CS446: Machine Learning **Fall 2015**

Problem Set 7

Dominik Durner *Handed In: December 3, 2015*

Question 1

1.a

$$P[w_j, d_i] = P[w_j|d_i]P[d_i] = P[d_i]\sum_{k=1}^{2} P[w_j, c_k|d_i] =$$

$$= \sum_{k=1}^{2} P[d_i]P[w_j, c_k|d_i] = \sum_{k=1}^{2} \frac{P[w_j|c_k, d_i]P[d_i|c_k]P[w_j, c_k]}{P[w_j|c_k]} =$$

$$= \sum_{k=1}^{2} \frac{P[w_j|c_k, d_i]P[c_k|d_i]P[d_i]P[w_j|c_k]P[c_k]}{P[w_j|c_k]P[c_k]} =$$

$$= \sum_{k=1}^{2} P[w_j|c_k, d_i]P[c_k|d_i]P[d_i]$$

$$\Rightarrow^{Ind.} \sum_{k=1}^{2} P[w_j|c_k]P[c_k|d_i]P[d_i]$$

1.b

$$P[c_k|w_j, d_i] = \frac{P[d_i, c_k|w_j]P[w_j]}{P[w_j, d_i]} = \frac{P[w_j|d_i, c_k]P[d_i, c_k]}{P[w_j, d_i]} =$$

$$= \frac{P[w_j|d_i, c_k]P[c_k|d_i]P[d_i]}{P[w_j, d_i]}$$

$$\Rightarrow^{Ind.} \frac{P[w_j|c_k]P[c_k|d_i]P[d_i]}{\sum_{m=1}^{2} P[w_j|c_m]P[c_m|d_i]P[d_i]} =$$

$$= \frac{P[w_j|c_k]P[c_k|d_i]}{\sum_{m=1}^{2} P[w_j|c_m]P[c_m|d_i]}$$

1.c

$$\forall i : P_\Theta[d_i] = \prod_j \prod_k P_\Theta[d_i, c_k, w_j]^{n(d_i,w_j)}$$

$$\Rightarrow \forall i : logP_\Theta[d_i] = \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} log(P_\Theta[d_i, c_k, w_j]) =$$

$$= \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} log(P_\Theta[w_j|d_i, c_k]P_\Theta[c_k|d_i]P_\Theta[d_i]) =$$

$$= \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} (log(P_\Theta[w_j|d_i, c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))$$

$$\Rightarrow^{Ind.} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} (log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))$$

Since we want to get the expected likelihood of the overall data we have to calculate the following.

$$E[LL] = E[\sum_{i=1}^{M} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} (log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))]$$

With Jensen's inequality we know that $E[log(X)] \leq log(E[X])$.

$$E[LL] = E[\sum_{i=1}^{M} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} (log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))] =$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} E[(log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))] \leq$$

$$\leq \sum_{i=1}^{M} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} P_\Theta[c_k|w_j, d_i](log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i])) =$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{V} n(d_i, w_j) \sum_{k=1}^{2} \frac{P[w_j|c_k]P[c_k|d_i]}{\sum_{m=1}^{2} P[w_j|c_m]P[c_m|d_i]} (log(P_\Theta[w_j|c_k]) + log(P_\Theta[c_k|d_i]) + log(P_\Theta[d_i]))$$

1.d

Since we have the following equations given from basic probability theory and what we want to maximize, we can use Lagrangian Multiplier to derive our deviation and maximization.

$$f(\Theta) : max_{\Theta} E[LL]$$
$$g_1 : \sum_{i=1}^{M} P[d_i] = 1$$
$$g_2 : \forall k : \sum_{j=1}^{V} P[w_j|c_k] = 1$$
$$g_3 : \forall i : \sum_{k=1}^{2} P[c_k|d_i] = 1$$

Using Lagrange we get the following equations.

$$\bigtriangledown L(\Theta, \lambda) = \begin{pmatrix} \bigtriangledown f(\Theta) + \lambda_1 \bigtriangledown g_1(\Theta) + \lambda_2 \bigtriangledown g_2(\Theta) + \lambda_3 \bigtriangledown g_3(\Theta) \\ g_1(\Theta) \\ g_2(\Theta) \\ g_3(\Theta) \end{pmatrix} =$$
$$= \begin{pmatrix} \sum_{j=1}^{V} \sum_{k=1}^{2} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i] \frac{1}{P_{\Theta}[d_i]} + \lambda_1 \\ \sum_{i=1}^{M} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i] \frac{1}{P_{\Theta}[w_j|c_k]} + \lambda_2 \\ \sum_{j=1}^{V} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i] \frac{1}{P_{\Theta}[c_k|d_i]} + \lambda_3 \\ (\sum_{i=1}^{M} P[d_i]) - 1 \\ \sum_{k=1}^{2} ((\sum_{j=1}^{V} P[w_j|c_k]) - 1) \\ \sum_{i=1}^{M} ((\sum_{k=1}^{2} P[c_k|d_i]) - 1) \end{pmatrix} = \overrightarrow{0}$$

Hence we get the following equations.

$$\forall i : P_{\Theta}[d_i] = \frac{-1}{\lambda_1} \sum_{j=1}^{V} \sum_{k=1}^{2} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i]$$
$$\forall j, k : P_{\Theta}[w_j|c_k] = \frac{-1}{\lambda_2} \sum_{i=1}^{M} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i]$$
$$\forall i, k : P_{\Theta}[c_k|d_i] = \frac{-1}{\lambda_3} \sum_{j=1}^{V} n(d_i, w_j) P_{\Theta}[c_k|w_j, d_i]$$

With our g_i we can further calculate the λ (insert into g_i, solve to λ and reinsert) and write the above equations in a fully calculated manner.

$$\forall i : P_\Theta[d_i] = \frac{\sum_{j=1}^{V}\sum_{k=1}^{2} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{\hat{i}=1}^{M}\sum_{j=1}^{V}\sum_{k=1}^{2} n(d_{\hat{i}}, w_j) P_\Theta[c_k|w_j, d_{\hat{i}}]} = \frac{\sum_{j=1}^{V} n(d_i, w_j)}{\sum_{\hat{i}=1}^{M}\sum_{j=1}^{V} n(d_{\hat{i}}, w_j)}$$

$$\forall j, k : P_\Theta[w_j|c_k] = \frac{\sum_{i=1}^{M} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{i=1}^{M}\sum_{\hat{j}=1}^{V} n(d_i, w_{\hat{j}}) P_\Theta[c_k|w_{\hat{j}}, d_i]}$$

$$\forall i, k : P_\Theta[c_k|d_i] = \frac{\sum_{j=1}^{V} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{j=1}^{V}\sum_{\hat{k}=1}^{2} n(d_i, w_j) P_\Theta[c_{\hat{k}}|w_j, d_i]} = \frac{\sum_{j=1}^{V} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{j=1}^{V} n(d_i, w_j)}$$

1.e

$$\forall i : P_\Theta[d_i] = \frac{\sum_{j=1}^{V} n(d_i, w_j)}{\sum_{\hat{i}=1}^{M}\sum_{j=1}^{V} n(d_{\hat{i}}, w_j)}$$

$P[d_i]$ can be seen as the total count of all words occurring in a document d_i over the total number of words across all documents.

$$\forall j, k : P_\Theta[w_j|c_k] = \frac{\sum_{i=1}^{M} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{i=1}^{M}\sum_{\hat{j}=1}^{V} n(d_i, w_{\hat{j}}) P_\Theta[c_k|w_{\hat{j}}, d_i]}$$

$P[w_j|c_k]$ is the weighted average of the probabilities of the word w_j being labeled with class c_k across all the documents in which it occurs divided by the same weighted average of all the worlds across all the documents.

$$\forall i, k : P_\Theta[c_k|d_i] = \frac{\sum_{j=1}^{V} n(d_i, w_j) P_\Theta[c_k|w_j, d_i]}{\sum_{j=1}^{V} n(d_i, w_j)}$$

$P[c_k|d_i]$ can be interpreted as the weighted sum of the probabilities of each word in the document d_i being classified as label c_k over the total number of words across all documents.

1.f

1) Initialization: We assign initial estimates to the parameters $P[d_i], P[w_j|c_k]$ and $P[c_k|d_i]$.

$$\forall i : P[d_i] = \frac{1}{M}$$
$$\forall i, k : P[c_k|d_i] = \frac{1}{C} = \frac{1}{2}$$
$$\forall j, k : P[w_j|c_k] = \frac{1}{V}$$

2) E-M step: Calculate the posterior probabilities $P[c_k|w_j, d_i]$.

$$\forall i, j, k : P[c_k|w_j, d_i] = \frac{P[w_j|c_k]P[c_k|d_i]}{\sum_{m=1}^{2} P[w_j|c_m]P[c_m|d_i]}$$

Now calculate the next iteration values of the parameters $P[d_i], P[w_j|c_k]$ and $P[c_k|d_i]$

$$\forall i : P_\Theta[d_i] = \frac{\sum_{j=1}^{V} n(d_i, w_j)}{\sum_{\hat{i}=1}^{M}\sum_{j=1}^{V} n(d_{\hat{i}}, w_j)}$$

$$\forall j, k : P_\Theta[w_j|c_k] = \frac{\sum_{i=1}^{M} n(d_i, w_j)P_\Theta[c_k|w_j, d_i]}{\sum_{i=1}^{M}\sum_{\hat{j}=1}^{V} n(d_i, w_{\hat{j}})P_\Theta[c_k|w_{\hat{j}}, d_i]}$$

$$\forall i, k : P_\Theta[c_k|d_i] = \frac{\sum_{j=1}^{V} n(d_i, w_j)P_\Theta[c_k|w_j, d_i]}{\sum_{j=1}^{V} n(d_i, w_j)}$$

3) Go back to step 2) until we have converged values (difference between the previous round and the current round is smaller a small threshold) for $P[d_i], P[w_j|c_k]$ and $P[c_k|d_i]$

Question 2

2.a

For showing that any two directed trees constructed from taking any node as the root and directing all edges away from the root are equivalent, we can just show that the joint probabilities are the same for both trees. Therefore, tree 1 with root $x_{r1} \in \{x_1, ..., x_n\}$ results into $P(x_1, x_2, ..., x_n)$. The same is true for tree 2 which has root $x_{r2} \in \{x_1, ..., x_n\}$ and also has the same joint distribution $P(x_1, x_2, ..., x_n)$. This is a result of writing the joint distribution as conditional probabilities, such that $P(x_1, x_2, ..., x_n) = P(x_1)P(x_2|x_1)\cdots$ with root x_1. The idea is that if I change the root some of the conditional probabilities change direction. For example a tree with root x_2 looks like $P(x_1, x_2, ..., x_n) = P(x_2)P(x_1|x_2)\cdots$. The task for question b is to show that for any root these two equation result into the same joint distribution.

2.b

As stated in 2.a., we have two trees which should be the same probability in the end but they have a different root $x_{r1} \neq x_{r2}$.

$$P(x_{r1}, ..., x_n) = P(x_{r1}) \prod_{i \neq r1} P(x_i|parent(x_i))$$
$$P(x_{r2}, ..., x_n) = P(x_{r2}) \prod_{i \neq r2} P(x_i|parent(x_i))$$

The goal is now the go from $P(x_{r1}) \prod_{i \neq r1} P(x_i, parent(x_i))$ to $P(x_{r2}) \prod_{i \neq r2} P(x_i, parent(x_i))$. The basic idea is to change the root we only need to change the path from x_{r1} to x_{r2}. Therefore, all other children of x_{r1} stay the same and all childrens of x_{r2} stay the same. This can be easily seen since for example the probability with parent x_{r2} is only dependent on x_{r2} and no other variable. Hence, we can still use the same way of going through this subtree.

$$\begin{aligned}
P(x_{r1}, ..., x_n) = P(x_{r1}) \prod_{x \in path(x_{r1}, x_{r2})} P(x|parent(x)) \prod_{x \in subtree(x_{r1})} P(x|parent(x)) \\
\prod_{x \in subtree(x_{r2})} P(x|parent(x)) = \\
= P_{r1} * P_{r2} * P(x_{r1}) \prod_{x \in path(x_{r1}, x_{r2})} P(x|parent(x)) = \\
= P_{r1} * P_{r2} * P(x_{r1})P(x_i|x_{r1})P(x_{i+1}|x_i)\cdots P(x_j|x_{j-1})P(x_{r2}|x_j) = \\
= P_{r1} * P_{r2} * P(x_{r1}|x_i)P(x_i)P(x_{i+1}|x_i)\cdots P(x_j|x_{j-1})P(x_{r2}|x_j) = \\
= P_{r1} * P_{r2} * P(x_{r1}|x_i)P(x_i|x_{i+1})P(x_{i+1})\cdots P(x_j|x_{j-1})P(x_{r2}|x_j) = \\
= P_{r1} * P_{r2} * P(x_{r1}|x_i)P(x_i|x_{i+1})\cdots P(x_j|x_{r2})P(x_{r2}) = \\
= P_{r1} * P_{r2} * P(x_{r2}) \prod_{x \in path(x_{r2}, x_{r1})} P(x|parent(x)) = \\
= P(x_{r2}, ..., x_n)
\end{aligned}$$