CS446: Machine Learning **Spring 2017**

Problem Set 7

Kexin Hui *Handed In: May 1, 2017*

1. Answer to problem 1 - EM Algorithm

   a.

$$
\begin{aligned}
\Pr(x^{(j)}) &= \Pr(\sum_i x_i^{(j)}) = \prod_i \Pr(x_i^{(j)}) \\
&= \prod_i \sum_z \Pr(x_i^{(j)}, Z) = \prod_i \sum_z \Pr(x_i^{(j)}|Z)\Pr(Z) \\
&= \prod_i (\ \Pr(z=1)\Pr(x_i^{(j)}|Z=1) + \Pr(z=2)\Pr(x_i^{(j)}|Z=2)\ ) \\
&= \alpha \prod_i \Pr(x_i^{(j)}|Z=1) + (1-\alpha)\prod_i \Pr(xd_i^{(j)}|Z=2) \\
&= \alpha \prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}
\end{aligned}
\tag{1}
$$

   b.

$$
\begin{aligned}
f_z^{(j)} &= \Pr(Z=z|x^{(j)}) = \frac{\Pr(x^{(j)}|Z=z)\Pr(Z=z)}{\Pr(x^{(j)})} \\
f_1^{(j)} &= \Pr(Z=1|x^{(j)}) = \frac{\Pr(x^{(j)}|Z=1)\Pr(Z=1)}{\Pr(x^{(j)})} \\
&= \frac{\alpha\prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}}}{\alpha\prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}} \\
f_2^{(j)} &= \Pr(Z=2|x^{(j)}) = \frac{\Pr(x^{(j)}|Z=2)\Pr(Z=2)}{\Pr(x^{(j)})} \\
&= \frac{(1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}}{\alpha\prod_{i=0}^{n} p_i^{x_i^{(j)}}(1-p_i)^{1-x_i^{(j)}} + (1-\alpha)\prod_{i=0}^{n} q_i^{x_i^{(j)}}(1-q_i)^{1-x_i^{(j)}}}
\end{aligned}
\tag{2}
$$

c.

$$
\begin{aligned}
LL &= log(\Pr(D;\tilde{\alpha},\tilde{p_i},\tilde{q_i})) \\
&= \sum_{j=1}^{m} log \sum_{Z} \Pr(x^{(j)}, Z;\tilde{\alpha},\tilde{p_i},\tilde{q_i}) \\
&= \sum_{j=1}^{m} log \sum_{Z} \Pr(Z|x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) \Pr(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) \\
&= \sum_{j=1}^{m} log \sum_{Z} f_z^{(j)} \Pr(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) \\
&\geq \sum_{j=1}^{m} \sum_{Z} f_z^{(j)} log \Pr(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i})
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
E[LL] &= E[\sum_{j=1}^{m} \sum_{Z} f_z^{(j)} log \Pr(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i})] \\
&= \sum_{j=1}^{m} E[f_1^{(j)} log \Pr_{Z=1}(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) + f_2^{(j)} log \Pr_{Z=2}(x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i})] \\
&= \sum_{j=1}^{m} f_1^{(j)} log \Pr(Z=1, x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) + f_2^{(j)} log \Pr(Z=2, x^{(j)};\tilde{\alpha},\tilde{p_i},\tilde{q_i}) \\
&= \sum_{j=1}^{m} f_1^{(j)} log(\tilde{\alpha} \prod_{i=0}^{n} \tilde{p_i}^{x_i^{(j)}} (1-\tilde{p_i})^{1-x_i^{(j)}}) + f_2^{(j)} log((1-\tilde{\alpha}) \prod_{i=0}^{n} \tilde{q_i}^{x_i^{(j)}} (1-\tilde{q_i})^{1-x_i^{(j)}}) \\
&= \sum_{j=1}^{m} f_1^{(j)} (log\tilde{\alpha} + \sum_{i=0}^{n} \left[x_i^{(j)} log\tilde{p_i} + (1-x_i^{(j)}) log(1-\tilde{p_i}) \right]) \\
&\quad + f_2^{(j)} (log(1-\tilde{\alpha}) + \sum_{i=0}^{n} \left[x_i^{(j)} log\tilde{q_i} + (1-x_i^{(j)}) log(1-\tilde{q_i}) \right])
\end{aligned}
\tag{4}
$$

d.

$$
\frac{\partial E[LL]}{\partial \tilde{\alpha}} = \sum_{j=1}^{m} [\frac{f_1^{(j)}}{\tilde{\alpha}} - \frac{f_2^{(j)}}{1-\tilde{\alpha}}] = 0 \Rightarrow \tilde{\alpha} = \frac{\sum_{j=1}^{m} f_1^{(j)}}{\sum_{j=1}^{m} f_1^{(j)} + f_2^{(j)}} = \frac{\sum_{j=1}^{m} f_1^{(j)}}{m} \tag{5}
$$

$$
\frac{\partial E[LL]}{\partial \tilde{p_i}} = \sum_{j=1}^{m} f_1^{(j)} \times (\frac{x_i^{(j)}}{\tilde{p_i}} - \frac{1-x_i^{(j)}}{1-\tilde{p_i}}) = 0 \Rightarrow \tilde{p_i} = \frac{\sum_{j=1}^{m} f_1^{(j)} x_i^{(j)}}{\sum_{j=1}^{m} f_1^{(j)}} \tag{6}
$$

$$
\frac{\partial E[LL]}{\partial \tilde{q_i}} = \sum_{j=1}^{m} f_2^{(j)} \times (\frac{x_i^{(j)}}{\tilde{q_i}} - \frac{1-x_i^{(j)}}{1-\tilde{q_i}}) = 0 \Rightarrow \tilde{q_i} = \frac{\sum_{j=1}^{m} f_2^{(j)} x_i^{(j)}}{\sum_{j=1}^{m} f_2^{(j)}} \tag{7}
$$

e. The updated $\tilde{\alpha} = \frac{\sum_{j=1}^m f_1^{(j)}}{m}$ is the average probability of $Z = 1$ given all data.
The updated $\tilde{p_i} = \frac{\sum_{j=1}^m f_1^{(j)} x_i^{(j)}}{\sum_{j=1}^m f_1^{(j)}}$ is the weighted average of $x_i = 1$ for all data with $Z = 1$.
The updated $\tilde{q_i} = \frac{\sum_{j=1}^m f_2^{(j)} x_i^{(j)}}{\sum_{j=1}^m f_2^{(j)}}$ is the weighted average of $x_i = 1$ for all data with $Z = 2$.

Algorithm 1 pseudocode

Initialization: $\alpha^0 = 0.5$, $p_i^0 = 0.5$, $q_i^0 = 0.5$, $\forall i$
Iteration: While α^t, p_i^t, q_i^t do not converge:

$f_1^{(j)} = \frac{\alpha^t \prod_{i=0}^n p_i^{t\, x_i^{(j)}} (1-p_i^t)^{1-x_i^{(j)}}}{\alpha^t \prod_{i=0}^n p_i^{t\, x_i^{(j)}} (1-p_i^t)^{1-x_i^{(j)}} + (1-\alpha^t) \prod_{i=0}^n q_i^{t\, x_i^{(j)}} (1-q_i^t)^{1-x_i^{(j)}}}$

$f_2^{(j)} = 1 - f_1^{(j)}$

$\alpha^{t+1} = \frac{\sum_{j=1}^m f_1^{(j)}}{m}$

$p_i^{t+1} = \frac{\sum_{j=1}^m f_1^{(j)} x_i^{(j)}}{\sum_{j=1}^m f_1^{(j)}}$

$q_i^{t+1} = \frac{\sum_{j=1}^m f_2^{(j)} x_i^{(j)}}{\sum_{j=1}^m f_2^{(j)}}$

Termination: Converge when $error = \|\alpha^{t+1} - \alpha^t\| + \sum_i \left\|p_i^{t+1} - p_i^t\right\| + \sum_i \left\|q_i^{t+1} - q_i^t\right\| < threshold$

f.

$$
\begin{aligned}
&\frac{\Pr(x_0 = 0|x_1, x_2, ..., x_n)}{\Pr(x_0 = 1|x_1, x_2, ..., x_n)} = \frac{\sum_z \Pr(x_0 = 0, Z = z|x_1, x_2, ..., x_n)}{\sum_z \Pr(x_0 = 1, Z = z|x_1, x_2, ..., x_n)} \\
=&\frac{\sum_z \Pr(x_0 = 0|Z = z, x_1, x_2, ..., x_n) \Pr(Z = z|x_1, x_2, ..., x_n)}{\sum_z \Pr(x_0 = 1|Z = z, x_1, x_2, ..., x_n) \Pr(Z = z|x_1, x_2, ..., x_n)} \\
=&\frac{\sum_z \Pr(x_0 = 0|Z = z, x_1, x_2, ..., x_n) \Pr(Z = z|x_1, x_2, ..., x_n) \Pr(x_1, x_2, ..., x_n)}{\sum_z \Pr(x_0 = 1|Z = z, x_1, x_2, ..., x_n) \Pr(Z = z|x_1, x_2, ..., x_n) \Pr(x_1, x_2, ..., x_n)} \\
=&\frac{\sum_z \Pr(x_0 = 0|Z = z, x_1, x_2, ..., x_n) \Pr(x_1, x_2, ..., x_n|Z = z) \Pr(Z = z)}{\sum_z \Pr(x_0 = 1|Z = z, x_1, x_2, ..., x_n) \Pr(x_1, x_2, ..., x_n|Z = z) \Pr(Z = z)} \\
=&\frac{\alpha(1 - p_0) \prod_{i=1}^n p_i^{x_i}(1 - p_i)^{(1-x_i)} + (1 - \alpha)(1 - q_0) \prod_{i=1}^n q_i^{x_i}(1 - q_i)^{(1-x_i)}}{\alpha p_0 \prod_{i=1}^n p_i^{x_i}(1 - p_i)^{(1-x_i)} + (1 - \alpha)q_0 \prod_{i=1}^n q_i^{x_i}(1 - q_i)^{(1-x_i)}}
\end{aligned}
\tag{8}
$$

Thus, predict 0 iff
$\frac{\Pr(x_0=0|x_1,x_2,...,x_n)}{\Pr(x_0=1|x_1,x_2,...,x_n)} = \frac{\alpha(1-p_0) \prod_{i=1}^n p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)(1-q_0) \prod_{i=1}^n q_i^{x_i}(1-q_i)^{(1-x_i)}}{\alpha p_0 \prod_{i=1}^n p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)q_0 \prod_{i=1}^n q_i^{x_i}(1-q_i)^{(1-x_i)}} \geq 1.$

g.

$$\frac{\alpha(1-p_0)\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)(1-q_0)\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)}}{\alpha p_0\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)q_0\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)}} \geq 1$$

$$\begin{aligned}
&\Rightarrow \alpha(1-p_0)\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)(1-q_0)\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)} \\
&\qquad \geq \alpha p_0\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)} + (1-\alpha)q_0\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)} \\
&\Rightarrow \alpha(1-2p_0)\prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{(1-x_i)} \geq -(1-\alpha)(1-2q_0)\prod_{i=1}^{n} q_i^{x_i}(1-q_i)^{(1-x_i)} \\
&\Rightarrow log\alpha + log(1-2p_0) + \sum_{i=1}^{n}[x_i log(p_i) + (1-x_i)log(1-p_i)] \qquad (9) \\
&\qquad \geq log(\alpha-1) + log(1-2q_0) + \sum_{i=1}^{n}[x_i log(q_i) + (1-x_i)log(1-q_i)] \\
&\Rightarrow log\alpha(1-2p_0) + \sum_{i=1}^{n}[x_i log\frac{p_i}{1-p_i} + log(1-p_i)] \\
&\qquad \geq log(\alpha-1)(1-2q_0) + \sum_{i=1}^{n}[x_i log\frac{q_i}{1-q_i} + log(1-p_i)] \\
&\Rightarrow \sum_{i=1}^{n} x_i \times log\frac{p_i(1-q_i)}{q_i(1-p_i)} + log\frac{\alpha(1-2p_0)}{(\alpha-1)(1-2q_0)} + \sum_{i=1}^{n} log\frac{1-p_i}{1-q_i} \geq 0
\end{aligned}$$

It is clear that from the final inequality, the decision surface is a linear function which can be written as $\sum_i w_i x_i + b$.

2. Answer to problem 2 - Tree Dependent Distributions

   a. Suppose we have two directed trees T_1 and T_2 obtained from undirected tree T, T_1 and T_2 are equivalent $\Leftrightarrow \Pr_{T_1}(x_1, x_2, ..., x_n) = \Pr_{T_2}(x_1, x_2, ...x_n)$. That is, given any event E, the probability of E happening based on T_1 is always equal to the probability of E happening based on T_2.

   b. Let x_i and x_j be two root nodes, and T_i and T_j be the two corresponding directed tree, where $i \neq j$ and $1 \leq i, j \leq n$. We need to prove the equivalence $\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n)$.
   Let $parent(x_k)$ be the parent node of x_k, $\forall k$. It is true that there exists a unique path from x_i to x_j in the tree structure denoted by $path_{ij}$.
   Proof:
   We can easily prove $\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n)$ if the length of the path is 1, $\|path_{ij}\| = 1$.

$$\begin{aligned}
\Pr_{T_i}(x_1, x_2, ..., x_n) &= \Pr(x_i) \prod_{k=1,k\neq i}^{n} \Pr(x_k|parent(x_k)) \\
&= \Pr(x_i)\Pr(x_j|x_i) \prod_{k=1,k\neq i,k\neq j}^{n} \Pr(x_k|parent(x_k)) \\
&= \Pr(x_i, x_j) \prod_{k=1,k\neq i,k\neq j}^{n} \Pr(x_k|parent(x_k)) \\
&= \Pr(x_j)\Pr(x_j|x_i) \prod_{k=1,k\neq i,k\neq j}^{n} \Pr(x_k|parent(x_k)) \\
&= \Pr(x_j) \prod_{k=1,k\neq j}^{n} \Pr(x_k|parent(x_k)) \\
&= \Pr_{T_j}(x_1, x_2, ..., x_n)
\end{aligned} \tag{10}$$

Assume that the equivalence holds if the length of path from x_i to x_j is l, $\forall l > 1$, given by

$$\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n) \qquad where \;\; \|path_{ij}\| = l > 1 \tag{11}$$

When the length of path from x_i to x_j is $l+1$, there exists a node x_t that satisfies the length of path from x_i to x_t is l, and the length of path from x_t to x_j is 1. We already know that

$$\begin{aligned}
\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_t}(x_1, x_2, ...x_n) \qquad where \;\; \|path_{it}\| = l \\
\Pr_{T_t}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n) \qquad where \;\; \|path_{tj}\| = 1
\end{aligned} \tag{12}$$

Thus, by the transitive relation of the equation, the equivalence also holds for $\|path_{ij}\| = l+1$, that is

$$\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n) \qquad where \;\; \|path_{ij}\| = l+1 \tag{13}$$

To conclude, $\Pr_{T_i}(x_1, x_2, ..., x_n) = \Pr_{T_j}(x_1, x_2, ...x_n)$ for all cases. ■