max_\sigma(4/5\log\sigma^3 + 1/5\log\sigma + 1/2\log\sigma + 1/2\log\sigma - 2\log(6\sigma + 2\sigma^3))$

2. Consider a $n \times n$ grid graph $G = (V, E)$ where V are vertices and E are edges of G. Each node $k \in V$ is a binary random variable y_k which takes value 1 or 0 depending on whether it is part of foreground or background. Each node is attached with a x_k that is a real-value denoting its propensity to be foreground. There are only three features in this UGM

$$\begin{aligned} f_1((y_k), (k), \mathbf{x}) &= x_k y_k \\ f_2((y_k), (k), \mathbf{x}) &= y_k \\ f_3((y_k, y_j), (k, j), \mathbf{x}) &= y_k y_j + (1 - y_k)(1 - y_j) \text{ if } (k, j) \in E, 0 \ \text{ otherwise.} \end{aligned} \tag{1}$$

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ denote the corresponding weights of these three features $\mathbf{f} = [f_1, f_2, f_3]$.

Also, consider an instance $(\mathbf{x}^i, \mathbf{y}^i)$ for a 3×3 grid for which the value of features x_k^i and correct label y_k^i are as given as:

$x_1 = 0.0, y_1 = 0$	$x_2 = 1.5, y_2 = 1$	$x_3 = 1.0, y_3 = 0$
$x_4 = 1.4, y_4 = 1$	$x_5 = 2.6, y_5 = 1$	$x_6 = 1.0, y_6 = 1$
$x_7 = 0.5, y_7 = 0$	$x_8 = 2.0, y_8 = 1$	$x_9 = 0.0, y_9 = 0$

(a) Write the expression for $\Pr(\mathbf{y}|\mathbf{x})$ in terms of $\theta_1, \theta_2, \theta_3, x_k, y_k$ for $k \in V$ [Do not use $f_k()$s but their defined values above. E.g. use $x_k y_k$ in place of $f_1()$ etc.] ..1

$\frac{1}{Z(\mathbf{x}^i)} \prod_{k \in V} (\exp(\theta_1 x_k y_k + \theta_2 y_k)) \prod_{(k,j) \in E} \exp(\theta_3(y_k y_j + (1 - y_k)(1 - y_j))$

(b) Compute the value of the normalizer $Z(\mathbf{x}^i)$ at $[\theta_1^t, \theta_2^t, \theta_3^t] = [0, 0, 0]$ ..2

Since all the θs are zero, we have that for all $\mathbf{y}$ the $\theta.\mathbf{f}$ term is zero, that is numerator above is 1. Thus, $Z(\mathbf{x}^i)$ = number of $\mathbf{y}$ combinations possible which is 2^9

(c) Compute the gradient of $\log \Pr(\mathbf{y}^i|\mathbf{x}^i, \theta^t)$ wrt θ_1 at $[\theta_1^t, \theta_2^t, \theta_3^t] = [0, 0, 0]$ ..2

Since all $\mathbf{y}$-s are equally likely, the marginal probability for each y_k is the same at $1/2$. Thus, the gradient: $f_1(\mathbf{y}^i, \mathbf{x}^i) - E_{\Pr(\mathbf{y}|\mathbf{x}^i, \theta^t)}[f_1(\mathbf{y}^i, \mathbf{x}^i)]$ can be easily computed as. $\sum_{k \in V}[x_k^i y_k^i - 1/2(x_k^i)] = \sum_{k \in V} x_k^i y_k^i - x_k^i/2)$

3. Consider the problem of training the parameters of a simple HMM of length two where the state and observation variables are binary. Thus, we have two state variables y_1 and y_2 and two observation variables x_1 and x_2 and all four variables can take one of two possible values. The parameters of the HMM are $\Pr(y_1)$ $\Pr(y_2|y_1)$ and $\Pr(x_1|y_1)$ and $\Pr(x_2|y_2)$. Assume $\Pr(x_1|y_1) = \Pr(x_2|y_2) = \Pr(x_t|y_t)$ We use the EM algorithm for training the parameters.

Let the initial values at $t = 0$ be

$$\Pr^t(y_1 = 0) = \theta_0^t = 0.5$$
$$\Pr^t(y_2 = 0|y_1 = 0) = \theta_1^t = 0.7, \quad \Pr^t(y_2 = 0|y_1 = 1) = \theta_2^t = 0.2$$
$$\Pr^t(x_t = 0|y_t = 0) = \theta_3^t = 0.1, \quad \Pr^t(x_t = 0|y_t = 1) = \theta_4^t = 0.8.$$

For a dataset D consisting of these two sequences $\mathbf{x}^1 = [0, 1]$, $\mathbf{x}^2 = [1, 1]$.

(a) E-step: Estimate the values of $\Pr(y_1|\mathbf{x}^1, \theta^t)$ ..3

For the E-step

The node potentials at y_1, call them $\psi(y_1) = \Pr(y_1)\Pr(x_1^i|y_1)$

The node potentials at y_2, call them $\psi(y_2) = \Pr(x_2^i|y_2)$

The edge potential $\psi(y_1, y_2) = \Pr(y_2|y_1)$.

Using this we get that for $\mathbf{x}^1$, $\psi(y_1) = [0.1, 0.8]$, $\psi(y_2) = [0.9, 0.2]$,

The message from y_2 to y_1 $= \sum_{y_2} \psi(y_2)\Pr(y_2|y_1) = [0.9 * 0.7 + 0.2 * 0.3, 0.9 * 0.2 + 0.2 * 0.8]$

Multiplying this message with $\psi(y_1)$ gives us $\Pr(y_1|\mathbf{x}^1)$.

(b) M-step: In the M-step write the formula for the maximum likelihood estimate of θ_0 in terms of $\Pr(y_1|\mathbf{x}^1, \theta^t)$ and $\Pr(y_1|\mathbf{x}^2, \theta^t)$. ..1

This will just be $(\Pr(y_1 = 0|\mathbf{x}^1, \theta^t) + \Pr(y_1 = 0|\mathbf{x}^2, \theta^t))/2$