Assignment 5

Mousumi Akter

March 2020

1 Mixture Models

a. $P(\text{"}the\text{"}) = P(H) * P(\text{"}the\text{"}|H) + P(T) * P(\text{"}the\text{"}|T)$
$= 0.8 * 0.3 + 0.2 * 0.3$
$= 0.3$

b. As each word is written independently $P(\text{"}the\text{"}) = 0.3$ whether it appears first or second.

c. $P(H|\text{"}data\text{"}) = \frac{P(H)*P(\text{"}data\text{"}|H)}{P(H)*P(\text{"}data\text{"}|H)+P(T)*P(\text{"}data\text{"}|T)}$
$= \frac{0.8*0.1}{0.8*0.1+0.2*0.1}$
$= 0.8$

d. I expect the word "data" to be less frequent as the $p(\text{"}data\text{"}|H)$ and $p(\text{"}data\text{"}|T)$ both are low.

e. $P(\text{"}Computer\text{"}|H) = \frac{c(\text{"}Computer\text{"})}{|D|} = \frac{3}{10} = 0.3$
$P(\text{"}game\text{"}|H) = \frac{c(\text{"}game\text{"})}{|D|} = \frac{2}{10} = 0.2$

2 PLSA

2.1 PLSA With and Without the Background Topic

a. If we set $\lambda = 0$ then, the second model will be equivalent to the first

b. If we assume that the background language model is generated from unigram language model then, the maximum likelihood estimate for this language model will be $\frac{c(w,D)}{|D|}$, where $c(w, D)$ refers to count of word w in document D.

c. If we set $\lambda < 0.5$ that means we are emphasizing on background topic. On the other hand, if we set $\lambda > 0.5$ that means we are emphasizing on mixture topic model. We can test our hypothesis by setting $\lambda = 0.5$ which equally emphasize both background topic and mixture topic model.

2.2 Deriving the EM Algorithm for PLSA with a Background Topic

a.

$$n_{d,k} = \sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)$$

$$n_{w,k} = \sum_{d \in D} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k).$$

b.

$$p(z = k \mid \pi_d^{(n+1)}) = \frac{n_{d,k}}{\sum_{k'=1}^{K} n_{d,k'}} = \frac{\sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)}{\sum_{k'=1}^{K} \sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k')}$$

and

$$p(w \mid \theta_k^{(n+1)}) = \frac{n_{w,k}}{\sum_{w' \in V} n_{w',k}} = \frac{\sum_{d \in D} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)}{\sum_{w' \in V} \sum_{d \in D} c(w',d) q_y(y_{d,w'} = 1) q_{z|y}(z_{d,w'} = k)}.$$

Derivation has been shown on next page and for the derivation I took help from the document Professor shares.

We thus have a log likelihood of

$$\log p(D \mid \Theta, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|d_i|} \log \left\{ \lambda p(d_{i,j} = w \mid D) + (1-\lambda) \sum_{k=1}^{K} p(z_{i,j} = k \mid \pi_i) p(d_{i,j} = w \mid \theta_k) \right\}$$

where we can observe again the problematic summation occurring within the logarithm. We thus turn to EM again for finding the maximum likelihood estimates for the model parameters Θ and Π.

E-step: Our main computation in the E-step is to estimate the joint distribution over the latent variables Y and Z given the observations D and our current model parameters $\Theta^{(n)}$ and $\Pi^{(n)}$.

First, we can observe that

$$p(y_{i,j} = \ell, z_{i,j} = k \mid D, \Theta^{(n)}, \Pi^{(n)}) = p(y_{i,j} = \ell \mid D, \Theta^{(n)}, \Pi^{(n)}) p(z_{i,j} = k \mid y_{i,j} = \ell, D, \Theta^{n}, \Pi^{(n)})$$

and thus we can break this problem down into estimating two distributions: q_y and $q_{z|y}$ for the first and second term, respectively.

Focusing on the first term, and noting that since $y_{i,j}$ is binary random variable we can focus on only one specific case, we have

$$p(y_{i,j} = 1 \mid D, \Theta^{(n)}, \Pi^{(n)}) = p(y_{i,j} = 1 \mid d_{i,j} = w, \Theta^{(n)}, \pi_i)$$

based on our independence assumptions, and

$$= \frac{p(d_{i,j} = w \mid y_{i,j} = 1, \Theta^{(n)}, \pi_i^{(n)}) p(y_{i,j} = 1 \mid \Theta^{(n)}, \pi_i^{(n)})}{p(d_{i,j} = w \mid \Theta^{(n)}, \pi_i^{(n)})}$$

by Bayes' rule. Substituting in our model distributions, we have

$$= \frac{(1-\lambda) \sum_{k=1}^{K} p(z_{i,j} = k \mid \pi_i^{(n)}) p(w \mid \theta_k)}{\lambda p(w \mid D) + (1-\lambda) \sum_{k=1}^{K} p(z_{i,j} = k \mid \pi_i^{(n)}) p(w \mid \theta_k)}$$

and we can then set $q_y(y_{i,j} = 1) = p(y_{i,j} = 1 \mid D, \Theta^{(n)}, \Pi^{(n)})$.[7]

Let's now focus on the second term. We know that $p(z_{i,j} = 0 \mid y_{i,j} = 0, \Theta^{(n)}, \Pi^{(n)}) = 1$ by our model definition, so we only need to concern ourselves with estimating $p(z_{i,j} = k \mid y_{i,j} = 1, \Theta^{(n)}, \Pi^{(n)})$. Notice, however, that if $y_{i,j} = 1$ then we know for certain that we are sampling from the PLSA mixture (and thus $p(z_{i,j} = 0 \mid y_{i,j} = 1, \Theta^{(n)}, \Pi^{(n)}) = 0$), so we will end up with the exact same estimate for $q_{z|y}$ as we had for q in the PLSA derivation. Specifically, we have, for $k > 0$,

$$p(z_{i,j} = k \mid y_{i,j} = 1, \Theta^{(n)}, \Pi^{(n)}) = \frac{p(z_{i,j} = k \mid \pi_i^{(n)}) p(w_{i,j} \mid \theta_k^{(n)})}{\sum_{k'=1}^{K} p(z_{i,j} = k' \mid \pi_i^{(n)}) p(w_{i,j} \mid \theta_{k'}^{(n)})}$$

and we simply let $q_{z|y}(z_{i,j} = k) = p(z_{i,j} = k \mid y_{i,j} = 1, \Theta^{(n)}, \Pi^{(n)})$.

M-step: We now need to re-estimate the parameters for our model $\Theta^{(n+1)}$ and $\Pi^{(n+1)}$ using the distributions q_y and $q_{z|y}$ that we estimated in the E-step. We again will take an "expected counts" view. Let $n_{d,k}$ be the number of times we expect to see a word in document d assigned to topic k from the PLSA mixture model, and let $n_{w,k}$ be the number of times we expect to see a specific word type w assigned to topic k from the PLSA mixture model.

We have

$$n_{d,k} = \sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)$$

and

$$n_{w,k} = \sum_{d \in D} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k).$$

Let's look at the product $q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)$. We see that

$$\begin{aligned} q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k) &= \left(\frac{(1-\lambda) \sum_{k'=1}^{K} p(z_{d,w} = k' \mid \pi_d^{(n)}) p(w \mid \theta_{k'}^{(n)})}{\lambda p(w \mid D) + (1-\lambda) \sum_{k'=1}^{K} p(z_{d,w} = k' \mid \pi_d^{(n)}) p(w \mid \theta_{k'}^{(n)})} \right) \\ &\quad \times \left(\frac{p(z_{d,w} = k \mid \pi_d^{(n)}) p(w \mid \theta_k^{(n)})}{\sum_{k'=1}^{K} p(z_{d,w} = k' \mid \pi_d^{(n)}) p(w \mid \theta_{k'}^{(n)})} \right) \\ &= \frac{(1-\lambda) p(z_{d,w} = k \mid \pi_d^{(n)}) p(w \mid \theta_k^{(n)})}{\lambda p(w \mid D) + (1-\lambda) \sum_{k'=1}^{K} p(z_{d,w} = k' \mid \pi_d^{(n)}) p(w \mid \theta_{k'}^{(n)})} \end{aligned}$$

which can be used to simplify the computation of the expected counts[8].

Finally, we can normalize the expected counts to come up with the new estimates of our model's parameters. Specifically,

$$p(z = k \mid \pi_d^{(n+1)}) = \frac{n_{d,k}}{\sum_{k'=1}^{K} n_{d,k'}} = \frac{\sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)}{\sum_{k'=1}^{K} \sum_{w \in d} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k')}$$

and

$$p(w \mid \theta_k^{(n+1)}) = \frac{n_{w,k}}{\sum_{w' \in V} n_{w',k}} = \frac{\sum_{d \in D} c(w,d) q_y(y_{d,w} = 1) q_{z|y}(z_{d,w} = k)}{\sum_{w' \in V} \sum_{d \in D} c(w',d) q_y(y_{d,w'} = 1) q_{z|y}(z_{d,w'} = k)}.$$