CS446: Machine Learning **Fall 2015**

Problem Set 7

Jifu Zhao *Handed In: 11/21/2015*

1. **Answer to problem 1**

(a). According to the definition, we can get that:

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

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

(b).

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

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

And, we also have:

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

So, finally:

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

(c). Likelihood should be:

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

$$LL = \sum_i \sum_j n(d_i, w_j) log P(d_i, w_j)$$

Since that:

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

$$LL = \sum_i \sum_j n(d_i, w_j) log[P(d_i)\sum_k P(w_j|c_k)P(c_k|d_i)]$$

Suppose that k can be 1 and 2, so:

$$\mathbf{E}[\mathbf{LL}] = \sum_i \sum_j n(d_i, w_j)\{P(c_1|w_j, d_i)log[P(d_i)P(w_j|c_1)P(c_1|d_i)]$$

$$+P(c_2|w_j, d_i)log[P(d_i)P(w_j|c_2)P(c_2|d_i)]\}$$

(d). Given the value to be maximized:

$$\mathbf{E}[\mathbf{LL}] = \sum_i \sum_j n(d_i, w_j)\{P(c_1|w_j, d_i)log[P(d_i)P(w_j|c_1)P(c_1|d_i)]$$

$$+P(c_2|w_j, d_i)log[P(d_i)P(w_j|c_2)P(c_2|d_i)]\}$$

Also notice that:

$$\sum_i P(d_i) = 1$$

$$\sum_k P(c_k|d_i) = 1$$

$$\sum_j P(w_j|c_k) = 1$$

Solve these equations, solve for $\mathbf{P(w_j|c_k)}$ and $\mathbf{P(c_k|d_i)}$, we can find that in order to maximize $\mathbf{E}[\mathbf{LL}]$, we should have that:

$$\mathbf{P(w_j|c_k)} = \frac{\sum_{i=1}^{M} n(d_i, w_j)P(c_k|d_i, w_j)}{\sum_{i=1}^{M}\sum_{l=1}^{V} n(d_i, w_l)P(c_k|d_i, w_l)}$$

$$\mathbf{P(c_k|d_i)} = \frac{\sum_{j=1}^{V} n(d_i, w_j)P(c_k|d_i, w_j)}{\sum_{j=1}^{V} n(d_i, w_j)}$$

Finally, estimate $P(d_i)$ to be $P(d_i) = 1/M$

(e). From (d), we can know the expression for $\mathbf{P(w_j|c_k)}$, $\mathbf{P(c_k|d_i)}$ and $\mathbf{P(d_i)}$. They are the corresponding update rule.

1, For $\mathbf{P(w_j|c_k)}$, the value of:

$$\frac{\sum_{i=1}^{M} n(d_i, w_j) P(c_k|d_i, w_j)}{\sum_{i=1}^{M} \sum_{l=1}^{V} n(d_i, w_l) P(c_k|d_i, w_l)}$$

means that we first iterate all documents (i = 1 to M), calculate how many times that w_j has appeared with category of c_k, then, this value is divided by the total number of documents that has category c_k.

2, For $\mathbf{P(c_k|d_i)}$, the value of:

$$\frac{\sum_{j=1}^{V} n(d_i, w_j) P(c_k|d_i, w_j)}{\sum_{j=1}^{V} n(d_i, w_j)}$$

means that first, when given the particular document d_i, we calculate how many times the category c_k appeared, and then this value is divided by the total number of words in the document d_i.

3, For $\mathbf{P(d_i)}$, we only need to randomly choose one document from all documents without any more information, so its value is: $P(d_i) = 1/M$.

(f). The pseudocode is shown as below:

1, Choose the initial value for $P(w_j|c_k)$, $P(c_k|d_i)$ and $P(d_i)$. They can be random numbers, or you can guess them based on experience.

2, Do iteration:

I), According to the expression of $E(LL)$ from part(a), (b), (c), calculate the current log likelihood.

II), According to part(d), calculate new values for $P(w_j|c_k)$, $P(c_k|d_i)$ and $P(d_i)$.

III), According to the new value for $P(w_j|c_k)$, $P(c_k|d_i)$ and $P(d_i)$, do the next iteration.

IV), Stop until the expression converged.

3, Output the final conclusion.

---

2. **Answer to problem 2**

(a). In this question, it means that no matter we choose which one as the root node, finally, the directed tree we finally built is equivalent.

So, no matter which one y as the root node, the probability

$$P(y, x_1, x_2, ...) = P(y) \prod_i (x_i | Parents(x_i))$$

should be the same for all y.

(b). The mutual information between x and y, is:

$$I(x, y) = \sum_{x,y} P(x, y) \frac{P(x, y)}{P(x)P(y)}$$

In this question, we can get that:

$$I(x_i, x_j) = \sum_{x_i, x_j} P x_i, x_j) \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$

So, it is easy to derive that: $I(x_i, x_j) = I(x_j, x_i)$, so, the weight for edge (x_i, x_j) and (x_j, x_i).

This means that, no matter which node in T is chosen as the root for the "direction" stage, since the weight for edge (x_i, x_j) and (x_j, x_i) are the same, so the resulting joint probability distribution is the same, so, the resulting directed trees are all equivalent.