CSE 538: Sample Mid-term

September 29, 2021

There will be three types of questions.

1) Test your ability to apply the ideas we learnt on a new task. 2) Test your understanding of the basic concepts we saw in class. 3) Test your ability to apply these concepts to solve a sub-problem of something that is close to what we saw in class.

Note:

You can bring a single sheet (2 pages) of hand-written notes to use so there is less pressure to memorize stuff.

1 Authorship Detection

Assume you are building a tool that can recognize sentences written by different authors. You will show how we can build such an application using techniques we learned in class.

Dataset: You are given a collection of 1000 books drawn from some digitized collection. Let's say there are only ten authors in all. Each author has written at least ten books. Some authors are more prolific than others.

Assume that you are using a **generative** language modeling approach. Walk through the steps involved in setting this up as a classification problem solved using Bayesian classification.

Note: You are simply using count based techniques here i.e. no word embeddings.

A. Write down the argmax function you will use to predict the author who wrote an input sentence S, which is a sequence of words $w_1, \cdots, w_n$. Use A_j to denote the j^{th} author. The argmax function should be defined over a probabilistic quantity. [5 points]

B. Use Bayes theorem and related independence assumptions to show how you can estimate the above quantity using the independence assumptions you want to make. Please provide the maximum likelihood estimation formulas. [10 points]

C. Suppose your friend from the Creative Writing department tells you that authors usually like to use some kind of signature patterns such as "not only [x] but also [y]", "in addition to". Suggest one way of incorporating this information in this model and why it might be difficult to do so.

2 Word Representations

Recall that when learning word embeddings, we simply use a single UNK token to represent words that we haven't seen in training data.

Suppose your linguist friend tells you that word endings often carry useful information about word meanings. For instance, words that look like "[x]ed" are most likely words that represent the past tense of a verb [x].

How will you use this information to improve representation of unknown words? Specify what changes you need to set up the training data, and how you will represent an unseen word during test time. Also list what is one draw back of your idea and give an example to illustrate.

3 Sentence Representations

1. State one reason why a learning a model that always outputs a single representation of the entire sentence is not a good idea. Give an example to illustrate.
2. We learnt word representations by trying to predict an outer word given a context word. Design a similar scheme for learning sentence representations. Given the representation of a context sentence, you want to predict the next sentence. Specify the following: (i) The probabilistic quantity you target (ii) What is the function form that you will use to approximate this probabilistic quantity. NOTE: Assume that you will use DAN to give you a representation of the sentence. (iii) Write down the full objective function you will use to train these representations.