CS 124 Spring 2021 Midterm

Instructions:

• The main exam consists of 17 questions, worth a total of 23 points.
• You have 80 minutes to complete this exam.
• You may use the calculator functions on your computer to compute values for functions such as cosine.

Stanford University Honor Code:

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

(a) The Honor Code is an undertaking of the students, individually and collectively:

(i) 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;

(ii) 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.

Your name: ____________________________________________

SUNet Email: ____________________________@stanford.edu Numerical SUID: ____________

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.

(Signed) ________________________________________________
Regular Expressions and Edit Distance (4 points)

1. (1 point) You are a barista at Starbucks and an alumni of CS 124. Your customer pronounces her name as “KATE-LIN.” You aren’t sure how it’s really spelled, so you cleverly write a regular expression on the cup

\[KCai?te?l[ly]nn?\]

in order to capture a variety of possible spellings such as:

Katelyn
Caitlyn
Caitlin
Kaitlyn

However, the regular expression is not perfect. Which of the following alternate spellings is not generated by the regular expression?

(a) Kaitlinn
(b) Catlin
(c) Kaytlin
(d) Caitelyn

The answer is (c). The regular expression does not match the “y” that appears before the “t”.

2. (1 point) Which of the following regular expressions matches only one of the strings “NLP124” or “nlp224n”? Select all that apply. Recall that the regular expression “\w” captures alphanumeric characters.

(a) [Nn][Ll][Pp]\d+
(b) [Nn]\w+\d?
(c) [Nn][Ll][Pp]\d\d\d\w?
(d) [Nn][Ll]\w?\d?\d?\w?

The answers are (a) and (d).
(b) and (c) match both. (a) and (d) match only the NLP124.

3. (2 points) Which path represents a valid minimum alignment between these two words? Recall that substitution costs 2.
The answer is (b).
The completed alignment table is shown below:
(a) is not correct since the very first move is to substitute HELLO to be EELLO, even though the E in both words should align. Similarly, (c) is not correct because it does not align the L from the two words. (d) is not correct because the path must always go up and towards the right. Otherwise, it cannot be a minimum path.
Language Modeling and Naive Bayes (5 points)

4. (1 point) Suppose we build a simple unigram language model (no add-one smoothing, no unknown words) from the following snippet of text.

<\s> the cat wore the hat </s>
<\s> the hat was on the cat </s>
<\s> i want a blue hat the cat said </s>
<\s> but alas the cat’s hat was red </s>

Suppose we see the word ‘the’, what is the probability that the next word is ‘hat’, i.e. $P(\text{hat} | \text{the})$? Don’t forget to include the tokens <\s> and </s> in your counts.

(a) $\frac{2}{17}$
(b) $\frac{3}{34}$
(c) $\frac{1}{3}$
(d) $\frac{4}{17}$

The answer is (a).

Note that we are using a simple unigram model, so no add-1 smoothing and no need to generate bigrams. Thus, we count the number of tokens (34) and simply divide the number of hat tokens by the number of tokens. Although the question asks for $P(\text{hat} | \text{the})$, this is just equal to $P(\text{hat})$ given our unigram model.

For the next two questions, please use the following information:

You build a Naive Bayes classifier and test it on a dataset containing 170 examples. Your tests yield the confusion matrix below:

<table>
<thead>
<tr>
<th></th>
<th>Predicted Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive (y=1)</td>
</tr>
<tr>
<td>True Class</td>
<td></td>
</tr>
<tr>
<td>Positive (y=1)</td>
<td>50</td>
</tr>
<tr>
<td>Negative (y=0)</td>
<td>30</td>
</tr>
</tbody>
</table>

5. (1 point) What is the recall?

(a) $\frac{5}{8}$
(b) $\frac{5}{7}$
(c) $\frac{5}{17}$
(d) $\frac{2}{7}$

The answer is (b).

Recall is the fraction of positive examples that are classified correctly $\frac{50}{70}$. 
6. (1 point) What is the precision?

(a) \( \frac{5}{8} \)
(b) \( \frac{5}{7} \)
(c) \( \frac{5}{17} \)
(d) \( \frac{1}{2} \)

The answer is (a).

Precision is the fraction of relevant instances (true positives = 50) among the retrieved instances (true positives + false positives = 50+30 = 80). The fraction of positive examples that are classified correctly is \( \frac{50}{80} \).

7. (2 points) We are given the following corpus:

<s> My name is John </s>
<s> John is my name </s>
<s> I am John </s>
<s> My name is John </s>
<s> My name is not John </s>

If we use linear interpolation smoothing between a maximum-likelihood unigram model and a maximum-likelihood bigram model with \( \lambda_1 = \frac{3}{4} \) and \( \lambda_2 = \frac{1}{4} \), what is \( P(\text{John}|\text{is}) \)? Don’t forget to include <s> and </s> in your counts as you would any other token.

(a) \( \frac{11}{24} \)
(b) \( \frac{5}{12} \)
(c) \( \frac{7}{12} \)
(d) \( \frac{1}{2} \)

The answer is (b).

This question follows the formula utilized by Question 3 on the Week 2 quiz:

\[
\frac{1}{4} \cdot P(\text{John}) + \frac{3}{4} \cdot P(\text{John}|\text{is}) = \frac{1}{4} \cdot \frac{5}{30} + \frac{3}{4} \cdot \frac{2}{4} = \frac{10}{24} = \frac{5}{12}.
\]
Logistic Regression and Sentiment Analysis (2 points)

8. (2 points) We are given the following tweet and we have to use logistic regression (using the softmax function) to classify it as either having positive sentiment ($y = 1$), negative sentiment ($y = -1$), or neutral sentiment ($y = 0$).

"What a great quarter that was. I had a blast in CS124!"

Positive lexicon: great, blast

Negative lexicon: boring, blast

Suppose:

$$W_{positive} = [0.3, 0.8, -0.7] \quad b_{positive} = 0$$
$$W_{neutral} = [0.5, 0, 0] \quad b_{neutral} = 0$$
$$W_{negative} = [0.5, -0.2, 0.7] \quad b_{negative} = 0$$

Use the following three features (for all classes) to calculate $P(y = 1|X)$, $P(y = 0|X)$, and $P(y = -1|X)$.

$X_1 = 1$ if '!' ∈ doc, 0 otherwise

$X_2 =$ count of positive lexicon words in doc

$X_3 =$ count of negative lexicon words in doc

Do not round off your numbers until the final solution.

(a) $P(y = 1|X) = 0.44$, $P(y = 0|X) = 0.20$, $P(y = -1|X) = 0.36$

(b) $P(y = 1|X) = 0.77$, $P(y = 0|X) = 0.62$, $P(y = -1|X) = 0.69$

(c) $P(y = 1|X) = 0.48$, $P(y = 0|X) = 0.20$, $P(y = -1|X) = 0.32$

(d) $P(y = 1|X) = 0.46$, $P(y = 0|X) = 0.23$, $P(y = -1|X) = 0.31$

The answer is (d).

$X = [1, 2, 1]$

$z_{positive} = W_{positive} \cdot X + b_{positive} = 0.3 + 1.6 - 0.7 = 1.2$

$z_{neutral} = W_{neutral} \cdot X + b_{neutral} = 0.5 + 0 + 0 = 0.5$

$z_{negative} = W_{negative} \cdot X + b_{negative} = 0.5 - 0.4 + 0.7 = 0.8$

$[P(y = 1|X), P(y = 0|X), P(y = -1|X)] = softmax([z_{positive}, z_{neutral}, z_{negative}])$

$= [0.461, 0.229, 0.309]$
Information Retrieval (6 points)

9. (1 point) We have a set of 600 documents in our corpus and the terms below have the following document frequencies:

<table>
<thead>
<tr>
<th>term</th>
<th>document frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>seahorse</td>
<td>100</td>
</tr>
<tr>
<td>shell</td>
<td>400</td>
</tr>
<tr>
<td>ocean</td>
<td>350</td>
</tr>
<tr>
<td>wave</td>
<td>200</td>
</tr>
<tr>
<td>sun</td>
<td>500</td>
</tr>
<tr>
<td>sea</td>
<td>250</td>
</tr>
</tbody>
</table>

What part of the boolean query "sea AND sun AND NOT (shell OR seahorse) AND NOT (wave OR ocean)" should we execute first?

(a) "NOT (shell OR seahorse)"
(b) "seahorse AND NOT wave"
(c) "NOT (wave OR ocean)"
(d) "sea AND sun"

The answer is (c).
Remember that we estimate (x OR y) by adding |x| + |y|, (x AND y) by taking min(|x|, |y|), and (NOT x) by subtracting N - |x| where N is the total number of documents.

10. (1 point) In ranked retrieval, we often utilize the ltc.lnn weighting and normalization scheme. How would the ranking of our results change if we were to switch to ltc.lnc weighting?

(a) Documents that contain all of the words in the query would be ranked highest.
(b) The ranking will no longer pay attention to the length of the documents.
(c) There will be no change in the ranking.
(d) Not enough information is provided to determine if the ranking would change.

The answer is (c)
The comparisons are all relative to the same query, so if we start normalizing by the query vector’s length, there will be no change in the overall ranking.

Note: Use the following documents for the next three questions.

Doc 1: “My favorite color is sky blue”
Doc 2: “I am really surprised blue cheese is really blue”
Doc 3: “There are so many clouds in the blue sky”
Doc 4: “The sky is really clear today”
11. (1 point) Build a positional inverted index for the above documents. Which of the following would appear in the positional inverted index?

(a) “sky” ⇒ {1 : [4], 2 : [2], 3 : [8], 4 : [1]}
(b) “really” ⇒ {2 : [2], 4 : [3]}
(c) “clouds” ⇒ {3 : [2]}
(d) "blue" ⇒ {1 : [5], 2 : [4, 8], 3 : [7]}

D. (a) "sky" does not actually show up in Document 2.
(b) "really" shows up twice in Document 2.
(c) "clouds" shows up at position 4 in Document 3.

12. (1 point) Which of the following is the result of a boolean retrieval for the query “blue sky”?

(a) Doc 1, Doc 2, Doc 3
(b) Doc 1, Doc 3, Doc 4
(c) Doc 3
(d) Doc 1, Doc 3

D. "blue" shows up in Documents 1, 2, and 3. "sky" shows up in Documents 1, 3, and 4. The common documents here are Documents 1 and 3.

13. (2 points) What would be the results of the phrase retrieval for the query “blue sky”?

(a) Doc 1, Doc 3
(b) Doc 3
(c) Doc 3, Doc 4
(d) Doc 1, Doc 2, Doc 3

B. The phrase "blue sky" only shows up in Document 3. Note that "sky blue" doesn’t count because it’s in the wrong order.
Vector Semantics (2 points)

14. (1 point) Good word embeddings, such as word2vec embeddings, often capture relational meanings. Which of the following equations do you think should hold for a good word embedding? Select all that apply.

(a) $\text{vector('boy')} - \text{vector('girl')} = \text{vector('brother')} - \text{vector('sister')}$
(b) $\text{vector('boy')} - \text{vector('brother')} = \text{vector('girl')} - \text{vector('sister')}$
(c) $\text{vector('boy')} - \text{vector('brother')} = \text{vector('sister')} - \text{vector('girl')}$
(d) $\text{vector('sister')} - \text{vector('brother')} = \text{vector('boy')} - \text{vector('girl')}$

The answer is (a)(b).

15. (1 point) Suppose we have the embedding for a word $a$, and we find another word $b$ that has an embedding with a very high cosine similarity. Which of the following is most likely to be true?

(a) Words $a$ and $b$ are synonyms.
(b) The embeddings for $a$ and $b$ are only alike due to randomness in the vector space.
(c) Words $a$ and $b$ appear in a lot of the same contexts.
(d) Words $a$ and $b$ can be used interchangeably without affecting the meaning of sentences.

The answer is (c).
Recall that the embedding for a word is generated by considering the words that appear near it. The contexts of words define the embeddings, not the meanings.
Neural Networks (4 points)
Note: Use the following neural network for the next two questions.

16. (2 point) Given the input $x = [2, 3]$ and using the ReLU activation function, what is the output $out_Y$ of the neural net?

(a) -1
(b) 0
(c) 15
(d) 16

The answer is (d). $out_A = 4$, $out_B = 0$, $out_Y = 16$

17. (2 point) Suppose you are using the sigmoid activation function and want to use backpropagation to update your weight $w_{x1,A}$. After running forward propagation, which inputs, weights, and/or neuron outputs do you need in order to update $w_{x1,A}$?

(a) $w_{x1,A}, out_A, w_{A,Y}, out_Y$
(b) $x_1, w_{x1,A}, out_A, w_{A,Y}, out_Y$
(c) $x_1, w_{x1,A}, out_A, w_{A,Y}, out_B, w_{B,Y}, out_Y$
(d) $x_1, w_{x1,A}, w_{x1,B}, out_A, w_{A,Y}, out_B, w_{B,Y}, out_Y$

The answer is (b). In order to update $w_{x1,A}$, we need to know the current value of $w_{x1,A}$ and $\frac{\partial \text{Loss}}{\partial w_{x1,A}}$. We can compute $\frac{\partial \text{Loss}}{\partial w_{x1,A}} = \frac{\partial \text{Loss}}{\partial out_Y} \ast \frac{\partial out_Y}{\partial out_A} \ast \frac{\partial out_A}{\partial in_A} \ast \frac{\partial in_A}{\partial w_{x1,A}}$. Recall that the derivative of the sigmoid function is equal to $\sigma \ast (1 - \sigma)$. Thus, we have $\frac{\partial out_Y}{\partial in_Y} = out_Y \ast (1 - out_Y)$, $\frac{\partial in_Y}{\partial out_A} = w_{A,Y}$, $\frac{\partial out_A}{\partial in_A} = out_A \ast (1 - out_A)$, $\frac{\partial in_A}{\partial w_{x1,A}} = x_1$. 

11