CS4248 Natural Language Processing Final Exam (for Archiving) AY 21/22 Sem 2

Kan Min-Yen / Christian von der Weth
No TF/MCQ/MRQ (external use ok)

1. During model training to convergence, we observe high training loss and almost equally high validation loss.

   (3 marks) Is the model suffering from overfitting or underfitting? Briefly justify your answer.

   (4 marks) Name two methods that help address the above-mentioned issue. For each method, use one or two sentences to explain how it functions.

   (For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)

2. For n-gram language models, what are advantages and disadvantages of setting a large n?

   Justify each aspect that you give.

3. **Activation Functions**

   (3 marks) Name two key properties of the softmax activation function.

   (7 marks) Design a replacement activation function that preserves the key properties of the original softmax without using any exponentials (i.e., *exp()*), justifying how your answer satisfies the properties.

   (For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)

4. **Attention Mechanism**

   (5 marks) Recall that we adopt the attention mechanism to compute a set of importance weights over the encoder hidden states to obtain a weighted aggregation. Alternatively, we can pass the encoder hidden states directly through a fully-connected learnable weight layer to obtain a weighted aggregation.

   What are key difference(s) between these two representation methods?

   (7 marks) Both of the above models' self-attention require quadratic cost $O(n^2)$ with respect to their input length. That is, given an n-length input sequence “it was such a nice day”, we compute a score for every pair of words in this sequence, visualized as all n^2 cells in the diagram below. This does not scale well for long input sequences.

Sketch a self-attention variant can attend to long sequences with reduced cost, briefly justifying your answer.

(Note: You need not specify a formal mathematical definition; a description with justification is sufficient)

(For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)

5. **Antonymic Embeddings**

(2 marks) Consider an antonymic pair, (student, teacher) as represented by their word embeddings. Would the embedding for teacher be more likely to appear in embeddings that are nearest to or farthest from student?

Briefly justify your answer.

Parts 2 and 3 concern recognizing antonymic pairs automatically, leveraging the contrast hypothesis as a means for detection, when antonymic resources are available.

Contrast Hypothesis: For a given antonymic word pair (A, B), there are other antonymic pairs (C, D), such that A and C are highly similar and B and D are highly similar (e.g., (student, teacher) and (learn, teach)).

(6 marks) Assume that we have a lexicon of antonym pairs L=<<w_1^+, w_1^->,<w_2^+, w_2^->, … ,<w_N^+, w_N^->, where each <w^+,w^-> is a pair of antonyms.

Given an appropriately-trained embedding model and an input pair of words <x, x'>, sketch a method to predict whether <x, x'> is a pair of antonyms.

(Hint: You can assume a threshold k deciding whether two words are similar)

(6 marks) Assume that we now additionally have access to a ground truth labeled corpus GT=<<x_1, x_1'>,<x_2, x_2'>, …,<x_N, x_N'>>, where each word pair <x, x'> has a binary label from {antonymic / not antonymic}.

Describe a framework to optimize an antonym pair prediction model by sketching a brief, 1-sentence description on how the framework handles each of the following three components:

a. Model selection
b. Feature engineering
c. Training and evaluation.

(For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)

6. **Minimum Edit Distance**

1. Calculate the Minimum Edit Distance (M.E.D.) values for the prefixes of "campus" and

"compass" below.

(For this part, assume that the cost of inserting, deleting, and substituting a character is 1, 1 and 2, respectively. You need not show intermediate steps or the M.E.D. table for full marks, as it may be difficult to input digitally)

(3 marks) M.E.D. between "campu" and "com": _BLANK_
(3 marks) the full M.E.D. between "campus" and "compass": _BLANK_

Answer: 4 and 5

2. (3 marks) If we reverse the two above-mentioned strings and re-calculate the M.E.D. for the reversed strings, can we reuse part of the previous computations with respect to the original pair of strings?

Answer: (Yes / No) _BLANK_

3. (3 marks) If we allow transposition (e.g. xyz→ xzy) with the cost of 1/2 per operation, what is the E.D. between "campus" and "compass"?

Answer: The edit distance with transposition is: _BLANK_.

7. **Naïve Bayes**

A collection of movie reviews contains the following keywords and binary sentiment label (pos and neg). The data are shown below.

doc	interesting	funny	upset	cried	class
d1	1	1	0	1	pos
d2	2	2	0	0	pos
d3	0	1	2	1	neg

Test sentence: "*This is an interesting movie -- didn't upset me at all. The plots were so funny that I cried out in the cinema.*"

(10 marks) Create a multinomial Naïve Bayes model with add-1 smoothing to predict the label of the test sentence. Show all results of the log-prior values, log-likelihood values and predicted label) for full credit.

Predicted label: POSITIVE.

(3 marks) If a negation detection algorithm is run on the test sentence that discards any occurrences of negated adjectives or verbs, would the prediction change? Justify your answer.

(Note: For this question, please present all log values in the form of log a/b. For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)

8. Generative Language Model

In this task, we generate strings using a language model. The strings are generated from left-to-right, sourced from the two type vocabulary V={w_1,w_2}. We consider a locally-normalized trigram language model; i.e., scores are normalized at each generation step via the equation shown below.

$$p(y_t \mid y_{t-2}, y_{t-1}) = \frac{score\ (y_t, y_{t-2}, y_{t-1})}{\sum_{y_l \in V} score\ (y_l, y_{l-2}, y_{l-1})}$$

The non-normalized scores of the language model are given in the below table. The beginning of a string is marked by the (<s><s>) sequence, and we consider neither a special end-of-sequence symbol nor the empty string.

Given the start sequence (<s><s>) and using beam search with beam size of 2,

(10 marks) Decode the sequences of length 3 (without counting the (<s>) tokens). For full credit, show the partial decoded sequence and their conditional probabilities at each iteration.

(4 marks) State the most likely output sequence of length 3 and calculate its probability.

(For multipart questions such as this one, please explicitly demarcate which part of your response addresses each part)å