CS446: Machine Learning	**Fall 2014**
Problem Set 7	
Chen Zhang	*Handed In: 12/03/2014*

1 Problem 1

1.1 Part A

It's obvious that

$$P(w_j, d_i, c_k) = P(d_i)P(c_k|d_i)P(w_j|c_k).$$

Then it could be derived that

$$P(w_j, d_i) = \sum_k P(d_i)P(c_k|d_i)P(w_j|c_k)$$

1.2 Part B

$$\begin{aligned} P(c_k|w_j, d_i) = \frac{P(w_j, d_i, c_k)}{P(w_j, d_i)} &= \frac{P(d_i)P(c_k|d_i)P(w_j|c_k, d_i)}{w_j, d_i} \quad \text{Note}(P(w_j|c_k, d_i) = P(w_j|c_k)) \\ &= \frac{P(d_i)P(c_k|d_i)P(w_j|c_k)}{\sum_k P(d_i)P(c_k|d_i)P(w_j|c_k)} \\ &= \frac{P(c_k|d_i)P(w_j|c_k)}{\sum_k P(c_k|d_i)P(w_j|c_k)} \end{aligned}$$

1.3 Part C

Suppose that if we know the latent, that is if we know the c_k of each observed data point, then we have

$$L = \prod_i \prod_j P(w_j, c_k, d_i)^{(n(d_i, w_j))}$$

$$\log L = \sum_i \sum_j n(d_i, w_j) \log \left[P(w_j, d_j, c_k)\right]$$

Since in fact we don't know the c_k value for data points, then we take the expectation, and we have

$$\begin{aligned} E_{(c_k|w_j,d_i)}[\log L] &= E_{(c_k|w_j,d_i)}[\sum_i \sum_j n(d_i, w_j) \log [P(w_j, d_j, c_k)]] \\ &= \sum_i \sum_j n(d_i, w_j) E_{(c_k|w_j,d_i)}[\log [P(w_j, d_j, c_k)]] \\ &= \sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i) \log [P(w_j, d_j, c_k)] \end{aligned}$$

We also have

$$P(w_j, d_i, c_k) = P(d_i)P(c_k|d_i)P(w_j|c_k).$$

Plugging in the equation and apply the Lagrange multiplier, we have

$$\begin{aligned} E_{(c_k|w_j,d_i)}[\log L] &= \sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i) \log [P(d_i)P(c_k|d_i)P(w_j|c_k)] \\ &+ \lambda[\sum_i P(d_i) - 1] + \sum_i \beta_i[\sum_k P(c_k|d_i) - 1] + \sum_k \eta_k[\sum_j P(w_j|c_k) - 1] \\ &= \sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i) \times [\log P(d_i) + \log P(c_k|d_i) + \log P(w_j|c_k)] \\ &+ \lambda[\sum_i P(d_i) - 1] + \sum_i \beta_i[\sum_k P(c_k|d_i) - 1] + \sum_k \eta_k[\sum_j P(w_j|c_k) - 1] \end{aligned}$$

Note that $P(c_k|w_j, d_i)$ has the expression calculated from Part B

$$P(c_k|w_j, d_i) = \frac{P(c_k|d_i)P(w_j|c_k)}{\sum_k P(c_k|d_i)P(w_j|c_k)}.$$

1.4 Part D

1.4.1 Unit 1

Taking derivative with respect to $P(d_1)$ and set to zero, we get

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(d_1)} = \frac{\sum_j \sum_k n(d_1, w_j) P(c_k|w_j, d_1)}{P(d_1)} + \lambda = 0$$

Similarly we have

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(d_2)} = \frac{\sum_j \sum_k n(d_2, w_j) P(c_k|w_j, d_2)}{P(d_2)} + \lambda = 0$$

Using these two equations we have

$$P(d_2) = \frac{\sum_j \sum_k n(d_2, w_j) P(c_k|w_j, d_2)}{\sum_j \sum_k n(d_1, w_j) P(c_k|w_j, d_1)} P(d_1).$$

Similarly we could have

$$P(d_i) = \frac{\sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i)}{\sum_j \sum_k n(d_1, w_j) P(c_k|w_j, d_1)} P(d_1).$$

Taking derivative with respect to the Lagrange Multiplier we have

$$P(d_1) + P(d_2) + \dots P(d_i) = 1.$$

Plug in the expression for $P(d_i)$, we could have

$$\frac{\sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i)}{\sum_j \sum_k n(d_1, w_j) P(c_k|w_j, d_1)} P(d_1) = 1.$$

Thus $P(d_1)$ is calculated as

$$P(d_1) = \frac{\sum_j \sum_k n(d_1, w_j) P(c_k|w_j, d_1)}{\sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i)}.$$

Similarly the expression of $P(d_i)$ can be generalized as

$$P(d_i) = \frac{\sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i)}{\sum_i \sum_j \sum_k n(d_i, w_j) P(c_k|w_j, d_i)}.$$

1.4.2 Unit 2

Taking derivative with respect to $P(c_1|d_i)$ and set to zero, we get

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(c_1|d_i)} = \frac{\sum_j n(d_i, w_j) P(c_1|w_j, d_i)}{P(c_1|d_i)} + \beta_i = 0$$

Similarly we have

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(c_2|d_i)} = \frac{\sum_j n(d_i, w_j) P(c_2|w_j, d_i)}{P(c_2|d_i)} + \beta_i = 0$$

Using these two equations we have

$$P(c_2|d_i) = \frac{\sum\limits_j n(d_i, w_j)P(c_2|w_j, d_i)}{\sum\limits_j n(d_i, w_j)P(c_1|w_j, d_i)}P(c_1|d_i)$$

Similarly we could have

$$P(c_k|d_i) = \frac{\sum\limits_j n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_j n(d_i, w_j)P(c_1|w_j, d_i)}P(c_1|d_i)$$

Taking derivative with respect to the Lagrange Multiplier we have

$$P(c_1|d_i) + P(c_2|d_i) + \ldots P(c_k|d_i) = 1.$$

Plug in the expression for $P(c_k|d_i)$, we could have

$$\frac{\sum\limits_k \sum\limits_j n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_j n(d_i, w_j)P(c_1|w_j, d_i)}P(c_1|d_i) = 1$$

Thus $P(c_1|d_i)$ is calculated as

$$P(c_1|d_i) = \frac{\sum\limits_j n(d_i, w_j)P(c_1|w_j, d_i)}{\sum\limits_k \sum\limits_j n(d_i, w_j)P(c_k|w_j, d_i)}$$

Similarly the expression of $P(c_k|d_i)$ can be generalized as

$$P(c_k|d_i) = \frac{\sum\limits_j n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_k \sum\limits_j n(d_i, w_j)P(c_k|w_j, d_i)}$$

1.4.3 Unit 3

Taking derivative with respect to $P(w_1|c_k)$ and set to zero, we get

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(w_1|c_k)} = \frac{\sum\limits_i n(d_i, w_1)P(c_k|w_1, d_i)}{P(w_1|c_k)} + \eta_k = 0$$

Similarly we have

$$\frac{\partial E_{(c_k|w_j,d_i)}[\log L]}{\partial P(w_2|c_k)} = \frac{\sum\limits_i n(d_i, w_2)P(c_k|w_2, d_i)}{P(w_2|c_k)} + \eta_k = 0$$

Using these two equations we have

$$P(w_2|c_k) = \frac{\sum\limits_i n(d_i, w_2)P(c_k|w_2, d_i)}{\sum\limits_i n(d_i, w_1)P(c_k|w_1, d_i)}P(w_1|c_k)$$

Similarly we could have

$$P(w_j|c_k) = \frac{\sum\limits_i n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_i n(d_i, w_1)P(c_k|w_1, d_i)}P(w_1|c_k)$$

Taking derivative with respect to the Lagrange Multiplier we have

$$P(w_1|c_k) + P(w_2|c_k) + \ldots P(w_j|c_k) = 1.$$

Plug in the expression for $P(w_j|c_k)$, we could have

$$\frac{\sum\limits_j \sum\limits_i n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_i n(d_i, w_1)P(c_k|w_1, d_i)}P(w_1|c_k) = 1$$

Thus $P(w_1|c_k)$ is calculated as

$$P(w_1|c_k) = \frac{\sum\limits_i n(d_i, w_1)P(c_k|w_1, d_i)}{\sum\limits_j \sum\limits_i n(d_i, w_j)P(c_k|w_j, d_i)}$$

Similarly the expression of $P(w_j|c_k)$ can be generalized as

$$P(w_j|c_k) = \frac{\sum\limits_i n(d_i, w_j)P(c_k|w_j, d_i)}{\sum\limits_j \sum\limits_i n(d_i, w_j)P(c_k|w_j, d_i)}$$

In conclusion, the Unit 1 Unit 3 are the update rules of the parameters.

1.5 Part E

The update rules make sense since for multinomial distribution, the distribution associated with one parameter is basically the ratio of the probability associated with this parameter over its marginal probability.
The pseudo-code would be
Step 1: Put a set of initial guess to $P(d_i)^0$, $P(c_k|d_i)^0$, and $P(w_j|c_k)^0$.
Step 2: Use $P(d_i)^m$, $P(c_k|d_i)^m$, and $P(w_j|c_k)^m$ to calculate $P(c_k|d_i, w_j)^m$, as in Part B.
Step 3: Use $P(d_i)^m$, $P(c_k|d_i)^m$, $P(w_j|c_k)^m$ and $P(c_k|d_i, w_j)^m$ to calculated a new set of $P(d_i)^{m+1}$, $P(c_k|d_i)^{m+1}$, $P(w_j|c_k)^{m+1}$, as in the update rule calculated in Part D.
Step 4: If the result from Step 3 converges, then stop. If it does not converge, set $m = m+1$ and repeat from Step 2.

2 Problem 2

2.1 Part A

For the two distribution to be the same, the probability of the set of events calculated from the two different trees need to be the same, which means

$$\prod_D P(r_1) \prod_i P(x_i|parents(x_i)) = \prod_D P(r_2) \prod_j P(x_j|parents(x_j))$$

In fact we only need the single event calculated from the two directed trees to be the same. The probability of one single event calculated with trees with root r_1 and r_2 are

$$P(r_1, x_1, x_2 \dots x_n) = P(r_1) \prod_i P(x_i|parents(x_i))$$
$$P(r_2, x_1, x_2 \dots x_n) = P(r_2) \prod_j P(x_j|parents(x_j))$$

For the two directed trees to be equivalent, the probability of the event need to be the same, which follows

$$P(r_1, x_1, x_2 \dots x_n) = P(r_2, x_1, x_2 \dots x_n).$$

Namely,

$$P(r_1) \prod_i P(x_i|parents(x_i)) = P(r_2) \prod_j P(x_j|parents(x_j))$$

2.2 Part B

Suppose that we have a directed tree with root r_1. The probability of this event can be calculated as

$$P(r_1, x_1, x_2 \dots x_n) = P(r_1) \prod_i P(x_i|parents(x_i)) \tag{1}$$

Now we change the root to r_2, and also change the corresponding directions of the edges. Now the probability of this event can be calculated as

$$P(r_2, x_1, x_2 \dots x_n) = P(r_2) \prod_i P(x_i|parents(x_i)) \tag{2}$$

Note that only the direction in the path of r_1 and r_2 are inverted. The direction of the rest of the edges remain the same. Thus the difference of (1) and (2) only lies in the part which describes the path from r_1 to r_2. Without loss of generality, we can assume that node a, b, and c lies in between r_1 and r_2. So the difference in (1) and (2) is that part of (1) is

$$P(1)P(a|1)P(b|a)P(c|b)P(2|c)$$

and part of (2) is (the inverted direction)

$$P(2)P(c|2)P(b|c)P(a|b)P(1|a).$$

But from the Bayes model we know that

$$P(1)P(a|1)P(b|a)P(c|b)P(2|c) = P(1,a,b,c,2) = P(2)P(c|2)P(b|c)P(a|b)P(1|a).$$

This means that the only different part in (1) and (2) ends up being the same, which mean (1) and (2) are the same. Using the definition from part A, we can see that these two directed trees obtained from T are equivalent.
Note that the proof is arbitrary, meaning that now matter how we change the root of the tree, we always end up with equivalent trees.