CS446: Machine Learning **Spring 2017**

Problem Set 7

Bangqi Wang *Handed In: May 1, 2017*

1. Answer to problem 1

   a.

$$
\begin{aligned}
P(x^{(j)}) &= \sum_{z\in\{1,2\}} \prod_{i=0}^{n} P(x_i^{(j)}|z) && (1)\\
&= \prod_{i=0}^{n} P(x_i^{(j)}|z=1)P(z=1) + \prod_{i=0}^{n} P(x_i^{(j)}|z=2)P(z=2) && (2)\\
P(x^{(j)}) &= \alpha \prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}} && (3)
\end{aligned}
$$

   b. By Bayes Theorem,

$$
\begin{aligned}
f_z^{(j)} &= P(Z=z|x^{(j)}) = \frac{P(Z=z,x^{(j)})}{P(x^{(j)})} = \frac{P(x^{(j)}|Z=z)P(Z=z)}{P(x^{(j)})} && (4)\\
f_1^{(j)} &= \frac{\alpha \prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}}}{\alpha \prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}} && (5)\\
f_2^{(j)} &= \frac{(1-\alpha) \prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}}{\alpha \prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}} && (6)\\
& && (7)
\end{aligned}
$$

   c.

$$
\begin{aligned}
L &= \prod_{j=1}^{m} P(x^{(j)}|p,q,\alpha) && (8)\\
LL &= \sum_{j=1}^{m} logP(x^{(j)}|p,q,\alpha) && (9)\\
& && (10)
\end{aligned}
$$

$$\begin{aligned}
E[LL] &= E[\sum_{j=1}^{m} logP(x^{(j)}|p,q,\alpha)] = \sum_{j=1}^{m} E[logP(x^{(j)}|p,q,\alpha)] && (11)\\
&= \sum_{j=1}^{m}\sum_{z=1}^{2} f_z^{(j)} logP(Z=z, x^{(j)}|\tilde{p},\tilde{q},\tilde{\alpha}) - \sum_{j=1}^{m}\sum_{z=1}^{2} f_z^{(j)} log f_z^{(j)} && (12)\\
&= \sum_{j=1}^{m} f_1^{(j)} log(\alpha \prod_{i=1}^{n} \tilde{p}_i^{x_i^{(j)}} (1-\tilde{p})^{1-x_i^{(j)}}) && (13)\\
&\quad + f_2^{(j)} log((1-\alpha)\prod_{i=1}^{n} \tilde{q}_i)^{x_i^{(j)}} (1-\tilde{q}^{1-x_i^{(j)}}) && (14)\\
&\quad - \sum_{j=1}^{m} (f_1^{(j)} log f_1^{(j)} + f_2^{(j)} log f_2^{(j)}) && (15)
\end{aligned}$$

d. The expected log likelihood $E[LL]$ will be maximized when the derivative is equal to 0.

$$\begin{aligned}
\frac{\partial E}{\partial \tilde{\alpha}} &= \sum_{j=1}^{m} \frac{f_1^{(j)}}{\tilde{\alpha}} - \frac{f_2^{(j)}}{1-\tilde{\alpha}} = 0 && (16)\\
\Rightarrow \tilde{\alpha} &= \frac{\sum_{j=1}^{m} f_1^{(j)}}{m} && (17)\\
\frac{\partial E}{\partial \tilde{p}_i} &= \sum_{j=1}^{m} \frac{f_1^{(j)} x_i}{\tilde{p}_i} - \frac{f_1^{(j)}(1-x_i)}{1-\tilde{p}_i} = 0 && (18)\\
\Rightarrow \tilde{p}_i &= \frac{\sum_{j=1}^{m} f_1^{(j)} x_i^{(j)}}{\sum_{j=1}^{m} f_1^{(j)}} && (19)\\
\frac{\partial E}{\partial \tilde{q}_i} &= \sum_{j=1}^{m} \frac{f_2^{(j)} x_i}{\tilde{q}_i} - \frac{f_2^{(j)}(1-x_i)}{1-\tilde{q}_i} = 0 && (20)\\
\Rightarrow \tilde{q}_i &= \frac{\sum_{j=1}^{m} (1-f_2^{(j)}) x_i^{(j)}}{\sum_{j=1}^{m} (1-f_2^{(j)})} && (21)
\end{aligned}$$

e. $\tilde{\alpha}$ is the estimated probability of generating a sample with $z = 1$.
$\tilde{p}$ is the estimated probability of getting $x_i = 1$ given that $z = 1$.
$\tilde{q}$ is the estimated probability of getting $x_i = 1$ given that $z = 2$.

Algorithm 1 Pseudocode:

procedure MyProcedure
  Initialize parameters p, q, α with random values as estimations $\tilde{p}$, $\tilde{q}$, $\tilde{\alpha}$.
  Calculate the posterior distribution $f_1^{(j)}$ and $f_2^{(j)}$, with the equation from part (b).
  Use the update rules in part (d) to update estimations $\tilde{p}$, $\tilde{q}$, $\tilde{\alpha}$.
  Repeat step ii. and iii. until the estimations $\tilde{p}$, $\tilde{q}$, $\tilde{\alpha}$ converge.
end procedure

f. The algorithm will predict x_0 as 1 if $\frac{P(X_0=1)}{P(X_0=0)} > 1$, otherwise, the algorithm will predict x_1 as 0. Therefore, $x_0 = sign(log\frac{P(X_0=1)}{P(X_0=0)})$.

$$\begin{aligned} P(X_0 = 0) \quad &= P(x_0|x_1, ..., x_n) && (22)\\ &= P(Z = 1|x_1, ..., x_n)P(X_0 = 0|Z = 1) && (23)\\ &\quad +P(Z = 2|x_1, ..., x_n)P(X_0 = 0|Z = 2) && (24)\\ &= f_1 p_0 + f_2 q_0 && (25)\\ P(X_0 = 1) \quad &= P(x_0|x_1, ..., x_n) && (26)\\ &= P(Z = 1|x_1, ..., x_n)P(X_0 = 1|Z = 1) && (27)\\ &\quad +P(Z = 2|x_1, ..., x_n)P(X_0 = 1|Z = 2) && (28)\\ &= f_1(1 - p_0) + f_2(1 - q_0) && (29) \end{aligned}$$

$$\begin{aligned} f_1 \quad &= \frac{\alpha \prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i}}{\alpha \prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}} && (30)\\ f_2 \quad &= \frac{(1-\alpha) \prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}}{\alpha \prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}} && (31) \end{aligned}$$

$$\begin{aligned} x_0 \quad &= sign(log\frac{P(X_0 = 1)}{P(X_0 = 0)}) && (32)\\ &= sign(log(\frac{f_1(1-p_0) + f_2(1-q_0)}{f_1 p_0 + f_2 q_0})) && (33)\\ &= sign(log(\frac{(1-p_0)\alpha \prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i} + (1-q_0)(1-\alpha)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}}{p_0\alpha \prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i} + q_0(1-\alpha)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}} && (34)\\ & && (35) \end{aligned}$$

g. According to the result from part (f), the decision surface for this prediction can transform to a linear function. After transforming,

$$
\begin{aligned}
x_0 &= sign(log(\frac{(1-2p_0)\alpha\prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i}}{(1-2q_0)(\alpha-1)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}})) && (36)\\
&= log((1-2p_0)\alpha\prod_{i=0}^{n} p_i^{x_i}(1-p_i)^{1-x_i}) - log((1-2q_0)(\alpha-1)\prod_{i=0}^{n} q_i^{x_i}(1-q_i)^{1-x_i}) && (37)\\
&= log(\frac{1-2p_0}{1-2q_0}) + log(\frac{\alpha}{\alpha-1}) + \sum_{i=0}^{n} x_i log(\frac{p_i}{q_i}) + \sum_{i=0}^{n}(1-x_i)log(\frac{1-p_i}{1-q_i}) && (38)
\end{aligned}
$$

2. Answer to problem 2

a. The statement means that the probabilities for every event E over variables $x_1, ..., x_n$ are equal for two directed trees T_0 and T_1. The joint probability distributions are the same. $P_{T_0}(x_1, ..., x_n) = P_{T_1}(x_1, ..., x_n)$.

b. Assume two directed trees T_i and T_j have different roots x_i and x_j from the undirected tree T. The resulting directred trees are all equivalent if $P_{T_0}(x) = P_{T_1}(x)$. So, $P(x_1|x_2)P(x_2) = P(x_2|x_1)P(x_1) = P(x_1, x_2)$.
First, we assume x_i and x_j are nodes in tree T, and they are connected by a path P with length 1.

$$
\begin{aligned}
P_{T_i}(x) &= P(x_i)\prod_{k\in\{N-i\}}^{n} P(x_k|Parent_{x_k}) && (39)\\
&= P(x_i)P(x_j|x_i)\prod_{k\in\{N-P\}} P(x_k|Parent_{x_k}) && (40)\\
&= P(x_i, x_j)\prod_{k\in\{N-P\}} P(x_k|Parent_{x_k}) && (41)\\
&= P(x_j)P(x_i|x_j)\prod_{k\in\{N-P\}} P(x_k|Parent_{x_k}) && (42)\\
&= P(x_j)\prod_{k\in\{N-j\}}^{n} P(x_k|Parent_{x_k}) && (43)\\
&= P_{T_j}(x) && (44)
\end{aligned}
$$

When the path P has length larger than 1, the hypothesis still holds. $P(x_{root})\prod_{k\in\{N-P\}}^{n} P(x_k|Parent_{x_k})$ remains the same. We can switch the edges between x_i ans x_j according the chain rule. By switching one edge at a time, the resulting directed trees are all equivalent.