Assignment 1

CMSC 673 — Natural Language Processing

Faisal Rasheed Khan

VB02734

vb02734@umbc.edu

Question 1--

1.

A. i. Directly quoting is acceptable if we use the information from any source by properly citing the source from where it has taken by following citing standards. Directly quoting is acceptable when the information is exact as of the original source (for example it includes theorem, famous quotations etc which can't be altered), In order to use it we need to refer the source from where it has been taken.
For example:
"Our lives sometimes depend on computers performing as predicted." - Philip Emeagwali
The above use of the information is valid as it is directly quoted with appropriate citation.

ii. For Paraphrasing, just changing some words from the source or lightly editing the source even with proper citation is not acceptable. The content from the source should be read and should be completely written in our own words with proper citation standards which would be acceptable.
For Example:
Machine Learning Challenges in the Learning
Machine learning algorithm improves with the help of learning. With a good learning model it performs the best for the data. There will be challenges for the model to learn as sometimes there could be noise in the learning data. The noise would be present both at the feature data or label data. These types of challenges will be there for the machine learning algorithms when they try to learn. (CIML Ch 2.3)
Reference: ciml-v0_99-ch02.pdf

iii. Improper use of the information by near copying from the source with/without citation or rewriting by lightly modifying the text with/without citation leads to academic dishonesty. To overcome this, understand and write the information entirely in your own words. Providing help / getting help from others also leads to academic dishonesty.
Example of improperly using the information from the source:
Not entire thing is Learnable
Although machine learning works well-in many cases, it is important to keep in mind that it is not magical. There are many answers why a machine learning algorithm might fail on some learning task. There could be noise in the training data. Noise can occur both at the feature level and at the label level. (CIML Ch 2.3)
Reference: ciml-v0_99-ch02.pdf

* Providing reference also doesn’t work out as entire information is exact although it is slightly re-written

B. Working in a group while avoiding issues of academic integrity is acceptable when you are not entirely copying each other’s work. Discussion of the ideas is acceptable when you are working in a group. Discussion forums assigned by the instructor and the permissible sharing of ideas are acceptable rather than presenting entire work. And after collaborating together (keeping in mind of academic integrity), The details should also be given with whom you have worked together, failing to provide the details may lead to academic integrity violation. Working in a group should be according to the instructors guidelines.

C. When working in the group everyone is responsible for the entire group work. Ever individual is responsible for the entire work done, can’t make excuses like I haven’t done that section and I don’t know from where my teammate has included that part of the work (without citation).

D. Academic integrity Related to Coding:
   i. Directly copying the code from online resources
   ii. Not providing the appropriate references when you code
   iii. Assisting/taking help(copying) from your friend with the code
   iv. Submitting code which is not written by you, and you didn’t understand.

E. The online resources can be used by maintaining integrity, when you want to understand the concept you can refer, and when you are stuck and getting error in the code you can resolve the issue with the help of the online resources by properly citing the source. You can also understand the solution and try solving in your own way by properly citing the source. Whenever you are using ideas from online (but not copying entire solution), cite those.

F. The online resources could result in academic dishonesty because of copying the entire work from the online platforms. It can also lead to dishonesty if you upload your work so that others can cheat. You just copy the work without learning anything and claiming the work as your own leads to academic dishonesty.

Question 2---

2.

a. The first sentence has 29 tokens and has been verified.
   Total number of sentences in training splits are 12543.
   Total number of sentences in validation splits are 2002.
   Average words per sentence for training sentences is 16.50745435701188.
   Average words per sentence for validation sentences is 12.726273726273726.

b. Total number of different tokens are 207053.
   Total number of different types are 20132.

c. The most common words are articles, punctuations, prepositions, verbs, pronouns, conjunction. Some of the frequent words which are neither from the above grammar are people, back, work etc.

d. The considering of items as distinct depends on a particular Natural Language Processing task. It depends on how we are going to define a model. The consideration of all the related words to its root word will also depend on what task needs that. Considering collapsing method lemmatization, which will assign the similar words to its root word might be helpful as it reduces the vocabulary size but on the other hand it might create not so efficient algorithm as it would miss the important information, for example apple is a fruit and Apple is a company. So, the collapsing method also depends on what type of NLP task you are solving.

e. i. The most common words are articles, punctuations, prepositions, verbs, pronouns, conjunction. From the count of 74 appearances of the words, they look like the standard words. Few words like not, also, yet, Is occurred in between 40-55 times. But the majority of them are present after 73 counts of a word.

ii. The words which has high frequency are the most common words and they make sentences meaningful, so treating them as distinct is reasonable. The words which have the moderate frequency are the words specific to that content, so this may also be reasonable to treat as distinct because content specific new words are learnt. For the rare frequency/rare occurring words, treating them as distinct is not helpful from computational point of view. For the error words which have less frequency, and their correct word has high frequency, decompose the word such that the language model learns these type of error words too. The solution should be based on the above frequency measures.

f. i. Total number of word tokens are 25478.

ii. Out Of Vocabulary (OOV) words are 1706.

iii. The number of OOV are 1706 which would impact the efficiency of the model. The new words are to be added based on their importance for the respective content. Sometimes the OOV increases the vocabulary space by adding those words. The OOV needs to be handled efficiently. With these OOV validation helps assess and sees how the trained model performs on those OOV as it helps in identifying the improvement areas in the model.

Question 3---

3.

a. torch.nn.Linear applies the linear transformation for the input data, it performs the dot product of weight matrix and feature vector, and also adds the bias term with it, if the bias is present.

$y = x * \text{Weight}^T + b$

b. the value of y for [1,0,0] is [3,1] which is the first column vector of the given matrix, similarly for [0,1,0] the value is the second column vector of the given matrix, for [0,0,1] the value is the third column vector of the given matrix. For the input tensor [1,1] its gives error(size 1x2).

The computation is done like a linear layer which is $y = x * \text{Weight}^T + b$, given input matrix is of size is 2x3 and the transpose size is 3x2, therefore the size of a feature vector should be 1x3, that's why for the vector of size 1x2 it gave an error as the computation is done in this linear way $y = x * \text{Weight}^T + b$

c. The value z is a scalar value and if we backpropagate the value z, it computes gradient to the tensor variables w.r.t z. so the tensor variables f.weight and x gradients are computed w.r.t z. The tensor variable represents with small change how it affects z.

Question 4---

4.

a. f_μ(x) = exp($-\frac{1}{3}$(x - μ)^3)

$\frac{\partial(f_\mu(x))}{\partial(\mu)}$= exp($-\frac{1}{3}$(x - μ)^3) * $\frac{\partial\left(-\frac{1}{3}(\mathrm{x}-\mu)\text{^}3\right)}{\partial(\mu)}$ *since,* ∂x[e^x] = e^x

$\frac{\partial(f_\mu(x))}{\partial(\mu)}$= exp($-\frac{1}{3}$(x - μ)^3) * -1 * (x - μ)^2 * $\frac{\partial((\mathrm{x}-\mu))}{\partial(\mu)}$By chain rule, $\frac{d}{d(x)}[f(g(x))]=f'(g(x))g'(x)$

$\frac{\partial(f_\mu(x))}{\partial(\mu)}$= exp($-\frac{1}{3}$(x - μ)^3) * (x - μ)^2 since, ∂x[x^k] = kx^{k-1}

b. x=1.5, $\frac{df}{d\mu}$ | μ=5

exp($-\frac{1}{3}$(1.5 - 5)^3) * (1.5 - 5)^2

19720968

c. f(x) = log(ax^4 + bx^2 -1)

$\frac{\partial(f(x))}{\partial(a)} = \frac{1}{\mathrm{a\,x\text{^}4 + b\,x\text{^}2 - 1}}$ * x^4 since, ∂x[log x] = $\frac{1}{x}$

$\frac{\partial(f(x))}{\partial(a)} = \frac{\mathrm{x\text{^}4}}{\mathrm{a\,x\text{^}4 + b\,x\text{^}2 - 1}}$

$\frac{\partial(f(x))}{\partial(b)} = \frac{1}{\mathrm{a\,x\text{^}4 + b\,x\text{^}2 - 1}}$ * x^2

$\frac{\partial(f(x))}{\partial(b)} = \frac{\mathrm{x\text{^}2}}{\mathrm{a\,x\text{^}4 + b\,x\text{^}2 - 1}}$

d. x=-2, $\frac{df}{da}$ | a=1, $\frac{df}{db}$ | b=1

$\frac{\partial(f(x))}{\partial(a)} = \frac{-2\text{^}4}{-2^4 + -2\text{^}2 - 1}$ = 0.8421

$\frac{\partial(f(x))}{\partial(b)} = \frac{-2\text{^}2}{-2^4 + -2\text{^}2 - 1}$ = 0.2105

e. Given two k-dimensional vectors u and v

$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$ and $\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$ of size kx1

$u^T . v$

$[u_1 \quad u_2 \quad \dots \quad u_k] \quad . \quad \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$

$u^T . v = u_1 * v_1 + u_2 * v_2 + \dots\dots + u_k * v_k$

$u^T . v = \sum_{i=1}^{k} u_i * v_i$

f. Given the matrix A = $\begin{bmatrix} 4 & 2 & 0 \\ -1 & 0 & -1 \end{bmatrix}$

$A^T = \begin{bmatrix} 4 & -1 \\ 2 & 0 \\ 0 & -1 \end{bmatrix}$

$A.A^T$=

C11 = 4 x 4 + 2 x 2 + 0 x 0 = 20

C12 = 4 x (-1) + 2 x 0 + (-1) x 0 = -4

C21 = (-1) x 4 + 0 x 2 + (-1) x 0 = -4

C22 = (-1) x (-1) + 0 x 0 + (-1) x (-1) = 2

$$\begin{bmatrix} 20 & -4 \\ -4 & 2 \end{bmatrix}$$

$A^T.A$=

R11 = 4 x 4 + (-1) x (-1) = 17

R12 = 4 x 2 + 0 x (-1) = 8

R13 = 4 x 0 + (-1) x (-1) = 1

R21 = 2 x 4 + 0 x (-1) = 8

R22 = 2 x 2 + 0 x 0 = 4

R23 = 2 x 0 + 0 x (-1) = 0

R31 = 0 x 4 + (-1) x (-1) = 1

R32 = 0 x 2 + (-1) x 0 = 0

R33 = 0 x 0 + (-1) x (-1) = 1

$$\begin{bmatrix} 17 & 8 & 1 \\ 8 & 4 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

References:

AcademicIntegrityOverview.mp4 - Google Drive

Welcome to PyTorch Tutorials — PyTorch Tutorials 2.0.1+cu117 documentation

5. Data Structures — Python 3.11.5 documentation

Sorting HOW TO — Python 3.11.5 documentation

math4ml.pdf (umd.edu)

ciml-v0_99-ch02.pdf