CS371R: Final Exam

Dec. 13, 2024

NAME: _____

UTEID: _____

INSTRUCTIONS:

- This exam has 8 problems and 15 pages. Before beginning, check that your exam is complete.
- You have 2 hours to complete the exam.
- The exam is closed book, closed notes, and closed computer, except for a scientific calculator and the provided equation sheets.
- Mark your answers **on the exam itself**. We will not grade answers on scratch paper.
- Make sure that your answers are legible and your handwriting is dark. We will be scanning the exams and grading them using Gradescope.
- In order to maximize your chance of getting partial credit, show all of your work and intermediate results.

Final grades will be available on Canvas on or before December 17.

Thank you for a great semester! Good luck and have a good break!

1. (8 points) Assume that the total number of documents in a corpus is 20,000 and that the following words occur in the following number of documents:

   - "and" occurs in 19,500 documents
   - "at" occurs in 18,000 documents
   - "Austin" occurs in 100 documents
   - "of" occurs in 18,000 documents
   - "Texas" occurs in 500 documents
   - "the" occurs in 19,500 documents
   - "state" occurs in 15,000 documents
   - "University" occurs in 10,000 documents

   You are given the following list of stop words:

   - "and"
   - "at"
   - "of"
   - "the"

   Assuming that term frequencies are normalized by the maximum frequency in the given document, calculate the TF-IDF weighted term vector for the following simple document:

   "University of Texas at Austin and the state of Texas"

   Perform stop word removal and order tokens in the vector alphabetically.

2. (12 points) Consider the following pages and the set of web pages that they link to:

   Page A points to page C.
   Page B points to page C.
   Page C points to page D.

   Consider running the PageRank algorithm on this subgraph of pages. Assume $\alpha = 0.15$. Simulate the algorithm for three iterations. Show the page rank scores for each page twice for each iteration, both before and after normalization, order the elments in the vectors in the sequence: A, B, C, D.

   (a) Show work for iteration 1 below and fill this table:

Page Rank	A	B	C	D
Before norm				
After norm				

(b) Show work for iteration 2 below and fill this table:

Page Rank	A	B	C	D
Before norm				
After norm				

(c) Show work for iteration 3 below and fill this table:

Page Rank	A	B	C	D
Before norm				
After norm				

3. (15 points) Consider examples described using the following three binary-valued features:

```
A:  0,1
B:  0,1
C:  0,1
```

Show the trace of a perceptron learning from the training set of [A,B,C] examples:

```
[0,0,0]: positive
[0,1,1]: positive
[1,1,1]: negative
```

Assume all of the weights and the threshold start at 0 and the learning rate is 1 and that during each epoch the examples are processed in the exact order given above. Show the weight vector (in the order [A,B,C]) and the threshold after every example presentation. The procedure should converge after only 4 epochs. **NOTE:** For the purposes of this problem, assume that the net input must be **strictly greater than** the threshold in order output a 1 (so this is slightly different from the equations in neural-net lecture slides).

If represented as a logical rule, what is the function learned?

(a) Iteration 1:

Example	Classification (right/wrong)	W(A)	W(B)	W(C)	Threshold
[0,0,0]					
[0,1,1]					
[1,1,1]					

(b) Iteration 2:

Example	Classification (right/wrong)	W(A)	W(B)	W(C)	Threshold
[0,0,0]					
[0,1,1]					
[1,1,1]					

(c) Iteration 3:

Example	Classification (right/wrong)	W(A)	W(B)	W(C)	Threshold
[0,0,0]					
[0,1,1]					
[1,1,1]					

(d) Iteration 4:

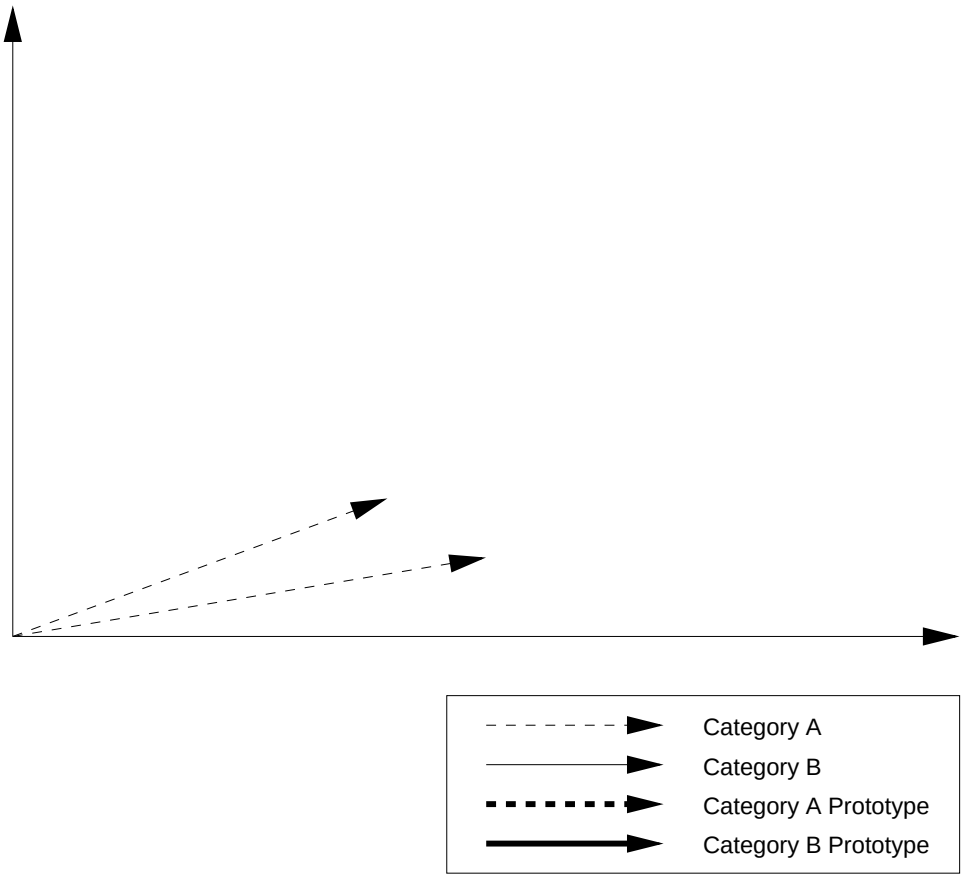
Example	Classification (right/wrong)	W(A)	W(B)	W(C)	Threshold
[0,0,0]					
[0,1,1]					
[1,1,1]					

What logical rule is learned for this concept?

4. (9 points)

(a) Assuming a two-dimensional vector space and the use of cosine similarity as the similarity metric, graphically illustrate why the Rocchio text categorization algorithm does not guarantee a consistent hypothesis (i.e. a hypothesis which correctly classifies all of the training examples).

More specifically, in the following graph, draw two document vectors for category B. Then using the document vectors in the graph, construct the prototype vectors for both categories A and B, which must be drawn **much thicker** than the document vectors. Show that at least one of the training documents would be misclassified.



(b) Does the 3 Nearest-Neighbor algorithm guarantee a consistent hypothesis (one that fits all of the training data)? Why or why not?

5. (14 points) Assume we want to categorize computer-science documents into the following categories: Systems, Theory, AI. Consider performing naive Bayes classification with a simple model in which there is a binary feature for each significant word indicating its presence or absence in the document. The following probabilities have been estimated by analyzing a corpus of preclassified training documents.

c	Systems	Theory	AI
$P(c)$	0.35	0.40	0.25
$P(\text{theorem}\|c)$	0.05	0.8	0.10
$P(\text{search}\|c)$	0.30	0.40	0.60
$P(\text{heuristic}\|c)$	0.05	0.01	0.50
$P(\text{disk}\|c)$	0.30	0.02	0.01
$P(\text{data}\|c)$	0.50	0.01	0.20

Assuming the probability of each evidence word is independent given the category of the text, compute the posterior probability for *each* of the possible categories for each of the following short texts. Assume the categories are disjoint and complete for this application. Ignore any words that are not in the table.

(a) Data on heuristic search for theorem proving

(b) Search for data stored on disk

6. (11 points) Consider training a naive Bayes classifier as in the previous problem and estimating the requisite conditional probability parameters from a set of training examples. Assume there are three possible colors: (red, blue, green) and two possible classes: (positive and negative). Assume the training data has the following properties.

   - There are 7 positive examples that are red
   - There are 4 positive examples that are blue
   - There are 0 positive examples that are green
   - There are 0 negative examples that are red
   - There are 3 negative examples that are blue
   - There are 9 negative examples that are green

   Assume Laplace smoothing is used to estimate parameters with $m = 1$ and $p = 1/3$ (i.e. a prior uniform distribution over the 3 colors). Calculate the conditional probability parameters in the table below.

c	positive	negative
$P(\text{red}\|c)$		
$P(\text{blue}\|c)$		
$P(\text{green}\|c)$		

7. (11 points) Consider the following item ratings to be used by collaborative filtering.

Item	User1	User2	User3	User4	Active User
A	1	5	8	7	9
B		4	3	3	5
C	9		2	1	1
D	4	1	9	8	
E	8	8	3	1	
c_{ij}	-1	1	0.933	0.982	

The Pearson correlation of each of the existing users with the active user (c_{ij}) is already given in the table. Compute the predicted rating for the active user for items D and E using standard significance weighting and the two most similar neighbors to make predictions using the method discussed in class.

8. (20 points) Provide short answers (1-3 sentences) for each of the following questions:

If only positive feedback is given during relevance feedback, will it mainly increase recall or precision? Why?

Why is k nearest neighbor typically better than plain nearest neighbor?

What is the "first rater" problem in collaborative filtering?

In addition to a similarity metric for instances, Hierarchical Agglomerative Clustering (HAC) also requires what *other* type of similarity metric (variations of which result in single-link, complete-link or group average versions of HAC)?

What is probably the best explanation for why Zipf's law applies to so many phenomenon such as personal wealth, city size, and book sales as well as word frequency? Give a brief explanation, not just a short name for the phenomenon.

What is the purpose of the $E(p)$ term in the PageRank algorithm?

What is stipulated by section 230 of the US Communication Decency Act?

What is the computational complexity of the self-attention mechanism of the transformer neural architecture, assuming it is given an input context of n tokens? Briefly explain *why* it has this complexity.

Why is naive Bayes text categorization typically implemented using logarithms of probabilities rather than probabilities themselves?

When is the perceptron algorithm guaranteed to converge to a function that accurately classifies all of the training examples?

(Extra credit) Who is the inventor of the World Wide Web and where was he working at the time he developed it?

(Extra credit) What neural-network pioneer recently won a Nobel Prize in Physics?

(Extra Credit) What fundamental problem in biochemistry was solved using deep learning, resulting in researchers at Deep Mind also winning the recent Nobel prize in chemistry?