CS446: Machine Learning **Spring 2017**

Problem Set 7

Kaizhao Liang(kl2) *Handed In: May 1, 2017*

1. a. $Pr(x^{(j)}) = \sum_{z=\{1,2\}} \prod_i Pr(x_i^{(j)}|z)$
$Pr(x^{(j)}) = \alpha \prod_i p_i^{x_i^{(j)}}(1-p_i)^{(1-x_i^{(j)})} + (1-\alpha)\prod_i q_i^{x_i^{(j)}}(1-q_i)^{(1-x_i^{(j)})}$

b. By Bayse rule:
$f_z^{(j)} = Pr(Z=z|x^{(j)}) = \frac{Pr(x^{(j)}|Z=z)Pr(Z=z)}{Pr(x^{(j)})}$
$f_1^{(j)} = \frac{\alpha \prod_i p_i^{x_i^{(j)}}(1-p_i)^{(1-x_i^{(j)})}}{\alpha \prod_i p_i^{x_i^{(j)}}(1-p_i)^{(1-x_i^{(j)})}+(1-\alpha)\prod_i q_i^{x_i^{(j)}}(1-q_i)^{(1-x_i^{(j)})}}$
$f_2^{(j)} = \frac{(1-\alpha) \prod_i q_i^{x_i^{(j)}}(1-q_i)^{(1-x_i^{(j)})}}{\alpha \prod_i p_i^{x_i^{(j)}}(1-p_i)^{(1-x_i^{(j)})}+(1-\alpha)\prod_i q_i^{x_i^{(j)}}(1-q_i)^{(1-x_i^{(j)})}}$

c. $E[LL] = E(\sum_{j=[1,m]} log(Pr(x^{(j)}|p,q,\alpha))) = \sum_{j=[1,m]} E(log(Pr(x^{(j)}|p,q,\alpha)))$
$Pr_z^j = Pr(Z=z|x^{(j)},p^0,q^0,\alpha^0) = f_z^{(j)}(p^0,q^0,\alpha^0)$ where p^0,q^0,α^0 are the original parameters.

$E[LL] = \sum_{j=[1,m]}\sum_{z=\{1,2\}} Pr_z^j log(Pr(Z=z,x^{(j)}|\tilde{p},\tilde{q},\tilde{\alpha})) - \sum Pr_z^j log(Pr_z^j)$

$E[LL] = \sum_{j=[1,m]} Pr_1^{(j)} log(\tilde{\alpha}\prod_i \tilde{p_i}^{x_i^{(j)}}(1-\tilde{p_i})^{(1-x_i^{(j)})}) + Pr_2^{(j)} log((1-\tilde{\alpha})\prod_i \tilde{q_i}^{x_i^{(j)}}(1-\tilde{q_i})^{(1-x_i^{(j)})} - \sum_j Pr_1^{(j)} log(Pr_1^{(j)}) + Pr_2^{(j)} log(Pr_2^{(j)})$

$E[LL] = \sum_{j=[1,m]}(Pr_1^{(j)}(log(\tilde{\alpha}) + \sum_{i=[1,n+1]} x_i^{(j)} log(\tilde{p_i}) + (1-x_i^{(j)})log(1-\tilde{p_i})) + Pr_2^{(j)}(log((1-\tilde{\alpha})) + \sum_{i=[1,n+1]} x_i^{(j)} log(\tilde{q_i}) + (1-x_i^{(j)})log(1-\tilde{q_i}))) - \sum_j (Pr_1^{(j)} log(Pr_1^{(j)}) + Pr_2^{(j)} log(Pr_2^{(j)}))$

d. Given that $Pr_1^j + Pr_2^j = 1$,
$where\ Pr_i^j = f_i^j$.
To maximize the E(LL):
$\frac{\partial E(LL)}{\partial \tilde{\alpha}} = \sum_{j=[1,m]} \frac{Pr_1^j}{\tilde{\alpha}} - \frac{Pr_2^j}{1-\tilde{\alpha}} = 0$
$\tilde{\alpha} = \frac{\sum_j Pr_1^j}{m}$
$\frac{\partial E(LL)}{\partial \tilde{p_i}} = \sum_{j=[1,m]} Pr_1^j(\frac{x_i^{(j)}}{\tilde{p_i}} - \frac{1-x_i^{(j)}}{1-\tilde{p_i}}) = 0$
$\tilde{p_i} = \frac{\sum_{j=[1,m]} Pr_1^j x_i^{(j)}}{\sum_{j=[1,m]} Pr_1^j}$
$\frac{\partial E(LL)}{\partial \tilde{q_i}} = \sum_{j=[1,m]}(1-Pr_1^j)(\frac{x_i^{(j)}}{\tilde{q_i}} - \frac{1-x_i^{(j)}}{1-\tilde{q_i}}) = 0$
$\tilde{q_i} = \frac{\sum_{j=[1,m]}(1-Pr_1^j)x_i^{(j)}}{\sum_{j=[1,m]}(1-Pr_1^j)}$

e. Update rules for:
α : Update the α as the average of the Probability that each sample was generated with z equal to 1 given current parameters.It is approximating the true average.
$\tilde{p}_i$: Update the $\tilde{p}_i$ as the Probability of all $x_i = 1$ given that $z = 1$ with the current parameters. It's approximating the true p_i
$\tilde{q}_i$: Update the $\tilde{q}_i$ as as the Probability of all $x_i = 1$ given that $z = 2$ with current parameters. It's approximating the true q_i

---

$Initialization:$
$set\ \alpha\ to\ random\ real\ number\ between\ 0\ and\ 1$
$set\ p\ to\ an\ array\ of\ length\ n+1\ with\ random\ real\ number\ between\ 0\ and\ 1$
$set\ q\ to\ an\ array\ of\ length\ n+1\ with\ random\ real\ number\ between\ 0\ and\ 1$
$set\ \theta\ to\ a\ small\ number\ as\ the\ termination\ criteria$
do
$\alpha \leftarrow \alpha'$
$p_i \leftarrow p'_i$
$q_i \leftarrow q'_i$
$\alpha' = 0$
$p' = [0]^{n+1}$
$q' = [0]^{n+1}$
  for all $x^{(j)}$ **do**
    $Pr^j_1 = f^j_1(\alpha, p, q)$
    $Pr^j_2 = f^j_2(\alpha, p, q)$
    $Pr^j_2 =$
    $\alpha' \leftarrow \alpha' + Pr^j_1$
    $p'_i \leftarrow p'_i + Pr^j_1 x_i$
    $q'_i \leftarrow q'_i + Pr^j_2 x_i$
  end for
$p'_i \leftarrow \frac{p'_i}{\alpha'}$
$q'_i \leftarrow \frac{q'_i}{m-\alpha'}$
$\alpha' \leftarrow \frac{\alpha'}{m}$
while $|q - q'| + |p - p'| + |\alpha - \alpha'| \geq \theta$

---

*Note that the $f^j_i = Pr(Z = i|x^{(j),\alpha,p,q})$ in the inner **for** loop, which is the equation derived in (b.).
*Also the update rules derived in (e.) are used after the **for** loop to update the $p', q'\ and\ \alpha'$.

f. $x_0 = sign(log(\frac{Pr(X_0=1)}{Pr(X_0=0)}))$
$Pr(X_0 = 1) = Pr(Z = 1|x_1, ..., x_n)p_0 + Pr(Z = 2|x_1, ..., x_n)q_0$
$Pr(X_0 = 0) = Pr(Z = 1|x_1, ..., x_n)(1 - p_0) + Pr(Z = 2|x_1, ..., x_n)(1 - q_0)$
$Pr(Z = 1|x_1, ..., x_n) = \frac{Pr(x_1,...,x_n|Z=1)Pr(Z=1)}{Pr(x_1,...,x_n)} = \frac{\alpha \prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}}{\alpha \prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i} + (1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}}$

$Pr(Z=2|x_1,...,x_n) = \frac{Pr(x_1,...,x_n|Z=2)Pr(Z=2)}{Pr(x_1,...,x_n)} = \frac{(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}}{\alpha\prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}+(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}}$
So,
$x_0 = sign(log(p_0\alpha\prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i} + q_0(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}) - log((1-p_0)\alpha\prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i} + (1-q_0)(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}))$

g. rewrite the decision surface, we have:
$x_0 = sign(log(\frac{p_0\alpha\prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}+q_0(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i}}{(1-p_0)\alpha\prod_{i=1}^n p_i^{x_i}(1-p_i)^{1-x_i}+(1-q_0)(1-\alpha)\prod_{i=1}^n q_i^{x_i}(1-q_i)^{1-x_i})}))$
$x_0 = sign(log(\frac{p_0\alpha+q_0(1-\alpha)\prod_{i=1}^n(\frac{q_i}{p_i})^{x_i}(\frac{1-q_i}{1-p_i})^{1-x_i}}{(1-p_0)\alpha+(1-q_0)(1-\alpha)\prod_{i=1}^n(\frac{q_i}{p_i})^{x_i}(\frac{1-q_i}{1-p_i})^{1-x_i})}))$
$x_0 = sign((2p_0-1)\alpha + (2q_0-1)(1-\alpha)\prod_{i=1}^n(\frac{q_i}{p_i})^{x_i}(\frac{1-q_i}{1-p_i})^{1-x_i}))$
if p_0 and q_0 are both greater or both less than $\frac{1}{2}$ Then the label is bound to be 1 or 0, and the decision is made independent of x_iś given the way we choose the label.
Otherwise,given that log is a concave and singluarly increasing function, rewrite the decision surface again:
$x_0 = sign(\sum_{i=1}^n x_i(log(\frac{q_i}{p_i}) - log(\frac{1-q_i}{1-p_i})) + \sum_{i=1}^n log(\frac{1-q_i}{1-p_i}) + log(|\frac{(1-2q_0)(1-\alpha)}{(1-2p_0)\alpha}|))$ Since we can rewrite the decision surface into $w^Tx+\theta$,where $w = [log(\frac{q_i}{p_i}) - log(\frac{1-q_i}{1-p_i})]^n$ and the $\theta = \sum_{i=1}^n log(\frac{1-q_i}{1-p_i}) + log(|\frac{(1-2q_0)(1-\alpha)}{(1-2p_0)\alpha}|)$, the decision surface is linear.

2. a. "The two directed trees obtained from T are equivalent" means that the probability of any event or conditional event will be the same no matter which tree it's computed by.

b. Base case: Pick two root nodes r_1 and r_2 which are next to each other, then we have two trees T_1 *and* T_2 with only difference on the edge (r_1, r_2). For any event X:
In the tree with r_1 as root:
$Pr(X) = Pr(r_1)\prod_{x_i\in subtree_{r_1}} Pr(x_i|Parents(x_i))\cdot Pr(r_2|r_1)\prod_{x_i\in subtree_{r_2}} Pr(x_i|Parents(x_i))$
In the tree with r_2 as root:
$Pr'(X) = Pr(r_1|r_2)\prod_{x_i\in subtree_{r_1}} Pr(x_i|Parents(x_i))\cdot Pr(r_2)\prod_{x_i\in subtree_{r_2}} Pr(x_i|Parents(x_i))$
By Bayse rule:
$Pr(r_1|r_2)Pr(r_2) = Pr(r_2|r_1)Pr(r_1)$
So, $Pr'(X) = Pr(X)$, i.e, the two trees are equivalent.
Induction hypothesis: if two roots are k steps away from each other, the two trees generated are equivalent.
For the two nodes that are k+1 steps away, the trees generated are also equivalent, because:
Pick two nodes r_1, r_2 that are k steps away, the two trees T_1, T_2 are equivalent by inductive hypothesis. By base case, if we pick a node r_3 one step from r_1 *or* r_2 away from the root that is not choosen, the tree T_3 is equivalent to T_1 *or* T_2. By transition, T_1, T_2, T_3 are all equivalent. So the hypothesis also holds for k+1.
In conclusion, no matter which root we pick, the directed tree generated will be always equivalent.

Given two events X and X' of $x_1, x_2, ..., x_n$ and $x'_1, x'_2, ..., x'_n$, the conditional probability, by Bayse Rule:

$Pr(X|X') = \frac{Pr(\{X+X'\})}{Pr(X')}$
$Pr(\{X + X'\})$ and $Pr(X')$ are proven not to be changed when we pick different nodes as root. So the probability of the conditional event also does not change.