Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2023-2024
M.Tech. in AIML

Mid-Semester Test
(EC-2 Makeup Paper)

Course No. : AIMLCZG530
Course Title : Natural Language Processing
Nature of Exam : Closed Book
Weightage : 30%
Duration : 2 Hours
Date of Exam : _FN

No. of Pages = 3
No. of Questions = 7

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1. [4 Marks] Introduction

a) In the following sentences: “She can fly airplanes” and “There was a banana fly in fruit basket.” What is this sort of ambiguity referred to as? **[1 mark]**
Solution : Lexical Ambiguity or Grammatical Ambiguity

b) In linguistic morphology, the word “ate” getting reduced to the root form of “eat” is an example of ____________, while the word “apples” getting reduced to the root form of “apple” is an example of ___________. **[1 mark]**
Solution : In linguistic morphology, the word “better” getting reduced to the root form of “good” is an example of **lemmatization,** while the word “friendly” getting reduced to the root form of “friend” is an example of **stemming.**

c) In the context of an NLP application that is processing a sequence of words, name the steps involved during the data pre-processing, which results in the sequence of words getting progressively processed. That is, fill up the right side column of the following table with the correct name of the pre-processing step. **[2 marks]**

Output after processing of sentence	**Name of the pre-processing step**
i. She is better in painting.	Given sentence
ii. She is better in painting	Remove punctuation
iii. She, is, better, in, painting	???
iv. She, better, painting.	???
v. She, good, painting	???
vi. She, good, paint	???

SOLUTION

0.5 marks each for each of the terms mentioned in blue colour below:

Output after processing of sentence		**Name of the pre-processing step**
i.	She is better in painting.	Given
ii.	She is better in painting	Remove punctuation
iii.	She, is, better, in, painting	Tokenization
iv.	She, better, painting.	Removal of Stop Words (is, at)
v.	She, good, painting	Lemmatization (better □ good)
vi.	She, good, paint	Stemming (running □ run)

Question 2. [5 Marks] n –gram language modeling

Akbar Birbal and Panchatantra are stories from India.

The stories of the Akbar Birbal and Panchatantra are translated in many languages

Compute the probability of the sentence using the bigram model and add 1 smoothing for the below sentence. Include start and end symbol in your calculations.

Akbar Birbal and Panchatantra are the stories from India translated in many languages.

Question 3. [5 Marks] Vector Semantics

The embedding matrix in a particular vector representation scheme for words is given below:

$$\begin{bmatrix} 59 & 76 & 64 & 10 & 12 & 31 & 53 \\ 53 & 41 & 79 & 5 & 57 & 43 & 21 \\ 21 & 7 & 36 & 27 & 65 & 31 & 45 \\ 48 & 47 & 30 & 57 & 16 & 20 & 27 \\ 49 & 74 & 34 & 43 & 11 & 64 & 9 \\ 44 & 65 & 13 & 76 & 79 & 80 & 44 \\ 42 & 9 & 82 & 41 & 28 & 65 & 85 \\ 28 & 5 & 21 & 42 & 22 & 60 & 40 \end{bmatrix}$$

The one-hot vector for two words are given below:

One-Hot_W1 = [0,1,0,0,0,0,0,0]

One_Hot_W2 = [0,0,0,0,0,0,1,0]

Answer the following questions:

3.1	Find out the vector embedding for words W1 and W2. Explain the steps.	2 marks
3.2	Calculate the Cosine Distance between these words. All major steps and calculations should be explained.	2 marks

Solutions

3.1	Find out the vector embedding for W1 and W2 Vector embedding for W1 is given by dot product {One_Hot_W1 . Embedding_matrix} W1 = [53,41,79,5,57,43,21] W2 = [42,9,82,41,28,65,85]
3.2	Calculate the Cosine Distance between these words W1 . W2 = [53x42 + 41x9 + 79x82 + 5x41 + 57x28 + 43x65 + 21x85] = 15454 |W1| = SQRT(53^2 + 41^2 + ...) = 127.6519 |W2| = sQRT(42^2 + 9^2 + ...) = 149.9467 Cosine Similarity = (W1 . W2) / (|W1| x |W2|) = 15454 / (127.6519 x 149.9467) = 0.807378

	Cosine Distance = 1 – Cosine Similarity = 1 – 0.807378 = 0.192622

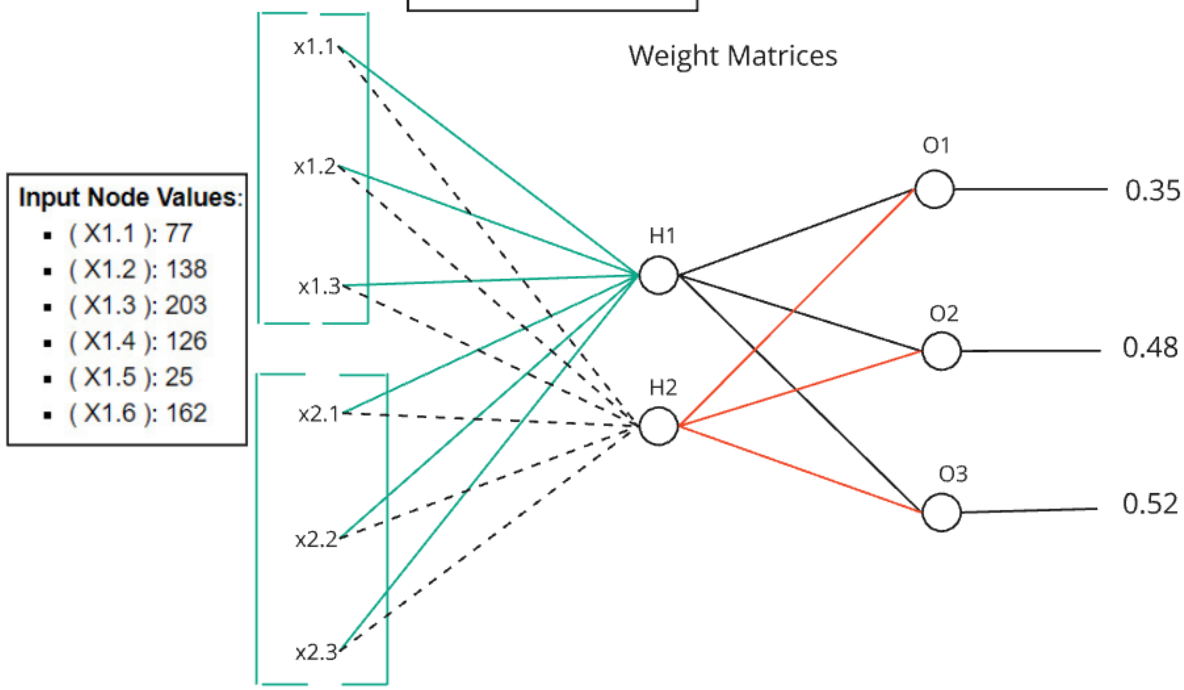
Question 4. [5 Marks]
In an NLP class, the concepts of Neural Network based word prediction models are being explained with the help of an example Neural Network shown below.
It is explained that:

- The vocabulary V consists of three words
- The Neural Network takes the embeddings of two words as inputs:
    - Word_1 = [X1.1, X1.2, X1.3]
    - Word_2 = [X2.1, X2.2, X2.3]
- There are 2 nodes in the hidden layer
- The weights associated with the links are as shown in the Weight Matrices
- There are three output nodes corresponding to the three words in the vocabulary
- Based on the embeddings of the two input (context) words, the NN is supposed to predict the most likely 'following' word.

	(H1)	(H2)
(X1.1)	0.743	0.456
(X1.2)	0.799	0.080
(X1.3)	0.739	0.410
(X2.1)	0.787	0.900
(X2.2)	0.107	0.270
(X2.3)	0.898	0.240

	(O1)	(O2)	(O3)
(H1)	0.435	0.407	0.255
(H2)	0.946	0.761	0.917

Weight Matrices



Now answer the following questions:

4.1	The students are informed by the instructor that the output node values are incorrect. Can you identify what is wrong and why?	1 mark
4.2	It is informed that the values output from the hidden nodes are [1, 1]. It is also mentioned that the hidden node outputs can assume the values in the range [-1,1]. What can you conclude from this information?	1 mark
4.3	Using the stated hidden node values [1,1] calculate the correct values of the output nodes.	2 marks

Answers

4.1	The students are informed by the instructor that the output node values are incorrect. Can you identify what is wrong and why? [1 Mark] The output node values are probabilities and they should add up to '1'. The given values add up to more than '1'	1 mark
4.2	It is informed that the values output from the hidden nodes are [1, 1]. It is also mentioned that the hidden node outputs can assume the values in the range [-1,1]. What can you conclude from this information? The activation function used for the hidden layer is TANH.	1 mark
4.3	Using the stated hidden node values [1,1] calculate the correct values of the output nodes. The activation function at the output layer: SOFTMAX Input to O1 = 1 x 0.435 + 1 x 0.946 = 1.381 Input to O2 = 1 x 0.407 + 1 x 0.761 = 1.168 Input to O3 = 1 x 0.255 + 1 x 0.917 = 1.172 exp(1.381) = 3.9788 exp(1.168) = 3.2155 exp(1.172) = 3.2284 Sum = exp(1.381) + exp(1.168) + exp(1.172) = 10.4227 Output at O1 = 3.9788/10.4227 = 0.3817 Output at O2 = 3.2155/10.4227 = 0.3085 Output at O3 = 3.2284/10.4227 = 0.3097	2 marks

Question 5.

Consider the following sentence:

"It always seems impossible until it's done."

Your task is to train a classifier such that, given the tuple ("impossible", "until") where "impossible" is the target word and "until" is the candidate context word, the classifier returns the probability that "until" is a real context word for "impossible". Give detail computation steps and provide the updated Input weight matrix for the Target Word after one iteration of the Word2Vec algorithm **[5 marks]**

The following information are provided:

- Use Word2Vec with Skip Gram Classifier with a Single Hidden Layer
- Use learning rate as 0.05 and Sigmoid as activation function
- Negative Sampling words have been specified as " Table", "Window"

Initial Embedding Matrix for the Single Hidden Layer

Impossible	0.05	0.01	0.03
Until	0.04	0.02	0.25
Table	0.01	0.05	0.05
Window	0.01	0.05	0.04

Initial Context Embedding Matrix for the Output Layer

Impossible	0.01	0.02	0.05
Until	0.04	0.03	0.04
Table	0.01	0.01	0.05
Window	0.02	0.01	0.04

Solution:

Step 1 – Forward Propagation (Hidden Layer) [<u>1 mark</u>]

- The One Hot Encoded Input Matrix: (I)
  [1000
  0100
  0010
  0001]

- Initial Embedding Matrix (W_{input})
  [0.05 0.01 0.03
  0.04 0.02 0.25
  0.01 0.05 0.05
  0.01 0.05 0.04]

	Context Embedding			Input word embedding	dot product	Sigmoid()	t	Sig - t	
Until	0.04	0.03	0.04	0.05	0.0035	0.500874999	1	-0.499125001	**Positive Context Word**
Table	0.01	0.01	0.05	0.01	0.0021	0.499475	0	0.499475	**Negative Word**
Window	0.02	0.01	0.04	0.03	0.0023	0.499425	0	0.499425	**Negative Word**
Context Embedding				Sig - t	C*(Sig-t)				
Until	0.04	0.03	0.04	-0.499125001	-0.019965	-0.01497375	-0.019965		**Positive Context Word**
Table	0.01	0.01	0.05	0.499475	0.00499475	0.00499475	0.02497375		**Negative Word**
Window	0.02	0.01	0.04	0.499425	0.0099885	0.00499425	0.019977		**Negative Word**
				Derivative of log loss with respect to input embedding	-0.00498175	-0.00498475	0.02498575		
				LR	0.05				
				Update step	-0.000249088	-0.000249238	0.001249287		
				Updated Embedd	**0.050249088**	**0.010249**	**0.028751**		
				Ans (3 decimals	**0.050**	**0.010**	**0.029**		

- Hidden Layer (h) for Target word "Impossible" = $W^T_{input} * I$
  = [0.05
  0.01
  0.03]

Step 2 – Forward Propagation (Sigmoid Output Layer) [2 marks]

- W_{output} (context) *for (until,table,window)*

  [0.04 0.03 0.04
  0.01 0.01 0.05
  0.02 0.01 0.04]

  Output Layer = W_{output} * h

  = [0.0035
  0.0021
  0.0023]

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

$$\boxed{\begin{aligned} P(-|w,c) &= 1 - P(+|w,c) \\ &= \sigma(-c \cdot w) = \frac{1}{1+\exp(c \cdot w)} \end{aligned}}$$

Applying Sigmoid Activation,

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma(x) = \frac{1}{1+e^{x}}$$

$\sigma(output\ layer)$ = [0.5009
0.4995
0.4994]

Step 3 – Prediction Error [1 mark]

Prediction Error = $\sigma(output\ layer)$ – 1-hot encoded vector for context

$$= \begin{bmatrix} 0.5009 \\ 0.4995 \\ 0.4994 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.499 \\ 0.499 \\ 0.499 \end{bmatrix}$$

Backward Propagation (computing W_{input}) step:

Derivative of Loss with respect to Input Word Embeddings for the target word “Impossible”:

[-0.0050 -0.0050 0.0250]

(W_{input})

Step 4 - Updated Weight Matrix by applying Learning Rate [1 mark]

Learning Rate = 0.05 (given)

W^{new}_{input} = [0.05   0.01   0.03] – 0.05 * W_{input} = **[0.050   0.010   0.029] Expected Answer**

This the updated Input weight matrix for the Target Word (Impossible) after one iteration of the Word2Vec algorithm

Question 6. [5 Marks]

a) Use the following (part of speech) lexicon:

adult	JJ	has	VBZ
adult	NN	just	RB
daughter	NN	my	PRP$
developed	VBD	programs	NNS
developed	VBN	programs	VBZ
first	JJ	tooth	NN
first	RB	whose	WP$

Consider the following sentence: my daughter whose first adult tooth has just developed programs
With this lexicon, how many different PoS tagging does this sentence have? Justify your answer.

<u>**Answer**</u>:

my	PRP$	
daughter	NN	
whose	WP$	
first	JJ	RB
adult	JJ	NN
tooth	NN	
has	VBZ	
just	RB	
developed	VBN	VBD
programs	NNS	VBZ

2x2x2x2=16 possible taggings
Examples:
my/PRP\$ daughter/NN whose/WP\$ first/JJ adult/JJ tooth/NN has/VBZ just/RB developed/VBN programs/VBZ
my/PRP\$ daughter/NN whose/WP\$ first/JJ adult/JJ tooth/NN has/VBZ just/RB developed/VBN programs/NN

Question 7. [4 Marks]

Fill up the Viterbi table for the sentence – ‘He can’. The tag transition probabilities and word emission probabilities, for the corpus used, are given below:

Tag transition probabilities	MD	VB	PRP
MD	0.02	0.6	0.003
VB	0.004	0	0.02
PRP	0.85	0.09	0.001
START	0.09	0.29	0.6

Word emission	He	can
MD	0	0.6
VB	0	0
PRP	0.9	0

Viterbi Table	He	can
VB		
MD		
PRP		

PRP: PERSONAL PRONOUN

MD:MODAL

VB:VERB BASE FORM

Answer:

Viterbi Table	He	Can
VB	0	0
MD	0	0.2754
PRP	0.54	0