FIT3181/5215 Deep Learning

Quiz for:
Word Embedding with Word2Vec

Teaching team

Department of Data Science and AI
Faculty of Information Technology, Monash University
Email: **trunglm@monash.edu**

Question 1

In Word2Vec with the skip-gram, consider the sentence:

"the mouse runs away from the cat"*. Which of the following statements is correct if we use a window size of 7?*

- A. 'The', 'mouse', 'runs', 'from', 'the', 'cat' are used to predict 'away'
- B. 'Runs' is used to predict 'The', 'mouse', 'away', 'from', 'the', 'cat'
- C. 'Runs', 'away' are used to predict 'The', 'mouse', 'from', 'the', 'cat'
- D. 'Away' is used to predict 'the', 'mouse', 'runs', 'from', 'the', 'cat'

Question 1

In Word2Vec with the skip-gram, consider the sentence:

“the mouse runs away from the cat”*. Which of the following statements is correct if we use a window size of 7?*

- A. ‘The’, ‘mouse’, ‘runs’, ‘from’, ‘the’, ‘cat’ are used to predict ‘away’
- B. ‘Runs’ is used to predict ‘The’, ‘mouse’, ‘away’, ‘from’, ‘the’, ‘cat’
- C. ‘Runs’, ‘away’ are used to predict ‘The’, ‘mouse’, ‘from’, ‘the’, ‘cat’
- D. ‘Away’ is used to predict ‘the’, ‘mouse’, ‘runs’, ‘from’, ‘the’, ‘cat’ **[x]**

Question 2

In Word2Vec with the CBOW, consider the sentence:

"the mouse runs away from the cat"*. Which of the following statements is correct?*

- ❑ A. 'The', 'mouse', 'runs', 'from', 'the', 'cat' are used to predict 'away'
- ❑ B. 'Runs' is used to predict 'The', 'mouse', 'away', 'from', 'the', 'cat'
- ❑ C. 'Runs', 'away' are used to predict 'The', 'mouse', 'from', 'the', 'cat'
- ❑ D. 'Away' is used to predict 'the', 'mouse', 'runs', 'from', 'the', 'cat'

Question 2

In Word2Vec with the CBOW, consider the sentence:

"the mouse runs away from the cat"*. Which of the following statements is correct?*

- A. 'The', 'mouse', 'runs', 'from', 'the', 'cat' are used to predict 'away' **[x]**
- B. 'Runs' is used to predict 'The', 'mouse', 'away', 'from', 'the', 'cat'
- C. 'Runs', 'away' are used to predict 'The', 'mouse', 'from', 'the', 'cat'
- D. 'Away' is used to predict 'the', 'mouse', 'runs', 'from', 'the', 'cat'

Question 3

You were given a corpus of texts which is assumed to be sufficiently large to learn semantic meaning of words. After applying word2vec embedding on this corpus, each vocabulary had now learned and associated with a vector. As you have learned from the lecture, this now allows you to perform analogical reasoning. Which answer do you expect if we reason

*"**? – prince = queen - princess**"*

- ❑A. King
- ❑B. Royal
- ❑C. Father
- ❑D. Dad

Question 3

You were given a corpus of texts which is assumed to be sufficiently large to learn semantic meaning of words. After applying word2vec embedding on this corpus, each vocabulary had now learned and associated with a vector. As you have learned from the lecture, this now allows you to perform analogical reasoning. Which answer do you expect if we reason

*“**? – prince = queen - princess**”*

- ❑A. King **[x]**
- ❑B. Royal
- ❑C. Father
- ❑D. Dad

Question 4

Consider an embedding approach in text analytics using the word2vec algorithm with a simple text corpus consisting of the following 9 words:

the quick brown fox jumps over the lazy dog

Tokenizing this corpus give us a vocabulary (i.e., unique words in the corpus) as well as their indices. Assume the above corpus returns the following

***word-to-indices dictionary = {'brown': 0, 'lazy': 1, 'over': 2, 'fox': 3, 'dog': 4, 'quick': 5, 'the': 6, 'jumps': 7}**.*

*What is the one-hot encoding vector for the word **dog**?*

- ❑A. [0,0,0,0,1,0,0,0]
- ❑B. [0,0,0,0,0,1,0,0]
- ❑C. [0,0,0,0,1,1,0,0]
- ❑D. [0,0,0,0,0,0,0,1]

Question 4

Consider an embedding approach in text analytics using the word2vec algorithm with a simple text corpus consisting of the following 9 words:

the quick brown fox jumps over the lazy dog

Tokenizing this corpus give us a vocabulary (i.e., unique words in the corpus) as well as their indices. Assume the above corpus returns the following

***word-to-indices dictionary = {'brown': 0, 'lazy': 1, 'over': 2, 'fox': 3, 'dog': 4, 'quick': 5, 'the': 6, 'jumps': 7}**.*

*What is the one-hot encoding vector for the word **dog**?*

- ❑A. [0,0,0,0,1,0,0,0] **[x]**
- ❑B. [0,0,0,0,0,1,0,0]
- ❑C. [0,0,0,0,1,1,0,0]
- ❑D. [0,0,0,0,0,0,0,1]

Question 5

Given a Wikipedia text corpus dataset with 1 billion sentences. We filtered the dataset and found 10k unique words. We train the word2vec model using the dataset, with the context window size 5. Which are correct statements? (MC)

- ❑ A. Dictionary size N=1 billion
- ❑ B. Dictionary size N=10k
- ❑ C. Each word will be represented by a vector size [5, 1]
- ❑ D. The embedding size is unknown with given information

Question 5

Given a Wikipedia text corpus dataset with 1 billion sentences. We filtered the dataset and found 10k unique words. We train the word2vec model using the dataset, with the context window size 5. Which are correct statements? (MC)

- ❑ A. Dictionary size N=1 billion
- ❑ B. Dictionary size N=10k **[x]**
- ❑ C. Each word will be represented by a vector size [5, 1]
- ❑ D. The embedding size is unknown with given information **[x]**

Question 6

Which are the drawbacks of skip-gram algorithm. (MC)

- ❑ A. The softmax output layer is computationally expensive.
- ❑ B. The values of prediction probabilities are small and hard to distinguish among classes.
- ❑ C. Skip-gram cannot capture the dependency between words.
- ❑ D. Skip-gram depends on the context window size.

Question 6

Which are the drawbacks of skip-gram algorithm. (MC)

- ❑ A. The softmax output layer is computationally expensive. **[x]**
- ❑ B. The values of prediction probabilities are small and hard to distinguish among classes. **[x]**
- ❑ C. Skip-gram cannot capture the dependency between words.
- ❑ D. Skip-gram depends on the context window size. **[x]**

Question 7

Which are correct statements about skip-gram with negative sampling. (MC)

- ❑ A. The output layer represents the probability over classes and sum of output equal to 1.
- ❑ B. We apply softmax activation function at the output layer.
- ❑ C. We apply sigmoid activation function at the output layer.
- ❑ D. Given a pair target word tw and context word cw, we need to randomly sample some negative words and maximize the probabilities to predict those negative words from the target word.
- ❑ E. Given a pair target word tw and context word cw, we need to randomly sample some negative words and minimize the probabilities to predict the pairs of target word and negative words as a positive pair.
- ❑ F. Given a pair target word tw and context word cw, we need to maximize the probability to predict the pair of target word and context word as a positive pair.

Question 7

Which are correct statements about skip-gram with negative sampling. (MC)

- ❑ A. The output layer represents the probability over classes and sum of output equal to 1.
- ❑ B. We apply softmax activation function at the output layer.
- ❑ C. We apply sigmoid activation function at the output layer. **[x]**
- ❑ D. Given a pair target word tw and context word cw, we need to randomly sample some negative words and maximize the probabilities to predict those negative words from the target word.
- ❑ E. Given a pair target word tw and context word cw, we need to randomly sample some negative words and minimize the probabilities to predict the pairs of target word and negative words as a positive pair. **[x]**
- ❑ F. Given a pair target word tw and context word cw, we need to maximize the probability to predict the pair of target word and context word as a positive pair. **[x]**

Question 8

Given an objective function below, $p(u|v)$ represents for the probability to base on the word v to predict the word u. Which are correct statements (MC)

$$J(\theta) = \prod_{t=1}^{T} \prod_{-C \le j \le C, j \ne 0} p(w_{t+j} \mid w_t; \theta)$$

- A. It is the objective function of CBOW algorithm
- B. It is the objective function of skip-gram algorithm
- C. **j** represents for the index of target (centre) word
- D. **t** represents for the index of target (centre) word
- E. We need to maximize this objective function to find parameter theta

Question 8

Given an objective function below, $p(u|v)$ represents for the probability to base on the word v to predict the word u. Which are correct statements (MC)

$$J(\theta) = \prod_{t=1}^{T} \prod_{-C \le j \le C, j \ne 0} p(w_{t+j} \mid w_t; \theta)$$

- ❑ A. It is the objective function of CBOW algorithm
- ❑ B. It is the objective function of skip-gram algorithm **[x]**
- ❑ C. **j** represents for the index of target (centre) word
- ❑ D. **t** represents for the index of target (centre) word **[x]**
- ❑ E. We need to maximize this objective function to find parameter theta **[x]**

Question 9

Given an objective function below, $p(u|v)$ represents for the probability to base on the word v to predict the word u. Which are correct statements (MC)

$$J(\theta) = \prod_{t=1}^{T} p(w_t \mid w_{t-C}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+C}; \theta)$$

- ❑ A. It is the objective function of CBOW algorithm
- ❑ B. It is the objective function of skip-gram algorithm
- ❑ C. T is the window size
- ❑ D. C is the window size
- ❑ E. $2C + 1$ is the window size

Question 9

Given an objective function below, $p(u|v)$ represents for the probability to base on the word v to predict the word u. Which are correct statements (MC)

$$J(\theta) = \prod_{t=1}^{T} p(w_t \mid w_{t-C}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+C}; \theta)$$

- ❑ A. It is the objective function of CBOW algorithm **[x]**
- ❑ B. It is the objective function of skip-gram algorithm
- ❑ C. T is the window size
- ❑ D. C is the window size
- ❑ E. $2C + 1$ is the window size **[x]**

Question 10

Given a skip-gram model with vocabulary size 200 and embedding size 100, we consider a pair of target and context words with indices 5 and 10 respectively. Let U and V be two weight matrices connecting input to hidden layers and hidden to output layers. What statements are correct? (MC)

- A. Shape of U is [200,100] and shape of V is [100,200]
- B. Shape of U is [100,200] and shape of V is [200,100]
- C. Input to the network is one-hot vector 1_5.
- D. Input to the network is one-hot vector 1_{10}.
- E. The hidden value h is the row 10 of the matrix U
- F. The hidden value h is the row 5 of the matrix U

Question 10

Given a skip-gram model with vocabulary size 200 and embedding size 100, we consider a pair of target and context words with indices 5 and 10 respectively. Let U and V be two weight matrices connecting input to hidden layers and hidden to output layers. What statements are correct? (MC)

- ❑ A. Shape of U is [200,100] and shape of V is [100,200] **[x]**
- ❑ B. Shape of U is [100,200] and shape of V is [200,100]
- ❑ C. Input to the network is one-hot vector 1_5. **[x]**
- ❑ D. Input to the network is one-hot vector 1_{10}.
- ❑ E. The hidden value h is the row 10 of the matrix U
- ❑ F. The hidden value h is the row 5 of the matrix U **[x]**

Question 11

Given a CBOW model with vocabulary size 100 and embedding size 150, we consider a target word with index 5 and context words with indices 10, 20, 30, 40 respectively. Let U and V be two weight matrices connecting input to hidden layers and hidden to output layers. What statements are correct? (MC)

- ❑ A. Shape of U is [100,100] and shape of V is [150,150]
- ❑ B. Shape of U is [100,150] and shape of V is [150,100] **[x]**
- ❑ C. Input to the network is one-hot vector 1_5.
- ❑ D. Input to the network is $\frac{1_{10}+1_{20}+1_{30}+1_{40}}{4}$. **[x]**
- ❑ E. The hidden value h is the average of rows $10,20,30,40$ of the matrix U **[x]**
- ❑ F. The hidden value h row 5 of the matrix U

Question 12

1xN	Nxd	1xd	dxN	1xN	1xN
quick	U	$\boldsymbol{h} = \boldsymbol{1}_t\boldsymbol{U}$	V	$\boldsymbol{o} = \boldsymbol{Vh}$	$p(cw\|tw)$

$$\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix}^T \quad \begin{bmatrix}.4 & .3 & .0\\.3 & .5 & .0\\.2 & .3 & .8\\.1 & .3 & .3\\.3 & .7 & .0\\.1 & .9 & .9\\.0 & .4 & .7\\.3 & .2 & .7\end{bmatrix} \Rightarrow h = [0.1\ 0.9\ 0.9] \quad \begin{bmatrix}.5 & .0 & .3 & .3 & .7 & .6 & .3 & .5\\.2 & .4 & .7 & .6 & .8 & .7 & .1 & .6\\.7 & .7 & .8 & .1 & .5 & .7 & .8 & .9\end{bmatrix} \quad \begin{bmatrix}0.45\\0.53\\0.89\\0.48\\0.67\\1.32\\0.77\\0.81\end{bmatrix}^T \xrightarrow{\text{softmax}} \begin{bmatrix}.09\\.10\\.14\\.09\\.11\\.21\\.12\\.13\end{bmatrix}^T$$

$w_t = 1_t$

dog: $w_{t+1} = \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix}^T$

the: $w_{t-1} = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\end{bmatrix}^T$

'fox': 4, 'dog': 5, **'quick': 6**, 'the': 7, 'jumps': 8

Consider the above skip-gram model, let denote $p_1 = p(the \mid quick)$ and $p_2 = p(brown \mid quick)$. What statements are correct? (MC)

- ❑ A. Vocabulary size is 8 and embedding size is 3.
- ❑ B. Vocabulary size is 3 and embedding size is 8.
- ❑ C. $p_1 = 0.09, p_2 = 0.10.$
- ❑ D. $p_1 = 0.45, p_2 = 0.77.$
- ❑ E. $p_1 = 0.12, p_2 = 0.11.$

Question 12

1xN	Nxd	1xd	dxN	1xN	1xN
quick	**U**	$\boldsymbol{h} = \boldsymbol{1}_t\boldsymbol{U}$	**V**	$\boldsymbol{o} = \boldsymbol{Vh}$	$p(cw \mid tw)$

$$w_t = 1_t = \begin{bmatrix}0\\0\\0\\0\\0\\1\\0\\0\end{bmatrix}^T \qquad U = \begin{bmatrix}.4 & .3 & .0\\ .3 & .5 & .0\\ .2 & .3 & .8\\ .1 & .3 & .3\\ .3 & .7 & .0\\ .1 & .9 & .9\\ .0 & .4 & .7\\ .3 & .2 & .7\end{bmatrix} \Rightarrow h = [0.1\ 0.9\ 0.9]$$

$$V = \begin{bmatrix}.5 & .0 & .3 & .3 & .7 & .6 & .3 & .5\\ .2 & .4 & .7 & .6 & .8 & .7 & .1 & .6\\ .7 & .7 & .8 & .1 & .5 & .7 & .8 & .9\end{bmatrix}$$

$$o = \begin{bmatrix}0.45\\0.53\\0.89\\0.48\\0.67\\1.32\\0.77\\0.81\end{bmatrix}^T \xrightarrow{\text{softmax}} \begin{bmatrix}.09\\.10\\.14\\.09\\.11\\.21\\.12\\.13\end{bmatrix}^T$$

dog: $w_{t+1} = \begin{bmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{bmatrix}^T$

the: $w_{t-1} = \begin{bmatrix}0\\0\\0\\0\\0\\0\\1\\0\end{bmatrix}^T$

'fox': 4, 'dog': 5, **'quick': 6**, 'the': 7, 'jumps': 8

Consider the above skip-gram model, let denote $p_1 = p(the \mid quick)$ and $p_2 = p(brown \mid quick)$. What statements are correct? (MC)

- ❑ A. Vocabulary size is 8 and embedding size is 3. **[x]**
- ❑ B. Vocabulary size is 3 and embedding size is 8.
- ❑ C. $p_1 = 0.09, p_2 = 0.10.$
- ❑ D. $p_1 = 0.45, p_2 = 0.77.$
- ❑ E. $p_1 = 0.12, p_2 = 0.11.$ **[x]**