HW #4

Doris Jung-Lin Lee (jlee782@illinois.edu)

1) Query likelihood model:

$$score(Q,D) = \sum logp(q_i|d)$$

$$= \sum c(w_i,q)logp(w_i|\theta_D)$$

We know that : $c(w_i,Q) = |Q|p(w|\theta_Q)$

$$score(Q,D) = \sum |Q|p(w|\theta_Q)logp(w_i|\theta_D)$$

$|Q|$ is a constant for a given query

$$score(Q,D) = \sum p(w|\theta_Q)logp(w_i|\theta_D)$$

$$= -D(\theta_Q||\theta_D)$$

$\therefore$ KL divergence covers query likelihood estimate as a special case when we set $p(w|\theta_Q) = \frac{c(w_i,Q)}{|Q|}$

2) a) Multinomial distribution

$$p(q|d) = \prod p(w_i = x_i|d)$$

Since the probability of each word is multiplied together, both w_1 and w_2 must occur. $\therefore$ conjuctive

b) We can extend the basic multinomial query likelihood model since we know that the subqueries have and AND relation (product of probabilities) and the words in the query terms have a OR relation (which means that the probabilities should be summed up)

$$p(Q|d) = \prod_{i=1}^{k} p(q_i = Q_i|d)$$

$$= \prod_{i=1}^{k} \sum_{j=1}^{n_i} p(w_i = x_i|d)$$

3)a)

$$p(\text{“}the\text{”}) = p(w = \text{“}the\text{”}|H) + p(w = \text{“}the\text{”}|T)$$

$$= 0.8 * 0.3 + 0.3 * 0.2 = \boxed{0.3}$$

b) It doesn't matter whether its the 2nd word or the 1st. Since we do not have constraint on what the 1st word must be, p=1 for whatever happens to the first word. $\therefore$ $\boxed{\text{p(w\_2=“the”)} = 0.3}$

c)

$$p(H|w = \text{“}data\text{”}) = \frac{p(w = \text{“}data\text{”}|H)}{p(w = \text{“}data\text{”})}$$

$$= \frac{p(w = \text{“}data\text{”}|H)}{p(w = \text{“}data\text{”}|H) + p(w = \text{“}data\text{”}|T)}$$

$$= \frac{0.1 * 0.8}{0.1 * 0.8 + 0.1 * 0.2} = \boxed{0.8}$$

d) Compare p(w):

$$p(w) = p(w|H) + p(w|T)$$

We expect that p(w="data") is lowest since both p(w|H), p(w|T) is lowest among the 5 words. So the word "data" is expected to occur least frequently.

e) From the data, we observe:

p("the") = 3/10
p("computer") = 3/10
p("data") = 2/10
p("game") = 2/10

Given that we know it was written only by H, we can then conclude that p("computer"|H) = 0.3 and p("game"|H) = 0.2.

4) a)

$$p(w) = (1 - \lambda)p(w|C) + \lambda p(w|\theta_1)$$

$$= (1 - \lambda)p(w|C) + \lambda \frac{c(w, D_1)}{|D_1|}$$

b) Let z_w be a binary latent variable that indicates which distribution w is drawn from (where $z_w = 0$ if background; $z_w = 1$ if from θ_1)

Our p(w) is piecewise, we are interested in the first term if z_w =0 , and we want the second term if z_w =1.

$$p(w) = [(1-\lambda)p(w|C)]^{(1-z_w)} + \left[\lambda \frac{c(w, D_1)}{|D_1|}\right]^{z_w}$$

$$logp(w) = \sum[(1-z_w)(1-\lambda)p(w|C)] + \left[z_w \lambda \frac{c(w, D_1)}{|D_1|}\right]$$

c) We need the hidden variable z_w which determines whether a word w is from the collection C or from the topic θ_1. We use $p(w|\theta_1)$ to estimate $p(z_w = 1)$ and vice versa. We have 2 documents, so there is a total of 2k parameters.

d) We can derive the EM updating formulas for $p(z_w)$ and $p^{(n+1)}(w|\theta_1)$ by either using the Lagrange multiplier approach or by minimizing the KL divergence between a lower bound function F with $p(w|\theta)$. Here we take the latter approach:

$$F(q, \theta_1) = \sum q(z) log \frac{p(z_w|\theta_1)}{q(z)}$$

where q(z) is the distribution over the binary latent variable z_w.

E step: We fix the model parameters and maximize F.

$$p(z_w = 1) = \frac{\lambda p^{(n)}(w|\theta_1)}{\lambda p^{(n)}(w|\theta_1) + (1-\lambda)p(w|C)}$$

$p(w|\theta_1) = \frac{c(w,D_1)}{|D_1|}$

$$p(z_w = 1) = \frac{\lambda p^{(n)}(w|\theta_1)}{\lambda \frac{c^{(n)}(w,D_1)}{|D_1|} + (1-\lambda)p(w|C)}$$

M-step: We fix the latent variable distribution q and maximize the log likelihood directly. The update equation is then:

$$p^{(n+1)}(w|\theta_1) = \frac{c(w, D_1)p(z_w = 1)}{\sum_{w \in V} c(w, D_1)p(z_{w'} = 1)}$$

where the numerator denotes the count/number of times we expect that w is drawn from the θ_1 distribution and the fraction is normalized by the term in the denominator which is the sum over that expectation for all words in document D_2.

We keep repeating E and M step until convergence, and estimate λ by :

$$\hat{\lambda} = argmax[p(z_w = 1)]$$