CS 6320.002: Natural Language Processing
Spring 2020

Homework 1 Written Component – 50 points
Issued 30 Aug. 2021
Due 11:59pm CDT 31 Sept. 2021

Deliverables: Answers are to be typed directly into Gradescope.

What does it mean to "show your work?" Write out the math step-by-step; we should be able to clearly follow your reasoning from one step to another. (You can combine "obvious" steps like simplifying fractions or doing basic arithmetic.) The point of showing your work is twofold: to get partial credit if your answer is incorrect, and to show us that you worked the problem yourself and understand it. We will deduct points if steps are missing.

1 Math Review — Multivariate Calculus

The problems in this section refresh your memory on concepts from classes you have taken previously that we will use later in this course.

1.1 Partial Derivatives (5 points)

$$f(x, y, z) = \frac{xz}{y^2} + yze^{x^2}$$

What are $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$? Show your work.

Since we know that $\frac{d}{dx}x^2 = 2x$, and that $\frac{d}{dx}ce^{f(x)} = f'(x)ce^{f(x)}$, it follows that

$$\frac{\partial f}{\partial x} = \frac{z}{y^2} + 2xyze^{x^2}$$

Additionally, we know that $\frac{d}{dy}cy^{-2} = -2cy^{-3}$, and so

$$\frac{\partial f}{\partial y} = \frac{-2xz}{y^3} + ze^{x^2}$$

Finally, we have that $\frac{d}{dz}cz = c$, which gives

$$\frac{\partial f}{\partial z} = \frac{x}{y^2} + ye^{x^2}$$

1.2 The Chain Rule (5 points)

$$
\begin{aligned}
f(x,y) &= xg(x,y) + 5y \\
g(x,y) &= x^2y - xh(x^2,y) \\
h(x,y) &= xy^2 + 2
\end{aligned}
$$

What are $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$? Show your work.

By the chain rule, we have

$$
\begin{aligned}
\frac{\partial f}{\partial x} &= g(x,y) + x\frac{\partial g}{\partial x} \\
&= (x^2y - x \cdot h(x^2,y)) + x(\frac{\partial g}{\partial x})
\end{aligned}
$$

$$
h(x^2,y) = h'(x,y) = x^2y^2 + 2
$$

$$
\frac{\partial h'}{\partial x} = 2xy^2
$$

$$
\begin{aligned}
\frac{\partial g}{\partial x} &= 2xy - (h(x^2,y) + x(\frac{\partial h'}{\partial x}) \\
&= 2xy - (x^2y^2 + 2 + x(\frac{\partial h'}{\partial x})) \\
&= 2xy - (x^2y^2 + 2 + x(2xy^2)) \\
&= 2xy - 3x^2y^2 - 2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial f}{\partial x} &= (x^2y - x \cdot h(x^2,y)) + x(2xy - 3x^2y^2 - 2) \\
&= (x^2y - x(x^2y^2 + 2)) + x(2xy - 3x^2y^2 - 2) \\
&= x^2y - x^3y^2 - 2x + 2x^2y - 3x^2y^2 - 2x \\
&= 3x^2y - 4x^3y^2 - 4x
\end{aligned}
$$

Similarly, for the other partial,

$$\frac{\partial f}{\partial y} = x\frac{\partial g}{\partial y} + 5$$

$$h(x^2, y) = h'(x, y) = x^2y^2 + 2$$
$$\frac{\partial h'}{\partial y} = 2x^2y$$

$$\begin{aligned}\frac{\partial g}{\partial y} &= x^2 - x(\frac{\partial h'}{\partial y}) \\ &= x^2 - 2x^3y\end{aligned}$$

$$\begin{aligned}\frac{\partial f}{\partial y} &= x(x^2 - 2x^3y) + 5 \\ &= x^3 - 2x^4y + 5\end{aligned}$$

1.2.1 Extrema (5 points)

$$f(x) = x\log_8(x) + (1-x)\log_8(1-x)$$

What are the values of x corresponding to the minima and maxima of $f(x)$ for $x \in [0, 1]$? Show your work (your math work; graphing it doesn't count!).
We are looking for the members of the set $\{x | x \in [0,1] f'(x) = 0, f''(x) \neq 0\}$
Recall that for $g(x) = \log_a(x)$, it follows that $g'(x) = \dfrac{1}{x\ln(a)}$. Thus, we find the following:

$$\begin{aligned}f'(x) &= \log_8(x) + \frac{1}{\ln 8} - \log_8(1-x) - \frac{1}{\ln 8} \\ &= \frac{\ln x}{\ln 8} + \frac{1}{\ln 8} - \frac{\ln(1-x)}{\ln 8} - \frac{1}{\ln 8} \\ &= \frac{\ln x - \ln(1-x)}{\ln 8} \\ &= \frac{\ln \frac{x}{1-x}}{\ln 8}\end{aligned}$$

It follows that $f'(x) = 0$ exactly when $\ln(\frac{x}{1-x}) = 0$, and this occurs when $\frac{x}{1-x} = 1$. That is, when $x = 1 - x$. This gives the solution of $x = 1/2$.
Now, we check the function at this point as well as at the bounds.

$$f(0) = 0 * 1 + 1 * 0 = 0 \quad | \; f(1/2) = (\tfrac{1}{2})(\tfrac{-1}{3}) + (\tfrac{1}{2})(\tfrac{-1}{3}) = \tfrac{-1}{3} \quad | \; f(1) = 1 * 0 + 0 * 1 = 0$$

$$
\begin{aligned}
f(0) &= 0 * 1 + 1 * 0 && = 0 \\
f(1/2) &= (\frac{1}{2})(\frac{-1}{3}) + (\frac{1}{2})(\frac{-1}{3}) && = \frac{-1}{3} \\
f(1) &= 1 * 0 + 0 * 1 && = 0
\end{aligned}
$$

Thus, we find that the minima of the function occur at $x = 1/2$ and that the maxima of the function occur at $x = 0, 1$.

2 Math Review — Probability and Statistics

The problems in this section refresh your memory on concepts from classes you have taken previously that we will use later in this course.

2.1 Conditional Probability (5 points)

Suppose there is a box containing 12 balls; six are orange, and six are green. You remove four balls at random, without replacing any. What is the probability that you remove four orange balls? Show your work.
Effectively, we are computing the product of the following probabilities

$$P(OOOO|OOO) * P(OOO|OO) * P(OO|O) * P(O)$$

That is, the probability of 4 orange balls given that we just drew 3, multiplied by the probability that we draw 3 orange balls given that we just drew 2, etc.

We are removing four balls. The probability that the first ball is orange is 6/12. After this, we have five orange balls and six green balls. The probability that the next ball is orange is 5/11.
This continues for the next two removals, and we get that

$$
\begin{aligned}
P(OOOO) &= \frac{6}{12} \cdot \frac{5}{11} \cdot \frac{4}{10} \cdot \frac{3}{9} \\
&= \frac{360}{11880} \\
&= \frac{1}{33} \\
&\approx 0.303
\end{aligned}
$$

2.2 Bayes's Rule (5 points)

Suppose you have two lab-mates. One (Friend A) talks about computer science 80% of the time, and linguistics 20% of the time; the other (Friend B) talks about linguistics 70% of the time, and computer science 30% of the time. One day, you find a typed note on your desk about computer science. Your lab-mates leave you notes equally often, so

you don't know who left this one. What is the probability the note is from Friend A? Show your work.
The following are given:

$$P(C|A) = .8 \quad P(L|A) = .2 \quad P(A) = .5$$
$$P(C|B) = .3 \quad P(L|B) = .7 \quad P(B) = .5$$

We know by Bayes's Rule that we are trying to find

$$P(A|C) = \frac{P(C \cap A)}{P(C)}$$

Observe that

$$\begin{aligned} P(C \cap A) &= P(A) * P(C|A) \\ &= .5 * .8 \\ &= .4 \end{aligned}$$

and also that

$$\begin{aligned} P(C) &= P(C \cap A) + P(C \cap B) \\ &= P(C|A) * P(A) + P(C|B) * P(B) \\ &= .8 * .5 + .3 * .5 \\ &= .55 \end{aligned}$$

Plugging these into the original definition for our answer gives

$$P(A|C) = \frac{.4}{.55} \approx 0.727$$

3 Language Modeling

The problems in this section are based on the material covered in Week 2.

Suppose we have a training corpus consisting of two sentences:

- The cat sat in the hat on the mat
- The dog sat on the log

3.1 Smoothing — Discounting and Katz Backoff (5 points)

If we train a bigram Katz backoff model on this corpus, using $\beta = 0.75$ and no end token, what is $p_{katz}(\text{sat}|\text{dog})$? What is $p_{katz}(\text{sat}|\text{fish})$? Show your work.
There is only one instance of the word dog, and so $c(\text{dog}) = 1$. This instance is immediately followed by "sat", so we find

$$c_d(\text{dog, sat}) = c(\text{dog, sat}) - \beta = 1 - \beta = 0.25$$

Since the $c(\text{dog, sat}) = 1 > 0$, we know that sat $\in A(\text{dog})$ so we don't need to back off to using unigram counts. Thus, $p_{katz}(\text{sat}|\text{dog}) = \dfrac{c_d(\text{dog, sat})}{c(\text{dog})} = \dfrac{0.25}{1} = 0.25$

There is no instance of the word fish, and so $c(\text{dog}) = 0$. This implies that $c(\text{fish, sat}) = 0$, meaning sat $\in B(\text{fish})$, and so we back off to using the unigram counts.
We need to find

$$\alpha(\text{fish}) = 1 - \sum_{w \in A(\text{fish})} \frac{c_d(\text{fish}, w)}{c(\text{fish})}$$

However, since there was no occurrence of "fish", $A(\text{fish}) = \emptyset$, and so the summation is simply 0. Thus, $\alpha(\text{fish}) = 1$.
This means that we will use

$$p_{\text{katz}}(\text{sat}|\text{fish}) = 1 \times \frac{p_{MLE}(\text{sat})}{\sum_{w' \in B(\text{fish})} p_{MLE}(w')}$$

For a unigram u, it's the case that $p_{MLE}(u) = c(u)$. Thus, $p_{MLE}(\text{sat}) = c(\text{sat})/|V| = 2/9$. We need to examine the set $B(\text{fish})$ which is defined is $\{w|c(\text{fish}, w) = 0\}$, and since "fish" doesn't occur in the corpus, all words in the vocabulary are in $B(v)$, and so the sum of their maximum likelihood probabilities will total 1.

$$p_{katz}(\text{sat}|\text{fish}) = \alpha(fish) \times \frac{\frac{2}{9}}{1} = \frac{2}{9}$$

3.1.1 Smoothing — Linear Interpolation (5 points)

If we use linear interpolation between a bigram model and a unigram model, using $\lambda_1 = \lambda_2 = 0.5$ and no end token, what is $p_{inter}(\text{dog}|\text{the})$? What is $p_{inter}(\text{dog}|\text{log})$? Show your work.

$$p_{\text{inter}}(\text{dog}|\text{the}) = \lambda_1 p_{MLE}(\text{dog}|\text{the}) + \lambda_2 p_{MLE}(\text{dog})$$

We have 5 occurrences of "the" (ignoring case), and only 1 of these is followed by "dog". Thus, we have that $p_{MLE}(\text{dog}|\text{the}) = \dfrac{c(\text{the, dog})}{c(\text{the})} = \dfrac{1}{5}$
There is 1 occurrence of "dog" and 9 words in the vocabulary, so $p_{MLE}(\text{dog}) = 1/9$. This gives

$$p_{\text{inter}}(\text{dog}|\text{the}) = 0.5 * \frac{1}{5} + 0.5 * \frac{1}{9} = \frac{7}{45} \approx 0.155$$

$$p_{\text{inter}}(\text{dog}|\text{log}) = \lambda_1 p_{MLE}(\text{dog}|\text{log}) + \lambda_2 p_{MLE}(\text{dog})$$

We have 1 occurrence of "log" (ignoring case), and it isn't followed by anything. Thus, we have that $p_{MLE}(\text{dog}|\text{log}) = \dfrac{c(\text{log, dog})}{c(\text{log})} = \dfrac{0}{1} = 0$

There is 1 occurrence of "dog" and 9 words in the vocabulary, so $p_{MLE}(\text{dog}) = 1/9$. This gives

$$p_{\text{inter}}(\text{dog}|\text{log}) = 0.5 * \frac{0}{1} + 0.5 * \frac{1}{9} = \frac{1}{18} \approx 0.055$$

3.2 Perplexity (5 points)

What is the maximum possible value that the perplexity score can take? What is the minimum possible value it can take? Explain your reasoning and give an example of a training corpus and two test corpora, one that achieves the maximum possible perplexity score and one that achieves the minimum possible perplexity score. (You can do this with a single short sentence for each corpus.)

3.3 Generation (5 points)

Use your code from the programming component of this assignment to train three language models on the provided data file, `shakespeare.txt`: one unigram model, one trigram, and one 5-gram. For each model, generate 5 random sentences with `max_length=10`. Show the sentences you generated with each model.

What are some problems you see with the generated sentences? How do the sentences generated by the different models compare with each other?

3.4 Applications (5 points)

Authorship identification is an important task in NLP. Can you think of a way to use language models to determine who wrote an unknown piece of text? Explain your idea and how it would work (you don't need to implement it). You must use language modeling the receive credit! Other approaches do not count.