Q1 Math Review --- Multivariate Calculus

15 Points

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

Q1.1 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.

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

Now, we need to calculate partial derivatives, $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, and $\frac{\partial f}{\partial z}$

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}\left(\frac{xz}{y^2} + yze^{x^2}\right)$$
$$= \frac{\partial}{\partial x}\left(\frac{xz}{y^2}\right) + \frac{\partial}{\partial x}\left(yze^{x^2}\right)$$

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

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}\left(\frac{xz}{y^2} + yze^{x^2}\right)$$
$$= \frac{\partial}{\partial y}\left(\frac{xz}{y^2}\right) + \frac{\partial}{\partial y}\left(yze^{x^2}\right)$$

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

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

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

Q1.2 The Chain Rule

5 Points

$$f(x, y) = xg(x, y) + 5y$$

$$g(x, y) = x^2y - xh(x^2, y)$$

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

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

Given, $f(x, y) = xg(x, y) + 5y$, $g(x, y) = x^2y - xh(x^2, y)$, $h(x, y) = xy^2 + 2$

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(xg(x, y) + 5y) = x\frac{\partial g(x, y)}{\partial x} + g(x, y)\frac{\partial x}{\partial x} + 0$$
$$= x\frac{\partial}{\partial x}(x^2y - xh(x^2, y)) + x^2y - xh(x^2, y)$$
$$= x(2xy - x\frac{\partial h(x^2, y)}{\partial x} - h(x^2, y)) + x^2y - x(x^2y^2 + 2)$$
$$= 2x^2y - x^2\frac{\partial(x^2y^2 + 2)}{\partial x} - x(x^2y^2 + 2) + x^2y - x^3y^2 + 2x$$
$$= 2x^2y - 2x^3y^2 - x^3y^2 - 2x + x^2y - x^3y^2 - 2x$$

$$\frac{\partial f}{\partial x} = 3x^2y - 4x^3y^2 - 4x$$

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(xg(x,y) + 5y) = x\frac{\partial g(x,y)}{\partial y} + 5$$
$$= x\frac{\partial}{\partial y}(x^2y - xh(x^2,y)) + 5 = x(x^2 - x\frac{\partial h(x^2,y)}{\partial y}) + 5$$
$$= x^3 - x^2\frac{\partial(x^2y^2+2)}{\partial y} + 5 = x^3 - 2x^4y + 5$$

$$\frac{\partial f}{\partial y} = x^3 - 2x^4y + 5$$

Substituting for $h(x^2, y)$ in $g(x, y)$ and for $g(x, y)$ in $f(x, y)$ is another way to approach the same problem and arrive at the same solution.
$g(x,y) = x^2y - xh(x^2,y) = x^2y - x(x^2y^2+2) = x^2y - x^3y^2 - 2x$
$f(x,y) = xg(x,y) + 5y = x(x^2y - x^3y^2 - 2x) + 5y = x^3y - x^4y^2 - 2x^2 + 5y$

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

Q1.3 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!).

Given, $f(x) = x\log_8(x) + (1-x)\log_8(1-x)$
The extrema (minima and maxima) values occur for a function at the value where its derivative is equal to zero.

Let, $g(x) = x\log_8(x)$ and $h(x) = (1-x)$ so that $f(x) = g(x) + g(h(x))$
Now, to find the points of extrema, we need to find x such that $f'(x) = 0$

$f'(x) = g'(x) + g'(h(x)) \cdot h'(x)$

$g'(x) = \log_8(x) + \frac{1}{ln8}$ and $h'(x) = -1$

$f'(x) = (\log_8(x) + \frac{1}{ln8}) + (-1) \cdot (\log_8(1-x) + \frac{1}{ln8}) = \log_8(x) - \log_8(1-x)$

Now, to find x at extrema, we equate $f'(x) = 0$
$\log_8(x) - \log_8(1-x) = 0$
Solving, we get $x = 1/2$

Now we have to find $f''(x)$ at $x = 1/2$ to determine whether this extrema is minima/maxima
$f''(x) = \frac{1}{xln8} + \frac{1}{(1-x)ln8}$
$f''(1/2) = \frac{2}{ln8} + \frac{2}{ln8} = \frac{4}{ln8} > 0$

Thus we can conclude that minima occurs at $x = 1/2$

In the given range $x \in [0, 1]$ we don't have a minima value that occurs for $f(x)$ such that $f'(x) = 0$ and $f''(x) < 0$
At $x = 0, 1$ we have $f(x) = 0$ and for $x \in (0, 1)$ we have $f(x) < 0$

Thus we conclude that maximum value for $f(x)$ occurs at $x = 0, 1$

Q2 Math Review --- Probability and Statistics

10 Points

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

Q2.1 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.

Probability of removing four orange balls (without replacement) =
probability of first ball removed being orange
* probability of second ball removed being orange given that the first one was orange
* probability of third ball removed being orange given that the first & second were orange
* probability of fourth ball removed being orange given that the first three were orange

$$P(b1, b2, b3, b4 = orange) = P(b1 = orange) * P(b2 = orange | b1 = orange) * P(b3 = orange | b1, b2 = orange) * P(b4 = orange | b1, b2, b3 = orange)$$

$$P(\text{4 orange balls}) = \frac{6}{12} \text{ x } \frac{5}{11} \text{ x } \frac{4}{10} \text{ x } \frac{3}{9} = \frac{1}{2} \text{ x } \frac{5}{11} \text{ x } \frac{2}{5} \text{ x } \frac{1}{3} = \frac{1}{33}$$

Probability of removing four orange balls (without replacement) =

$$\frac{1}{33}$$

Q2.2 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.

We define the following probabilities:
Probability of interaction by Friend A = $P(A)$

Probability of interaction by Friend B $= P(B)$
Probability that topic is computer science $= P(CS)$
Probability that topic is linguistics $= P(LG)$

We have been given the following probabilities:
$P(CS|A) = 0.80$, $P(LG|A) = 0.20$
$P(CS|B) = 0.30$, $P(LG|B) = 0.70$

Also, we know $P(A) = P(B) = p$ (some x to not assume $1/2$)

Now, the probability the note is from Friend A given that we find a typed note on your desk about computer science is defined as $P(A|CS)$

From Bayes' Rule, we get $P(A|CS) = \frac{P(CS|A)P(A)}{P(CS|A)P(A)+P(CS|B)P(B)} = \frac{0.80 \cdot p}{0.80 \cdot p + 0.30 \cdot p}$

$$P(A|CS) = \frac{8}{11}$$

Hence, the probability the note is from Friend A given that we find a typed note on your desk about computer science $= 0.72727$

Q3 Language Modeling

25 Points

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

Q3.1 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.

To find: $p_{katz}(\text{sat}|\text{dog})$

Case 1: context (v) = dog, word (w) = sat
$c(v) = 1, c(v, w) = 1$
$c_d(v, w) = c(v, w) - \beta = 1 - 0.75 = 0.25$

We have $c(v, w) = 1 > 0$ for $w \in A(v)$ where $A(v) = \{w|c(v, w) > 0\}$

So $p_{katz}(\text{sat}|\text{dog}) = \frac{c_d(v,w)}{c(v)} = \frac{0.25}{1}$

$$p_{katz}(\text{sat}|\text{dog}) = 0.25$$

To find: $p_{katz}(\text{sat}|\text{fish})$

Case 2: context (v) = fish, word (w) = sat
To make sure of non-zero context counts, we need to retrain the model by introducing the $< unk >$ token which I am considering in place of the last word of each sentence.
This way, when we encounter fish as a context, it registers as $< unk >$

Now we have $c(v) = 2$ but in the absence of the $< /s >$ token, $c(v, w) = 0$
$A(v) = \emptyset$ where $A(v) = \{w|c(v, w) > 0\}$

Hence, the missing probability mass $\alpha(v)$ is given by:
$\alpha(v) = 1 - \Sigma_{w \in A(v)} \frac{c_d(v,w)}{c(v)} = 1 - 0 = 1$

Now, $B(v)$ contains every word in the training set where $B(v) = \{w|c(v, w) = 0\}$
Since for $w \in B(v), p_{katz}(\text{v}|\text{w}) = \alpha(v)$ x $\frac{p_{mle}(w)}{\Sigma_{w' \in B(v)} p_{mle}(w)}$

$$p_{katz}(\text{sat}|\text{fish}) = 1 \text{ x } \frac{2}{15} = \frac{2}{15}$$

Q3.2 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.

We have $\lambda_1 = \lambda_2 = 0.5$
and $p_{inter}(w|v) = \lambda_1 p_{mle}(w|v) + \lambda_2 p_{mle}(w)$

$p_{inter}(\text{dog}|\text{the}) = \lambda_1 p_{mle}(\text{dog}|\text{the}) + \lambda_2 p_{mle}(\text{dog}) =$
$0.5 \text{ x } \frac{1}{5} + 0.5 \text{ x } \frac{1}{15}$

$$p_{inter}(\text{dog}|\text{the}) = \frac{2}{15}$$

$p_{inter}(\text{dog}|\text{log}) = \lambda_1 p_{mle}(\text{dog}|\text{log}) + \lambda_2 p_{mle}(\text{dog}) =$
$0.5 \text{ x } \frac{0}{1} + 0.5 \text{ x } \frac{1}{15}$

$$p_{inter}(\text{dog}|\text{the}) = \frac{1}{30}$$

Q3.3 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.)

Perplexity is a strictly positive value given by the formula: $pp = 2^{-l}$
where l is the average log probability. The value l is a negative

number that has lower absolute value if the probability of maximum likelihood is closer to 1.0 or 100%.

The lowest probability possible is equal to 0.0, which results in a log probability of $-\infty$. On the other hand, the highest probability that can be acheived is equal to 1.0, which gives a log probability of 0. Thus we can say that $l \in (-\infty, 0]$.

Since $pp = 2^{-l}$, we can say that the value of perplexity is lower bounded by 1 and has no upper bound, hence its range can be given by $[1, \infty)$. The lowest value of $pp = 1$ indicates that the training corpus is best representation of, or basically the same as, the test corpus. On the other hand, $pp \to \infty$ when the training dataset is the worst representation of, or has nothing in common with, the testing dataset.

An example to demonstrate the same:
Training corpus = "we had two pizzas for lunch"
Test corpus 1 = "my dog stepped on a bee"
Test corpus 2 = "we had two pizzas for lunch"

Now we train our model with the training corpus. When this model is tested against test corpus 1, it would give us a 0% probability, or $l \to -\infty$, or $pp \to +\infty$, the maximum possible perplexity score achievable.

However, if we test our trained model on test corpus 2, which is the exact same as the training corpus that was used, we would get a 100% probability, that is $l = 0$, or $pp = 1$, the minimum possible perplexity score achievable.

Q3.4 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?

GENERATED UNIGRAM SENTENCES
Sent 1 : are decay Ah but join . do BISHOP of </s>
Sent 2 : eyes the fair one is . me unjust CHIRON ;
Sent 3 : . extremes 'll Congeal , . more were heaven Having
Sent 4 : and when can due Lead youth know youth reports That
Sent 5 : aim shine My I live apprehend I way we pay

GENERATED TRIGRAM SENTENCES
Sent 1 : Second thunder-like belied Wicked Maudlin Circe blesses Commotions Ieur Error
Sent 2 : First swim damnation mirror Abuse Ascribe Judas desires unhack bow-strings
Sent 3 : Lavinia Prodigal James lament extremity Ptolemy outweighs Rating charters practising
Sent 4 : 'Bless benediction Bad miles school-boy shrift wreathe Voiced anthem crystalline
Sent 5 : First Soldier quarreller salute o' Be't uprighteously well-wish cite new-shed

GENERATED 5-GRAM SENTENCES
Sent 1 : I possets grew Phrygia deputation overtopp Obey sleave-silk ROMEO destroys
Sent 2 : CLEOPATRA bitter-searching Join holier Unfeeling induction bedlam re-salute Wherefore abbey-walls
Sent 3 : RODERIGO fur watered fain Blind eye widow-comfort withstand fat-brained predecessors
Sent 4 : CARDINAL lid daffed Varlet 'Will't finished remember'st edition knowest Confederates
Sent 5 : TOUCHSTONE : abrogate Saw gaps Else hull Allaying divers-colour Officious

Some problems I noticed with the generated sentences is that they all have grammatical errors to an extent, with those created by the unigram model are perhaps making least sense. In fact, the unigram model has many more errors like not capitalizing the first word of the generated sentence, capitalizing words in the middle

of a sentence, or adding periods in the start or middle, or even after other punctuation marks.

The trigram and 5-gram models have nearly similar performance, with the 5-gram model performing slightly better. Although, some words generated by all 3 models have been capitalized even if they're not exactly in the beginning of the sentence. While the trigram or the 5-gram models aren't perfect either, they perform significantly better than the unigram model, showing just how important context is to a sentence, as it should be for any situation.

Q3.5 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.

Authorship identification could probably be performed using the perplexity values determined for the language model on the test corpus, which is the unknown piece of text.
The factors used to manually identify which author created a given piece of text are the terminology utilized and the "style" of writing integrated. Seasoned readers can typically tell what author created a certain book or abstract, and sometimes they can also tell when an author's writing style changes.
To achieve authorship identification in NLP, numerous language models, one for each author, would need to be trained with adequate material provided by those authors. After completing this work, the unknown piece of text may be used as test corpus and used to determine the perplexity for each trained LM model. In this instance, the model with the lowest perplexity score should be able to tell us who wrote the unknown text. In other words, the model that's deviated from its test corpus the least, is the model that represents the source of the unknown piece of text.

Language modeling can therefore be employed for the important task of authorship identification in NLP.

Q4 Late Penalty

0 Points

This problem intentionally left blank.

Homework 1 Written

● UNGRADED

STUDENT

Adithya Iyer

TOTAL POINTS

- / 50 pts

QUESTION 1

Math Review --- Multivariate Calculus		15 pts
1.1	Partial Derivatives	5 pts
1.2	The Chain Rule	5 pts
1.3	Extrema	5 pts

QUESTION 2

Math Review --- Probability and Statistics		10 pts
2.1	Conditional Probability	5 pts
2.2	Bayes's Rule	5 pts

QUESTION 3

Language Modeling		25 pts
3.1	Smoothing --- Discounting and Katz Backoff	5 pts
3.2	Smoothing --- Linear Interpolation	5 pts

3.3	Perplexity	5 pts
3.4	Generation	5 pts
3.5	Applications	5 pts

QUESTION 4

Late Penalty	0 pts