Midterm Examination

CS498-CXZ Introduction to Text Information Systems
Professor ChengXiang Zhai

Time: 12:35–1:15pm, Oct. 12, 2004
Place: Room 1105, Siebel Center

Name:________________________

NetID:____________________
1. [6 points] Decide whether the following statements are true. Circle the answer.

(a) Removing stop words always reduces the size of the inverted index. true  false
(b) Relevance feedback improves average retrieval precision for every query. true  false
(c) In the classic probabilistic retrieval model (i.e., Robertson Sparck-Jones model), a query and a document are treated differently. true  false
(d) For any retrieval method, precision at 20 documents can never be higher than precision at 10 documents. true  false
(e) When there are multiple local maxima, because the EM algorithm may not always converge, it will not necessarily find the global maxima. true  false
(f) An inverted index makes it possible to quickly find all the documents containing a particular term, but it makes it impossible to find all the terms contained in any given document. true  false

2. [4 points] Logistic Regression

Briefly explain why logistic regression is better than linear regression for retrieval (i.e., modeling $p(R = 1|Q, D)$ with features defined on $Q$ an $D$).

3. [10 points] Retrieval Evaluation

(a) [5 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 8 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 two documents are relevant, while the third is non-relevant, etc. Compute the precision, recall, and the (non-interpolated) mean average precision for this result.

(b) [5 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 nor reflect this difference, and propose a new measure that would reflect this difference.
4. [20 points] Statistical Disambiguation 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 three prepositions “in”, “with”, and “of”, and only three nouns “telescope”, “book”, “UIUC”. We thus deal with three 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: $F_p \in \{\text{with, of}\}$ indicates which preposition is used in the PP under consideration.
- Head noun feature: $F_n \in \{\text{telescope, UIUC}\}$ indicates which noun is the head noun in the PP under consideration.

We observe values of $F_p$ and $F_n$, and want to infer the value for $C$. That is, we want to compute $p(C|F_p, F_n)$. We model the joint distribution of all three variables using the decomposition $p(F_p, F_n, C) = p(F_p, F_n|C)p(C)$, i.e., to generate values for $F_p$, $F_n$, and $C$, we first sample a value for $C$ according to $p(C)$, and then sample values for $F_p$ and $F_n$ according to $p(F_p, F_n|C)$ based on the value of $C$ that we obtained.

Suppose some of the conditional probabilities $p(F_p, F_n|C)$ and prior probability $p(C)$ are known as shown below.

\[
p(C = 0) = 0.5 \\
p(F_p = \text{with}, F_n = \text{telescope}|C = 0) = 0.08 \\
p(F_p = \text{with}, F_n = \text{UIUC}|C = 0) = 0.22 \\
p(F_p = \text{of}, F_n = \text{telescope}|C = 0) = 0.32 \\
p(F_p = \text{with}, F_n = \text{telescope}|C = 1) = 0.62 \\
p(F_p = \text{with}, F_n = \text{UIUC}|C = 1) = 0.18 \\
p(F_p = \text{of}, F_n = \text{telescope}|C = 1) = 0.08
\]

(a) [6 points] Compute the following probabilities using the basic probability rules:

\[
p(F_p = \text{of}|C = 0) = \\
p(F_p = \text{with}|F_n = \text{telescope}, C = 0) = \\
p(C = 0|F_p = \text{of}) =
\]

(b) [6 points] Are the two variables $F_p$ and $F_n$ conditionally independent given $C$?
(c) [8 points] Use Bayes rule to compute \( p(C = 0 | F_p = of, F_n = UIUC) \) and \( p(C = 1 | F_p = of, F_n = UIUC) \) and decide whether to attach the PP “of UIUC” to the verb or the noun in the sentence “John saw a student of UIUC”. How do you make the PP attachment decision if we do not observe the head noun “UIUC” (i.e., “John saw a student of ______")?

5. [50 points] Dirichlet prior smoothing and retrieval

Suppose we have a document collection with an extremely small vocabulary with only 8 words \( w_1, \ldots, w_8 \). The following table shows the estimated reference language model \( p(w|REF) \) using the whole collection of documents (2nd column) and the word counts in a document \( d \) (3rd column). \( c(w, d) \) is the count of word \( w \) in document \( d \). 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 | \( p(w|REF) \) | \( c(w, d) \) | \( p_{ml}(w|d) \) | \( p_{p}(w|d) \) |
|------|---------------|--------------|-----------------|---------------|
| \( w_1 \) | 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            |                 |               |

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

(b) 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.

i. [5 points] What is the value of \( \mu \)?

ii. [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.

(c) [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?
(d) **[15 points]** Let \( q_1 = w_3 \) be a query. If we use Dirichlet prior smoothing method (with \( \mu = 10 \)) to smooth the document language model for document \( d \), what is the probability of \( q_1 \) according to this smoothed language model? If we increase the value of \( \mu \), would the probability of \( q_1 \) become larger, smaller, or change in an unpredictable way?

(e) Suppose there is another document \( d' \) which is identical to document \( d \) except that one occurrence of \( w_1 \) has been changed to word \( w_4 \), thus giving \( c(w_1, d') = 1 \) and \( c(w_4, d') = 5 \). Let \( q_3 = w_1 w_4 \) be a query.

i. **[5 points]** Using the query likelihood retrieval method with Dirichlet prior smoothing, which of the two documents \( d \) and \( d' \) would be ranked higher? Does smoothing affect the ranking of \( d \) and \( d' \)?

ii. **[5 points]** Suppose we use the following simplest TF-IDF retrieval formula in which neither TF normalization nor document length normalization is performed:

\[
\text{score}(D, Q) = \sum_{w \in Q \cap D} c(w, D) \text{IDF}(w)
\]

Assuming that \( \text{IDF}(w_4) > \text{IDF}(w_1) \) since \( p(w_1 | \text{REF}) > p(w_4 | \text{REF}) \) (i.e., \( w_4 \) is more discriminative/rare than \( w_1 \)). According to this method, which of the two documents \( d \) and \( d' \) would be ranked higher?
iii. [5 points] In this particular case (i.e., query $q_3$, document $d$ and $d'$), which method seems to make more sense intuitively? Dirichlet prior smoothing or the simplest TF-IDF formula? Briefly explain why.

6. [10 points] Compression and Information Theory

(a) [2 points] Write down the gamma code for the integer 9.

(b) [3 points] Suppose the integers we want to encode are uniformly distributed, which of gamma-coding, binary-coding, and delta-coding would you recommend for compression? Why?