Dragon Star Course Final Examination

Information Retrieval: Foundation and Challenges

Instructor: ChengXiang Zhai (UIUC)
Teaching Assistants: Hongfei Yan (Peking University), Bo Peng (Peking University)

Time: 1:30-4:30pm, June 30, 2008
Place: Room 101, 2nd Classroom Building, Peking University

Name:_________________________

NetID:_____________________

1. [10 points] Evaluation

(a) [3 points] Suppose we have a topic (i.e., a query) with a total of 10 relevant documents in a collection with 100 documents. A system has retrieved 6 documents whose relevance status is

\[+, -, -, +, -, +\]

in the order of ranking. A “+” (or “-“) indicates that the corresponding document is relevant (or non-relevant). For example, the first document is relevant, while the second is non-relevant, etc. Compute the precision, recall, and the (non-interpolated) mean average precision for this result.

(b) [4 points] Now suppose our collection has a total of 10,000 documents, instead of 100 documents, but the topic has the same number of relevant documents (i.e., 10) in the collection, and the system has retrieved exactly the same results as in the previous question for this topic. Intuitively, the system’s performance is better in this case than in the previous case since now we have more distracting non-relevant documents and thus the retrieval task is much harder. Explain why precision and recall can not reflect this difference, and propose a reasonable new measure that would reflect this difference. How important is it to capture this difference in our evaluation of retrieval methods?
(e) [3 points] When reusing a retrieval test collection created using the “pooling” strategy to evaluate new retrieval functions that have not contributed to the “pool”, why is there a potential bias against the new functions?
2. **[20 points] Global Corpus Statistics**

Consider a document collection $C_0 = \{D_1, ..., D_N\}$, where $D_i$ is a document. We construct two other collections $C_1$ and $C_2$ in the following way:

1. $C_1$ is formed by deleting $D_1$ from $C_0$, thus it has $N - 1$ documents.
2. $C_2$ is formed by replacing $D_1$ in $C_0$ with a new document $D_1'$, which contains all the distinct terms in $D_1$ but all the terms only occurs once. $C_2$ thus has the same number of documents as $C_0$.

Let $t$ be a term occurring more than once in $D_1$ (i.e. $c(t, D_1) > 1$) and $IDF(t, C_i)$ be the IDF value of $t$ in collection $C_i$. Let $\theta(B(C_i))$ be a collection language model estimated based on collection $C_i$.

(a) **[5 points]** What can we say about $IDF(t, C_1)$ and $IDF(t, C_0)$? Are they the same? If not, which is higher? Why?

(b) **[5 points]** What can we say about $IDF(t, C_2)$ and $IDF(t, C_0)$? Are they the same? If not, which is higher? Why?

(c) **[5 points]** Is $p(t | \theta(C_1))$ higher than, lower than, or equal to $p(t | \theta(C_0))$? Why?

(d) **[5 points]** Is $p(t | \theta(C_2))$ higher than, lower than, or equal to $p(t | \theta(C_0))$? Why?
3. [15 points] Entropy and Bayes Rule

Consider the following structurally ambiguous sentence

*John saw a student with a telescope.*

The prepositional phrase (PP) “with a telescope” can be attached to either the verb “saw” or the noun “student”. One approach to resolving such an ambiguity (called “PP-attachment”) is to treat it as a statistical classification task. The instance to classify is the PP “with a telescope”, and the class labels are “attached to the verb” and “attach to the noun” respectively. We will represent the PP with two features that may help predict the class label: (1) the preposition (i.e., “with” in this case); (2) the head noun (i.e., “telescope” in this case). For example, we know that if the preposition is “of” then the PP should almost always be attached to the noun. Similarly, if the word “telescope” were replaced by “book”, then the PP “with a book” is also unlikely attached to the verb. Suppose we consider only two prepositions “with” and “of”, and only two nouns “telescope” and “book”. We thus deal with three binary random variables:

- Class label: \( C \in \{0, 1\} \) with \( C = 1 \) (or \( C = 0 \)) indicates that the PP should be attached to the verb (or the noun).
- Preposition feature: \( F_p \in \{\text{with}, \text{of}\} \) indicates which preposition is used in the PP under consideration.
- Head noun feature: \( F_n \in \{\text{telescope}, \text{book}\} \) indicates which noun is the head noun in the PP under consideration.

Suppose we have 10 example sentences for which we know the correct PP attachment. Their class labels and feature values are show in the following table:

<table>
<thead>
<tr>
<th>PPAttachedTo (C)</th>
<th>PrepositionFeature (F_p)</th>
<th>Head noun feature (F_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>with</td>
<td>telescope</td>
</tr>
<tr>
<td>Verb</td>
<td>with</td>
<td>telescope</td>
</tr>
<tr>
<td>Verb</td>
<td>with</td>
<td>book</td>
</tr>
<tr>
<td>Verb</td>
<td>of</td>
<td>telescope</td>
</tr>
<tr>
<td>Verb</td>
<td>with</td>
<td>telescope</td>
</tr>
<tr>
<td>Noun</td>
<td>of</td>
<td>book</td>
</tr>
<tr>
<td>Noun</td>
<td>of</td>
<td>book</td>
</tr>
<tr>
<td>Noun</td>
<td>of</td>
<td>book</td>
</tr>
<tr>
<td>Noun</td>
<td>of</td>
<td>telescope</td>
</tr>
<tr>
<td>Noun</td>
<td>of</td>
<td>book</td>
</tr>
</tbody>
</table>

(a) (4 points) Use the maximum likelihood estimator to compute the following probabilities:

- \( p(C = 1) = \)
- \( p(F_p = \text{with}) = \)
- \( p(F_p = \text{with}|C = 1) = \)
- \( p(F_p = \text{with}, F_n = \text{telescope}|C = 1) = \)

(b) (3 points) Consider the entropies of the two binary random variables \( F_p \), and \( F_n \) (i.e., \( H(F_p), H(F_n) \)). Without calculating the entropies, can you tell which is larger? Why?
(e) (3 points) Now consider the three conditional distributions $p(F_p|C = 1)$, $p(F_n|C = 1)$, and $p(F_p|C = 0)$. We may also compute their entropies (i.e., $H(F_p|C = 1)$, $H(F_n|C = 1)$, $H(F_p|C = 0)$). Which of the three entropies is the smallest? Why?

(d) (5 points) Given a new sentence $Q$ = “John saw the price of a book” and assuming that the observation of both $F_p$ and $F_n$ is independent of each other given a particular PP attachment, can you use Bayes Rule to infer whether the preposition “of a book” should be attached to the verb “saw” or the noun “price”? Show your calculation.
4. [25 points] Dirichlet prior smoothing and retrieval

(a) [5 points] Let \( q = q_1 \ldots q_m \) be a query and \( d \) be a document and \( p(q_i|d) \) be the probability of query word \( q_i \) according to a smoothed document language model estimated based on \( d \). Does the formula \( p(q|d) = \sum_{i=1}^m p(q_i|d) \) correctly describe the query likelihood retrieval model? If not, write down the correct formula.

Suppose we have a document collection with an extremely small vocabulary of only 8 words \( w_1, \ldots, w_8 \). The following table shows the estimated reference language model \( p(w|REF) \) using the whole collection (2nd column) and the word counts \( c(w, d) \) in document \( d \) (3rd column). The 4th and 5th columns are two unigram language models for document \( d \) estimated using the unsmoothed maximum likelihood estimator and Dirichlet prior smoothing (with parameter \( \mu \)), respectively.

| Word | \( p(w|REF) \) | \( c(w, d) \) | \( p_{ml}(w|d) \) | \( p_\mu(w|d) \) |
|------|----------------|-------------|-----------------|-----------------|
| \( w_1 \) | 0.3 | 2 | | |
| \( w_2 \) | 0.15 | 1 | | |
| \( w_3 \) | 0.1 | 2 | | 0.125 |
| \( w_4 \) | 0.1 | 4 | | |
| \( w_5 \) | 0.05 | 1 | | |
| \( w_6 \) | 0.1 | 0 | | |
| \( w_7 \) | 0.1 | 0 | | |
| \( w_8 \) | 0.1 | 0 | | |

(b) [5 points] Fill in the values for unsmoothed maximum likelihood probabilities \( p_{ml}(w|d) \) for all the eight words (column 4).

(c) [5 points] Column 5 is the probability of a word after applying Dirichlet prior smoothing with prior sample size parameter \( \mu \). In the table, only the smoothed probability for word \( w_3 \) is shown, which is 0.125. What is the value of \( \mu \)?

(d) [5 points] For the rest seven words on column 5, without actually computing the smoothed probability values, decide whether the smoothed probability is larger than, equal to, or smaller than the original unsmoothed maximum likelihood estimate on column 4. Use one of \( \{ >, =, < \} \) to finish the empty cells in column 5.

(e) [5 points] What condition should \( c(w, d) \) satisfy so that the smoothed probability of word \( w \) would always be the same as the unsmoothed maximum likelihood estimate regardless what value \( \mu \) is?
5. **[10 points] Feedback**

Web search engines can record which pages in the search results are viewed and which pages are skipped by a user. Such information is often called “clickthroughs”. Over time, we can thus collect a set of tuples of the following form:

- (query1, URL1, skipped)
- (query1, URL2, viewed)
- (query1, URL3, viewed)
  ...
- (query2, URL4, skipped)
- (query2, URL2, viewed)
  ...
- (query3, URL5, skipped)
- (query3, URL3, skipped)
  ...
- (query1, URL1, skipped)
- (query1, URL2, skipped)
- (query1, URL3, viewed)
  ...

Suppose a current user has just entered a query which is identical to “query1”. A straightforward way to use the past clickthrough data to improve ranking of pages for this query is to retrieve all the clickthroughs for “query1” from the query log as shown above and then simply rank the URLs that are viewed frequently. However, this strategy does not take advantage of the clickthroughs of related queries. For example, if “query1” and “query2” are similar, the clickthroughs for “query2” may also be potentially interesting to the user who entered “query1”. Can you propose an algorithm to further improve performance by exploiting the clickthroughs of such related queries? Briefly sketch the algorithm, preferably with some formulas.
6. **[10 points] Mixture models**

Suppose a mixture model for documents has two component unigram language models \( \theta_1 \) and \( \theta_2 \) and the mixing coefficients are \( \lambda_1 \) and \( \lambda_2 \) for \( \theta_1 \) and \( \theta_2 \), respectively. Of course \( \lambda_1 + \lambda_2 = 1 \). Consider the following two different ways to “generate” a document \( D = d_1 d_2 \ldots d_n \).

- **Model 1**
  \[
p(D|\theta_1, \theta_2, \lambda_1, \lambda_2) = \lambda_1 \prod_{i=1}^{n} p(d_i|\theta_1) + \lambda_2 \prod_{i=1}^{n} p(d_i|\theta_2)
\]

- **Model 2**
  \[
p(D|\theta_1, \theta_2, \lambda_1, \lambda_2) = \prod_{i=1}^{n} (\lambda_1 p(d_i|\theta_1) + \lambda_2 p(d_i|\theta_2))
\]

Given a collection of \( m \) documents \( C = \{D_1, \ldots, D_m\} \) generated using such a mixture model, we often use the EM algorithm to estimate all the parameters of our mixture model, in which we augment the observed data with hidden variables. Each hidden variable is binary and indicates whether \( \theta_1 \) or \( \theta_2 \) is chosen. In which case, Model 1 or Model 2, would we need to introduce more *distinct* hidden variables? How many distinct hidden variables do we introduce exactly for estimating Model 1?
7. **[10 points] Open question (proximity)**

When two documents match the same number of query terms, intuitively we should prefer a document in which all the matches are close to each other to a document in which the matches are far away (e.g., one at the very beginning of the document and one at the end) because in the latter case, the matches may be due to chances as opposed to a coherent discussion of the topic. Can you propose a way to extend the query likelihood retrieval function to incorporate such a proximity heuristic?
scratch