

CS 276B: Text Information Retrieval, Mining, and Exploitation
Open Book Midterm Examination
Thursday, February 6, 2003

This exam consists of 10 pages, 7 questions, and 75 points. We would like you to write your answers on the exam paper, in the spaces provided. To give you plenty of room, some pages are largely blank. If there isn't sufficient room, write on the back of a page, but please put an arrow or PTO on the front to tell us to look there. You have 75 minutes to complete the exam. Exams turned in after the end of the examination period will either be penalized or not graded at all.

Stanford University Honor Code:

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

Name (printed): _____

Signature: _____ SUID: _____

Question	Possible	Score
1 K-means	5	
2 Cluster	12	
3 NB Travel	12	
4 Smooth	10	
5 NB XOR	8	
6 Features	12	
7 IE	16	
Total	75	

The standard of academic conduct for Stanford students is as follows:

A. The Honor Code is an undertaking of the students, individually and collectively:

1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
2. that they will do their share and take an active part in seeing to it that they as well as others uphold the spirit and letter of the Honor Code.

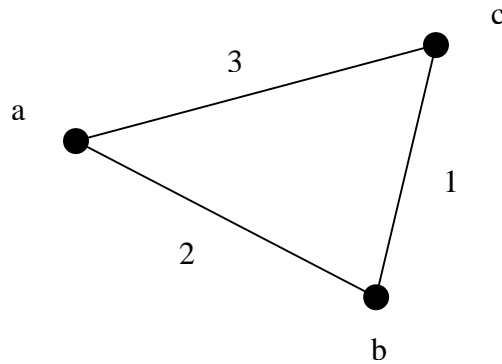
B. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

C. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Part I: Clustering

1. (5 points) Consider a run of $(N-1)$ -means clustering on N documents each represented as a unit vector in term space. Suppose further that all N pairwise cosine distances are different. We run the algorithm with $N-1$ of the input documents as the initial centroids. Give an example in two dimensions to show that the choice of starting centroids will determine the identities of the clusters we end up with.

Assume we have a set of 3 documents, $\{a,b,c\}$ with cosine distances as shown in the diagram. If a and b are chosen as starting centroids then the resulting clusters will be $\{a\}, \{b,c\}$. If b and c are chosen as starting centroids then the resulting clusters will be $\{a,b\}, \{c\}$.



2. (12 points) Consider clustering N points on a line situated at coordinates equaling the powers of two (1, 2, 4 etc). Suppose we use agglomerative clustering by merging the two closest centroids at each step.

a. Write down an expression for the centroid of the largest cluster after i steps.

$$\frac{2^{i+1} - 1}{i + 1}$$

b. Suppose that the cost of a cluster is the sum, over its members, of the distance from each member to the centroid. Write down an expression for the total cost of the clustering after i steps.

$$\sum_{j=0}^i \left| \frac{2^{i+1} - 1}{i + 1} - 2^j \right|$$

c. Suppose that we add a cost for having more clusters. Say, the cost of each distinct cluster is C ; thus, the initial clustering (consisting of N independent points) has cost NC , while after the first step it is $(N-1)C$ because we're down to $N-1$ clusters.

- (i) Write down the total cost of the clustering after step i , including the cluster costs computed in b. above.
- (ii) How many minima does this function have?
- (iii) What is the optimal clustering?

$$(N - i)C + \sum_{j=0}^i \left| \frac{2^{i+1} - 1}{i + 1} - 2^j \right|$$

The first term is linearly decreasing while the second one is monotonically increasing in i , so the expression has one minimum. If C is tiny, the minimum occurs at N clusters. If it's large, it occurs at one cluster.

Part II: Classification

3. (12 points) Consider the problem of classifying the origination point of passenger travel itineraries. Suppose we have the following training set of travel itineraries:

Itinerary	Document	Class
1	"smith: new york - chicago - san francisco - new york"	JFK
2	"chen: san francisco - london - paris - san francisco"	SFO
3	"chen: san francisco - tokyo - singapore- san francisco"	SFO
4	"o'brien: chicago - buenos aires - new york - chicago"	ORD

a. Assume that we use a Bernoulli (i.e., binary) Naive Bayes model. Compute the following feature probabilities:

$$P(X_{\text{francisco}}=\text{true} \mid \text{Class}=\text{SFO}) = 1.0$$

$$P(X_{\text{london}}=\text{true} \mid \text{Class}=\text{SFO}) = 0.5$$

$$P(X_{\text{francisco}}=\text{true} \mid \text{Class}=\text{JFK}) = 1.0$$

b. Assume that we use a multinomial NB model instead. Compute the following probabilities:

$$P(X=\text{francisco} \mid \text{Class}=\text{SFO}) = 4/14 \text{ (assuming no tokenization of punctuation)}$$

$$P(X=\text{london} \mid \text{Class}=\text{SFO}) = 1/14$$

$$P(X=\text{francisco} \mid \text{Class}=\text{JFK}) = 1/8$$

c. Consider a standard Naive Bayes classifier trained on the training set and applied to a similar test set. How accurate is this classifier for:

- (i) the Bernoulli model, and
- (ii) the multinomial model?

- (i) *Not very accurate, because it ignores frequency information, which is important in this domain.*
- (ii) *More accurate, because it uses frequency information. However, it ignore position information, so doesn't distinguish between a city name occurring at the beginning/end of the itinerary from one occurring in the middle*

d. Construct a non-standard feature representation that is 100% accurate for either model.

Use as a feature the term that occurs in the last position of each document.

4. (10 points) This problem concerns smoothing Naïve Bayes classifiers.
- a. Suppose we build a Naive Bayes classifier (multinomial or Bernoulli) with no smoothing of the respective $P(\text{word} \mid \text{class})$ probabilities. If a word was unseen in a class, it will thus have a probability of 0. Describe in words the decision procedure of this classifier (emphasizing the effect of the lack of smoothing, and how its decisions will differ from a smoothed Naive Bayes classifier).

It will never choose a category unless all words in a document were seen for that category for the training set (unless there is no category for which all words were seen, and then all categories are tied for the classifier). It will rank between classes for which all words were seen similarly to the smoothed classifier (but with possible differences due to the smoothing).

- b. Suppose we take a smoothed multinomial classifier and double the amount of smoothing (e.g., for a variant of “add 1 smoothing”, add 2 to each count, and add to the denominator $2k$, where k is the number of samples). What qualitative effect will this have on decisions of the classifier?

It'll be more likely to choose categories for which some/many of the words in the document were unseen.

5. (8 points) Recall that the exclusive-OR function is a function that outputs true if and only if the values of its inputs differ. The outputs of the exclusive-OR function are given in the following input/output table (here we represent true with a 1, and false with a 0):

x_1	x_2	Class
0	0	0
0	1	1
1	0	1
1	1	0

Suppose that we want to use a Naive Bayes classifier to classify samples in accordance with the exclusive-OR function. Assume that we use the feature representation suggested by the table: a two-dimensional feature vector \mathbf{x} , with two possible values (0 and 1) on each dimension.

a. If we train a Naive Bayes classifier on the 4 data points shown in the table using this representation, what is its expected accuracy on new data? Assume that each of the four cases 0/1, 1/0, 0/0, and 1/1 is equally likely to occur in new data, and that if two classes tie in probability, that ties are broken randomly. Accuracy is defined as number of correct decisions divided by all decisions.

Expected accuracy is 1/2 (random) because

$$\begin{aligned}
 &P(\text{class}=0 \mid \text{dim1}=x, \text{dim2}=y) \sim = \\
 &P(\text{dim1}=x, \text{dim2}=y \mid \text{class}=0) P(\text{class}=0) = \\
 &P(\text{dim1}=x \mid \text{class}=0) P(\text{dim2}=y \mid \text{class}=0) P(\text{class}=0) = \\
 &P(\text{dim1}=x \mid \text{class}=1) P(\text{dim2}=y \mid \text{class}=1) P(\text{class}=1) \sim = \\
 &P(\text{class}=1 \mid \text{dim1}=x, \text{dim2}=y)
 \end{aligned}$$

for all x, y .

b. Give as realistic an example as you can of a problem that occurs in text categorization that is similar in nature to the exclusive-OR problem mentioned above, and describe how a Naive Bayes classifier would perform on it.

It's impossible to model a class with one of two different terms, but not both of them. For example, if we wanted to model a class with either the term "operations" or the term "research" but not both terms, we could not do it.

6. (12 points) We want to train a Naive Bayes classifier to determine whether an earnings report originated from the Tokyo stock exchange or not. Earnings reports have the following format:

"[COMPANYNAME] today reported revenues of \$[NUMBER] million for the [ORDINAL] quarter. [COMPANYNAME] is traded on the [STOCKEXCHANGECITY] stock exchange."

Only the variables in square brackets change, all other words are fixed. Possible expansions for [STOCKEXCHANGECITY] are "Tokyo", "London", and "Paris". For example:

"ACME Inc. today reported revenues of \$755 million for the third quarter. ACME is traded on the Tokyo stock exchange."

We want to represent each news story with a single feature. The numerical quantity used for feature selection is either *document frequency* or *chi-square*. Feature selection is performed either on a per category basis or jointly for all categories. Follow how Yang and Pedersen perform joint feature selection for chi-square: they compute chi-square for each category individually, and then select the feature with the maximum chi-square. Thus we are considering four different feature selection strategies as shown in the two tables below.

a. In the table below, enter in each cell the feature that would be selected by the corresponding method for the category "Tokyo", and why. If there are ties, list all tied features. Assume that reports from Tokyo, London and Paris all have the same frequency. Use the following stop list: *ago, for, in, is, of, on, the, to, today*; and assume that there is no other term normalization. You *must* include in the box a qualitative argument as to which feature(s) will be selected to get full credit; you do not need to compute chi square scores for all words.

	Document Frequency	Chi-Square
Per-category Selection (selected feature can be different for each category)	<i>today, reported, revenues, million, quarter, traded, Tokyo, stock, exchange (all tied)</i>	<i>Tokyo</i>
Joint Selection (one feature selected for all categories)	<i>today, reported, revenues, million, quarter, traded, stock, exchange (all tied)</i>	<i>Tokyo, Paris, London (all tied)</i>

b. In the table below, enter in each cell the expected accuracy (number of correct decisions divided by total number of decisions) of the Naive Bayes classifier using the single feature chosen in B by the corresponding feature selection method. If there are ties, list the worst possible accuracy (assuming you always resolve ties adversely).

	Document Frequency	Chi-Square
Per-category Selection (selected feature can be different for each category)	$1/3$	1 (we always get Tokyo/not Tokyo correct using tokyo)
Joint Selection (one feature selected for all categories)	$1/3$	0 (we always get Tokyo/not Tokyo wrong using paris or london)

Part 3: Information Extraction

7. (16 points) Here is a small (valid!) HTML document:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
  <head>
    <title>People</title>
    <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
  </head>

  <body>
    <h1>People</h1>

    <table>
      <tr><th><b>Name</b></th><th>Office</th><th>Phone</th><th>Mail</th></tr>
      <tr><td>Chris Manning</td><td>418</td><td>3-7683</td><td>9040</td></tr>
      <tr><td>Teg Grenager</td><td>454</td><td>1-2345</td><td>9040</td></tr>
      <tr><td>Hector Garcia-Molina</td><td>276</td><td>3-9745</td><td>9025</td></tr>
      <tr><td>Jennifer Widom</td><td>422</td><td>5-4321</td><td>9040</td></tr>
    </table>
  </body>
</html>
```

a. Suppose one wanted to write individual LR (Kushmerick Left Right) context wrappers for each of the fields Name, Office, Phone, Mail. Which ones can you do it for, and which ones can you not? If you can do it, give a LR wrapper that would work, and if you cannot, briefly explain why.

Name: L: <tr><td> R: </td> works
Office: can't distinguish office and phone
Phone: can't distinguish office and phone
Mail: L: <td> R: </td></tr> works

b. Suppose you enhanced the LR wrapper framework so that they also did a regular expression match on the field content. Now, which fields can you do it for, and which ones can you not? If you can do it, give a LR wrapper that would work, and if you cannot, briefly explain why.

Name: as above (filler can be .)*
Office: L: <td> F: [0-9]+ R: </td><td>
Note that you need the trailing <td> to differentiate from Mail
Phone: L: <td> F:[0-9]-[0-9]+ R: </td>
Mail, as above

c. Alternatively (i.e., not using a regular expression match on field content), suppose that one had a notion of relation, and assumed that fields were ordered in the wrapper for a relation. Under the assumption of a known ordering, which fields can you do it for, and which ones can you not? If you can do it, give a LR wrapper that would work, and if you cannot, briefly explain why.

Assume that relation field order is Name, Office, Phone Mail. Then, each wrapper could just be L: <td> R:</td>, though this is maximally dangerous (any inconsistency, and you could get arbitrarily out of sync). Continuing to check for a match on <tr> and </tr> for Name and Mail would be safer.

d. Suppose one ignored all the markup, by replacing all HTML markup tags by whitespace. In other words, the document would now look like this:

```
People
People
Name Office Phone Mail
Chris Manning 418 3-7683 9040
Teg Grenager 454 1-2345 9040
Hector Garcia-Molina 276 3-9745 9025
Jennifer Widom 422 5-4321 9040
```

Now assume that one were using Rapier patterns to do the extraction – but that you are hand-specifying them rather than having a system try to learn suitable patterns. Starting from this input, which fields can you do it for, and which ones can you not? If you can do it, give a Rapier extraction rule that would work, and if you cannot, briefly explain why. To the extent that you have seen in the slides only a rough picture of what Rapier patterns look like, you may need to make some assumptions, but be careful to say clearly what your assumptions are. One thing you may well want to know is that there is only one "number" part-of-speech tag CD and all of 418, 3-7683 and 9040 would be tagged with it. Also, WordNet does not contain multidigit numbers. Finally, assume that line ends aren't represented, but the end of file is as a special token \$EOF\$.

Name: filler: NNP+ (i.e., some number of proper nouns)

Office: prefiller: pos:NNP filler: pos: CD

Phone: It seems not. If you can only specify one item, and then a list which is interpreted as between 0 and N words, then a pattern like:

Prefiller: pos:NNP list: 1 Filler: CD Postfiller: CD

fails because it could still match the room number or the phone number

Mail code is possible (given the suggested treatment of line ends and end of file, with a pattern like this:

Filler: CD Postfiller: { NNP | \$EOF\$}

Disjunctions are specifically licensed in the patterns.