Student on the left: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Your name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Your NetID: \_\_\_\_\_\_**

Student on the right: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**MIDTERM EXAM**

**CPSC 477/577:**

**Natural Language Processing**

**March 26, 2019**

**Instructor: Dragomir R. Radev**

|  |  |  |
| --- | --- | --- |
| **Question 1** | **Vector Similarity** | **10** |
| **Question 2** | **Naïve Bayes** | **10** |
| **Question 3** | **Recurrent Neural Networks** | **10** |
| **Question 4** | **Multilayer Neural Networks** | **10** |
| **Question 5** | **Multiple Choice (10 \* 2)** | **20** |
| **Question 6** | **Short Answers (10 \* 2)** | **20** |
| **Question 7** | **Terminology (5 \* 4)** | **20** |
| **TOTAL** |  | **100** |

**(Note: The back of this page and last page can be used as scratch paper. )**

**QUESTION 1.**

Use the following documents for this problem, where the frequency of word appearing (and not just the word’s presence) in a document matters.

D1 = “cat, dog, fox”

D2 = “fish, tiger, cat”

D3 = “cat, fox, dog”

D4 = “fish, cat, fish”

You may find it helpful to transform the documents into frequency vectors using the table below:

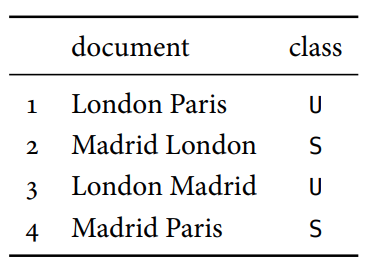
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | fish | cat | fox | tiger | dog |
| D1 |  |  |  |  |  |
| D2 |  |  |  |  |  |
| D3 |  |  |  |  |  |
| D4 |  |  |  |  |  |

1. What is the Jaccard Similarity between D1 and D2?
2. What is the Euclidean Distance between D3 and D4?
3. Suppose we decide to use Cosine Similarity. Which Document is most similar to D2?
4. What are the ranges (in general, not for this particular problem) of the above similarity and distance functions: Jaccard Similarity, Euclidean Distance and Cosine Similarity? You may assume that all document vectors only consist of non-negative components.

**QUESTION 2.**

A Naive Bayes classifier has to decide whether document number 5 ‘London Paris’ is news about the United Kingdom (class U) or news about Spain (class S). You can think of documents 1 through 4 as independent Bernoulli trials.

(a) Estimate the probabilities that are relevant for this decision from the following four documents:. Answer with fractions.



(b) Based on the estimated probabilities, which class does the classifier predict? Assume a uniform prior distribution over the classes U and S. Explain your answer, showing that you have understood the Naïve Bayes classification rule, and show your work.

(c) Practical implementations of a Naive Bayes classifier often use log probabilities. Explain why.

**QUESTION 3.**

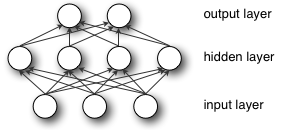
a. Describe the difference between *recurrent* neural networks and *recursive* neural networks. In particular, which can be thought of as a generalization of the other?

b. Suppose you start running into the “vanishing gradient problem” with a vanilla recurrent neural network. What RNN variants should you try to fix the problem?

c. Suppose you want to write a neural network that predicts the sentiment of sentences based on the dependency parse of the sentence. Would you want to use a *recurrent* neural network or a *recursive* neural network? Justify your choice.

d. Suppose you have some time-series data of stock prices, and want to use it to predict the future movement of stock prices. Would you want to use a *recurrent* neural network or a *recursive* neural network? Justify your choice.

**QUESTION 4.**



We use a 3-layer neural network for sentiment classification of words. The architecture of the neural network is shown in the picture. Words are represented by 3-dimensional embeddings as inputs, and the network outputs a probability distribution over the positive and negative classes as a 2 by 1 vector.

The hidden layer and the output layer use Rectified Linear Units (ReLU) as the activation function.

Suppose your parameters are

a) Suppose the input word has an embedding of and calculate the activation of the hidden layer.

b) Following 1), what is the probability distribution after we apply softmax over output layer activations?

**QUESTION 5 (10 MULTIPLE CHOICE QUESTIONS)**

M1. The dependency representation of the sentence "Jane has a cat" is the following (with the arrows pointing from parent to child node):

a. has → cat, has → Jane, cat → a

b. cat → has, Jane → cat, a → cat

c. has → Jane, Jane → cat, cat → has

d. Jane → has, has → a, a → cat

e. has → Jane, Jane → cat, a → cat

M2. Which of the following statements about ambiguity in (English) part of speech tagging is the most accurate? The lecture gave an example from the Brown corpus.

a. The percentage of ambiguous tokens is about the same as the percentage of ambiguous types.

b. The percentage of ambiguous tokens is slightly smaller than the percentage of ambiguous types.

c. The percentage of ambiguous tokens is significantly larger than the percentage of ambiguous types.

d. The percentage of ambiguous tokens is slightly larger than the percentage of ambiguous types.

e. The percentage of ambiguous tokens is significantly smaller than the percentage of ambiguous types.

M3. Which of the following part of speech categories is open class?

a. adjectives

b. prepositions

c. interjections

d. articles

e. conjunctions

M4. Which of the following features are not typically used for sentence boundary recognition?

a. punctuation

b. capitalization

c. word length

d. case (upper case, lower case)

M5. Which of these languages is closest to English from a historical, evolutionary perspective?

a. Greek

b. Norwegian

c. Russian

d. Korean

M6. Probabilistic Context Free Grammars (PCFG), by design, make certain independence assumptions that can hurt the performance of a natural language parser. Techniques such as parent annotation, constituent splitting, and vertical markovization are often used to mitigate such shortcomings. In this example, we will split the NP constituent into NP-SUBJ and NP-OBJ.

Let’s consider the following probabilities:

p = P (NP → PRP)

n = 1-p

ps = P (NP-SUBJ → PRP)

ns = 1-ps

po = P (NP-OBJ → PRP)

no = 1-po

Notes: PRP is a personal pronoun (such as “I”, “she”, and “they”). NP-SUBJ is a noun phrase that serves as the subject of the sentence. NP-OBJ is a noun phrases that serves as the object of the sentence.

Assume a PCFG trained on the Wall Street Journal portion of the Penn Treebank. Which of the following inequalities is/are likely to be accurate? Pick 0, 1, or more answers.

a. <

b. >

c. >

d. >

M7. A dependency tree for a sentence with N words includes this many dependencies:

a. N/2

b. N-1

c. N

d. 2N

M8. What is the semantic type of the sentence fragment "*walks(Janet)*"?

a. e

b. t → t

c. t

d. e → t

M9. If is coffee beans and is ground coffee, what is ?

a. cup of water

b. cup of green tea

c. cup of black coffee

d. cup of venti non-fat ice caramel macchiato with extra foam, to go

M10. Which one of the following can be considered as a "universal function approximator"?

a. a two-layer neural network

b. a hidden markov model

c. a push-down automaton

d. a finite-state automaton

**QUESTION 6. Short Answer Questions (TEN short answer questions)**

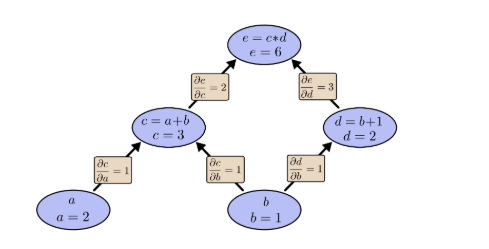
S1. What is the difference between the Viterbi algorithm and the Forward algorithm for POS tagging. Give pseudo-code for both and also explain the difference in plain English.

S2. Derive the derivative of the logistic function:

S3. Draw a neural network with a hidden layer that can compute X XOR Y for the Boolean variables X and Y. Show all biases and other weights and explain why your network works. Note that there are many possible answers.

S4. One of the lectures used sentences containing the word "acerola" to illustrate a concept. What is that concept?

S5. For the given computational graph below, calculate the following partial derivatives:



S6. Give three examples of syntactic constituency test, other than the “intuition” test.

S7. Which of the following words start(s) with a voiced consonant?

tear, cat, three, gate, part, fuse, deer

S8. In the following table we consider four candidates for when , which is the most likely correction? Why?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **|c** | **|c)** |  |
| deea | deer | a|r | 0.0001 | 0.003 |
| deea | idea | de|id | 0.0007 | 0.0009 |
| deea | dear | ea|ar | 0.0003 | 0.007 |
| deea | deea |  | 0.85 | 0.00000006 |

S9. A corpus contains 1,000,000 tokens, including 6,000 instances of the word "the". The next four most frequent words are "of", "and", "to", and "a". What are the expected counts for each of these words in the corpus?

S10. Translate into plain English the following FOL sentence:

**QUESTION 7. Terminology and examples (five items)**

Explain each of these terms and give an example (or formula, if appropriate) of each:

a. LCS (lowest common subsumer)

b. confusion matrix for binary classification

c. the semantic compositionality principle

d. negative sampling (for word embeddings)

e. backpropagation over time for RNN (recurrent neural networks)