University of Illinois at Urbana-Champaign

Midterm Examination Practice

CS598CXZ Advanced Topics in Information Retrieval (Fall 2013)

Professor ChengXiang Zhai

1. **Basic IR evaluation measures:** The following table shows the search results on the first page of a Web search engine along with binary relevance judgments where a "+" (or "-") indicates that the corresponding document is relevant (or non-relevant). Suppose there are 20 relevant documents in total in the whole collection.

Document ID	Relevance Judgment
D1	+
D2	+
D3	+
D4	-
D5	-
D6	-
D7	-
D8	-
D9	-
D10	+

Compute the following measures for this list of retrieval results (no need to reduce an expression to a decimal or the simplest form of a fraction):

(a) Precision =

(b) Recall =

(c) Average Precision =

(d) Precision at 5 documents =

(e) F1 =

2. **Pooling:** Suppose we use the pooling strategy to create a test collection for evaluating forum search, and we have 5 different systems to contribute to our pool. If we judge the top 10 documents from each of the systems, what would be the maximum possible number and minimum possible number of documents that the assessors would have to judge?

3. **Bayes Rule**

Helen and Tom are trying to surf the Web to find some ideas about what they are going to do together this weekend. They decided to use Google with a query written collectively in the following way:

(a) Each query word is written independently.

(b) When writing a word, they would first toss a coin to decide who will write the word. The coin is known to show up as HEAD 80% of the time. If the coin shows up as HEAD, then Helen would write the word, otherwise, Tom would write the word.

(c) If it is Helen's turn to write, she would "write" the word by simply drawing a word according to multinomial word distribution θ_H. Similarly, if it is Tom's turn to write, he would "write" the word by drawing a word according to multinomial word distribution θ_T.

Suppose the two distributions θ_H and θ_T are defined as follows:

Word w	$p(w\|\theta_H)$	$p(w\|\theta_T)$
the	0.3	0.3
football	0.1	0.2
game	0.1	0.1
opera	0.2	0.1
music	0.2	0.1
...	...	...

(a) What is the probability that they would write "the" as the first word of the query? Show your calculation.

(b) Suppose we observe that the first word they wrote is "music", what is the probability that this word was written by Helen (i.e., $p(\theta_H|w ='' music'')$)? Show your calculation (providing a formula is sufficient).

(c) Suppose we don't know θ_H, but observed a query Q known to be written *solely* by Helen. That is, the coin somehow always showed up as HEAD when they wrote the query. Suppose $Q =$"the opera music the music game the opera music game" and we would like to use the maximum likelihood estimator to estimate multinomial word distribution θ_H. Fill in the values for the following estimated probabilities:

Word w	$p(w\|\theta_H)$
the	
football	
game	
opera	
music	

4. **Probability ranking principle:** One assumption made by the probability ranking principle is that the relevance (or usefulness) of a document is independent of that of other documents that a user may have already seen. Give an example/scenario of search results that are optimal according to the probability ranking principle, but not ideal from a user's perspective.

5. **Vector space model:** Point out three most important reasons why the following vector-space retrieval function is unlikely effective:

$$score(Q, D) = \sum_{w \in Q, w \in D} c(w, Q) * c(w, D) * \log \frac{N_w + 1}{M + 1}$$

where $c(w, Q)$ and $c(w, D)$ are the counts of word w in query Q and document D, respectively, N_w is the number of documents in the collection that contain word w, and M is the total number of documents in the whole collection.

6. **Robertson-Sparck-Jones (RSJ) model:** The following equations show the initial steps in deriving the RSJ model based on probability ranking principle. $O(R = 1|Q, D)$ denotes the odds ratio that document D is relevant to query Q.

$$
\begin{aligned}
O(R = 1|Q, D) &= \frac{p(R = 1|Q, D)}{p(R = 0|Q, D)} && (1)\\
&= \frac{p(Q, D|R = 1)p(R = 1)}{p(Q, D|R = 0)p(R = 0)} && (2)\\
&\propto \frac{p(Q, D|R = 1)}{p(Q, D|R = 0)} && (3)\\
&= \frac{p(D|Q, R = 1)p(Q|R = 1)}{p(D|Q, R = 0)p(Q|R = 0)} && (4)\\
&\propto \frac{p(D|Q, R = 1)}{p(D|Q, R = 0)} && (5)\\
& && (6)
\end{aligned}
$$

According to the probability ranking principle, we would like to rank documents based on $p(R = 1|Q, D)$, but this is equivalent to ranking documents based on $O(R = 1|Q, D)$, so we started the derivation with $O(R = 1|Q, D)$, instead of $p(R = 1|Q, D)$. Applying Bayes rule to both the numerator and the denominator allows us to obtain Equation (2) from Equation (1).

(a) What is the justification for going from Equation (2) to Equation (3)?

(b) What is the justification for going from Equation (4) to Equation (5)?

(c) Now suppose our vocabulary has just four words: w_1, w_2, w_3, w_4. Using the Bernoulli model, we would model the presence and absence of each word in a document D. Thus we can represent document D as a bit vector (d_1, d_2, d_3, d_4), where d_i indicates whether w_i occurs in document D. Write the bit vector for document $D_0 = w_1w_2$, i.e., D_0 just contains two words.

(d) Let $p_i = (d_i = 1|Q, R = 1)$ be the probability of seeing w_i in a relevant document and $q_i = p(d_i = 1|Q, R = 0)$ be the probability of seeing term w_i in a non-relevant document. Assume again $D_0 = w_1w_2$. Express $\frac{p(D_0|Q,R=1)}{p(D_0|Q,R=0)}$ in terms of q_i's and p_i's.

(e) Suppose we have the following table with examples of relevant and non-relevant documents for query Q where we show the relevance status of each example and whether each word occurs in a document. "Rel" and "NonRel" mean that the corresponding document is relevant or non-relevant, respectively; a value of 1 (or 0) indicates that the corresponding word occurs (or does not occur) in the corresponding document.

Document ID	Relevance Status	w_1	w_2	w_3	w_4
D1	Rel	1	0	1	0
D2	Rel	1	1	1	0
D3	Rel	1	0	1	0
D4	Rel	1	0	0	0
D5	NonRel	1	1	0	1
D6	NonRel	1	1	0	1
D7	NonRel	1	1	1	1
D8	NonRel	1	1	1	1

Write down formulas to show how you can estimate p_1, p_3, p_4 and q_1, q_3, q_4 based on these examples. (You don't need to calculate the exact values for them.)

7. **KL-divergence retrieval model**
   The (negative) KL-divergence retrieval function is

$$score(Q, D) = -\sum_{w \in V} p(w|\theta_Q) \log \frac{p(w|\theta_Q)}{p(w|\theta_D)}$$

   where V is the set of all the words in our vocabulary, and $p(w|\theta_Q)$ and $p(w|\theta_D)$ are query language model and document language model, respectively. Show that this ranking function is equivalent to the basic multinomial query likelihood retrieval function if we estimate the query language model $p(w|\theta_Q)$ based on the empirical word distribution in the query, i.e., $p(w|\theta_Q) = \frac{c(w,Q)}{|Q|}$ where $c(w, Q)$ is the count of word w in query Q, and $|Q|$ is query length.

8. **Query likelihood and Dirichlet prior smoothing**
Let $Q = q_1...q_m$ be a query and D be a document.

(a) Show that if we use the multinomial query-likelihood scoring method (i.e., $p(Q|D) = p(q_1|D)...p(q_m|D)$) and the Dirichlet prior smoothing method, we can rank documents based on the following scoring function:

$$score(Q, D) = \left[\sum_{w \in Q, w \in D} c(w, Q) \log(1 + \frac{c(w, D)}{\mu p(w|C)}) \right] + |Q| \log \frac{\mu}{\mu + |D|}$$

where $c(w, D)$ and $c(w, Q)$ are the counts of word w in D and Q respectively, $|D|$ is the length of document D, $p(w|C)$ is a collection/reference language model, and μ is the smoothing parameter of the Dirichlet prior smoothing method.

(b) Briefly explain why the formula above is more efficient to compute with an inverted index than the original query likelihood formula.

9. **Mixture model**

Instead of modeling documents with multinomial distributions, we may also model documents with multiple multivariate Bernoulli distributions where we would represent each document D as a bit vector indicating whether a word occurs or does not occur in the document. Specifically, suppose our vocabulary set is $V = \{w_1, ..., w_N\}$ with N words. A document D will be represented as $D = (d_1, d_2, ..., d_N)$, where $d_i \in \{0, 1\}$ indicates whether word w_i is observed ($d_i = 1$) in D or not ($d_i = 0$). Suppose we have a collection of documents $C = \{D_1, ..., D_M\}$, and we would like to model all the documents with a mixture model with two multi-variate Bernoulli distributions θ_1 and θ_2. Each of them has N parameters corresponding to the probability that each word would show up in a document. For example, $p(w_i = 1|\theta_1)$ means the probability that word w_i would show up when using θ_1 to generate a document. Similarly, $p(w_i = 0|\theta_1)$ means the probability that word w_i would NOT show up when using θ_1 to generate a document. Thus, $p(w_i = 0|\theta_i) + p(w_i = 1|\theta_i) = 1$. Suppose we choose θ_1 with probability λ_1 and θ_2 with probability λ_2 (thus $\lambda_1 + \lambda_2 = 1$).

(a) Write down the log-likelihood function for $p(D)$ given such a two-component mixture model.

(b) Suppose λ_1 and λ_2 are fixed to constants. Write down the E-step and M-step formulas for estimating θ_1 and θ_2 using the Maximum Likelihood estimator.

scratch