CS698O: Quiz-2

Name: ______________________________

Roll No.: ______________________________

Please fill the square in the answer sheet with blue ink. Check/Tick marks are NOT allowed.

---

1. Consider the sigmoid activation function defined by $\sigma(\theta) = \frac{1}{1+exp(-\theta)}$. What is the derivate of $\sigma(\theta)$ w.r.t θ, i.e. $\sigma'(\theta) := \frac{d\sigma(\theta)}{d\theta}$ Fill all that you think are correct.

   A: $1 - \sigma(\theta)$

   B: $\sigma(\theta) * (1 + \sigma(\theta))$

   C: $\sigma(\theta) * (1 - \sigma(\theta))$

   D: $1 + \sigma(\theta)$

---

2. Which of the following are **NOT** true about distributed representations? Fill all that you think are correct.

   A: Number of units in a distributed representation scales linearly (i.e. $O(N)$) with the number of concepts (N).

   B: Distributed representations of words capture the semantic and syntactic properties of the words.

   C: Distributed representations can be explained by the phrase “All for one and one for all”.

   D: Distributed representations can easily be interpreted.

---

3. Consider the distributed representation for the word “January”:

2	0	3	1	0	2	0	0

   Which of the following would be the closest distributed representation for the word “June”?

   A:

3	0	2	2	0	1	0	0

   B:

1	0	1	1	0	1	0	0

   C:

5	0	-2	1	0	2	0	0

   D:

2	0	3	1	0	-10	0	0

4. Consider a corpus with vocabulary $\mid \mathcal{V} \mid = 10000$. A trigram language model is developed using this corpus. What is the range of values that the perplexity ($\mathcal{P}$) of this trigram model can take? Fill all that you think are correct.

A: $0 \leq \mathcal{P} \leq \infty$ **B:** $1 \leq \mathcal{P} \leq \infty$

C: $-\infty \leq \mathcal{P} \leq \infty$ **D:** $10 \leq \mathcal{P} \leq 10^3$

Consider the following training corpus ($\mathcal{D}$):

```
--------------------------------------
I am John
I am out today
John I am
Mary I am
The cat ran
John and cat ran
The cat ran after the mouse
--------------------------------------
```

For all the questions related to this corpus assume $\mathcal{V} = \{\mathcal{V} \cup \texttt{STOP}\}$

5. For a trigram language model trained on corpus ($\mathcal{D}$), approximately how many parameters need to be estimated?

A: $\sim 15^3$ **B:** $\sim 14^3$

C: $\sim 13^3$ **D:** $\sim 12^3$

6. What will be the MLE based probability estimate (i.e. $p(S_1)$) for the following sentence using a unigram language model trained on $\mathcal{D}$?

$S_1 = \texttt{John ran after the cat}$

A: $\frac{189}{33^6}$ **B:** $\frac{567}{33^6}$

C: $\frac{189}{26^6}$ **D:** $\frac{567}{26^6}$

7. What will be the MLE based probability estimate (i.e. $p(S_2)$) for the following sentence using a Laplace smoothed unigram language model trained on $\mathcal{D}$?

$S_2 = \texttt{Lion and Bear ran}$

A: 0 **B:** $\frac{64}{47^5}$

C: $\frac{32}{47^4}$ **D:** $\frac{32}{46^5}$

8. What is the MLE estimate of $p(\texttt{cat}|\texttt{and})$ and $p(\texttt{The}|\texttt{START})$ using laplace smoothed bigram language model trained on on $\mathcal{D}$? Fill all that you think are correct.

A: $\frac{2}{14}, \frac{3}{20}$

B: $\frac{2}{15}, \frac{3}{21}$

C: $1, \ 0$

D: $\frac{2}{15}, \frac{2}{7}$

9. For a corpus with vocabulary size, $\mathcal{V} = 1000$, two n-gram based language models M_1, and M_2 are trained. Perplexity of model M_1 is 970 and perplexity of model M_2 is 200. Which of the following statements are correct? Fill all that you think are correct.

A: The conditional n-gram probability distribution of model M_2 is close to uniform.

B: Model M_2 is not a good language model as compared to M_1

C: Model M_1 is not a good language model as compared to M_2

D: Cross Entropy of model M_2 is $\log_2 200$

10. Which of the following are true about Entropy $(H(X))$ of probability distributions? Fill all that you think are correct.

A: Entropy of uniform distribution between the interval $(0, 10)$ is more than the entropy of a gaussian distribution with mean $= 0$ and standard deviation $= 2$

B: For a discrete distribution, entropy is minimized when each discrete event is equally likely.

C: For any distribution over random variable X, $-\infty \leq H(X) \leq \infty$

D: Entropy is upper bounded by Cross Entropy.

11. Which of the following are true about log-linear models used in NLP? Fill all that you think are correct.

A: Log-linear models used in majority of NLP applications use indicator feature functions. Reason for using indicator features is to make the computation of conditional label probability efficient.

B: Log-linear models used in majority of NLP applications use indicator feature functions. Reason for using indicator features is that linguistic features can only be represented using indicator functions.

C: Log-linear models suffer from feature vector sparsity problems.

D: Log-linear models can overfit.

12. Word frequencies vs word rank in natural languages follows which of the following relationships? Fill all that you think are correct.

A: Power law relationship

B: Exponential relationship

C: Square relationship

D: Linear relationship

13. Log linear models are typically regularized. Reason for doing so is: (Fill all that you think are correct)

A: Regularization helps to overcome overfitting.

B: Regularization helps to let the parameter values to grow and have large values.

C: Regularization helps to overcome underfitting.

D: Regularization helps to checks the growth of parameters corresponding to infrequent features.

14. Consider an NLP classification problem. We would like to use log-linear model for classification. Give an input x we want to predict three classes $y = 1, y = 2, y = 3$. Let the parameters of the log-linear model be $\Theta = [\theta_1, \theta_2, \theta_3, \theta_4]^T$. The feature vectors corresponding to different classes are as follows:

$\phi(x, y = 1) = [1, 0, 0, 0]^T$
$\phi(x, y = 2) = [1, 0, 1, 0]^T$
$\phi(x, y = 3) = [1, 0, 0, 1]^T$

Which of the following are correct? (Fill all that you think are correct)

A: $p(y = 3 \mid x, \Theta) = \frac{exp(\theta_1+\theta_4)}{exp(\theta_1)+exp(\theta_1+\theta_3)+exp(\theta_1+\theta_4)}$

B: $p(y = 3 \mid x, \Theta) = \frac{exp(\theta_1+\theta_3)}{exp(\theta_1)+exp(\theta_1+\theta_3)+exp(\theta_1+\theta_4)}$

C: $p(y = 3 \mid x, \Theta) = \frac{exp(\theta_1)}{exp(\theta_1)+exp(\theta_1+\theta_3)+exp(\theta_1+\theta_4)}$

D: $p(y = 3 \mid x, \Theta) = \frac{exp(\theta_1+\theta_2)}{exp(\theta_1)+exp(\theta_1+\theta_3)+exp(\theta_1+\theta_4)}$

15. Which of the following is NOT true about MaxEnt models? (Fill all that you think are correct)

A: MaxEnt models try to minimize the entropy over subset of events not seen during training.

B: MaxEnt models try to maximize the entropy over subset of events not seen during training.

C: In MaxEnt models the expected value (w.r.t. conditional distribution) of an observed feature vector should match the average value of that feature vector observed in training.

D: For covering longer context in language models, maxent models can be used.

16. Consider a classification problem in NLP with 4 classes. We use a neural network with final softmax activation for doing the classification. Let, that the final hidden layer vector is x and the softmax weight parameter vector is θ. Let $z_j = \theta_j^T x$. Which of the following are correct? (Fill all that you think are correct)

A: $p(y = j \mid x) = \frac{exp(z_j)}{\prod_{i=1}^{4} exp(z_i)}$

B: $p(y = j \mid x) = \frac{exp(z_j)}{\sum_{i=1}^{4} exp(z_i)}$

C: $p(y = j \mid x) = exp\left(z_j - \log\left(\sum_{i=1}^{4} exp(z_i)\right)\right)$

D: $p(y = j \mid x) = exp(z_j) - \log\left(\sum_{i=1}^{4} exp(z_i)\right)$

17. Consider the softmax function ($\sigma(\mathbf{z})$). Which of the following are true? (Fill all that you think are correct)

A: $\sigma(\mathbf{z} + \mathbf{c_z}) = \sigma(\mathbf{z})$, where c_z is a function of vector z.

B: $\frac{d\sigma(\mathbf{z})}{dz} = \sigma(\mathbf{z}) * (1 + \sigma(\mathbf{z}))$

C: As temperature increases softmax function tends to become more uniform and spread out

D: As temperature increases softmax function tends to become more peaky at the maximum value

18. Consider sequence of R.V.s in the following order: $X_1 \to X_2, \ldots, \to X_n$. Which of the following are correct? (Fill all that you think are correct)

A: $P(X_k \mid X_{k-1}, X_{k-2}, X_{k-3}) = P(X_k \mid X_{k-1})$

B: $P(X_1, X_2, \ldots, X_n) = P(X_1) * \prod_{i=2}^{n} P(X_i \mid X_{i-1})$

C: $P(X_1, X_2, \ldots, X_n) = P(X_i \mid X_{i-1}, \cdots, X_1) * P(X_{i-1} \mid X_{i-2}, \cdots, X_1) \cdots P(X_2 \mid X_1) * P(X_1)$

D: $P(X_2 \mid X_1) = \frac{P(X_2, X_1)}{P(X_1)}$

19. For a neural network classifier with softmax activation, at the test time the biggest bottleneck is the normalization at the output. What are some of the techniques that can be used to overcome this? (Fill all that you think are correct)

A: Dividing words into classes and using class conditional model

B: Mapping words below certain frequency threshold to a special symbol

C: Mapping each word to its distributed representation

D: Using negative sampling

20. Which of the following is true about word classes? (Fill all that you think are correct)

A: List of open class words is static and does not change over time.

B: Closed class words includes nouns and verbs.

C: Nouns are open class words.

D: Verbs belong to closed class words.

21. Which of the following is/are examples of prepositional phrase attachment ambiguity? (Fill all that you think are correct)

A: `He saw a man with a telescope.`

B: `He saw people standing in the queue.`

C: `He ate dinner with spoon.`

D: `Jack asked him to leave with her.`

22. Consider three R.V.s: X, Y, Z. Which of the following are true? (Fill all that you think are correct)

A: $P(Y \mid X) = \sum_Z P(Y \mid X, Z)$

B: $P(Y \mid X) = \sum_Z P(Y, Z \mid X)$

C: $P(Y \mid X) = \sum_Z P(Z \mid X, Y)$

D: $P(Y \mid X) = \sum_Z P(Y \mid X) P(Z \mid X)$

23. Language models suffer from sparsity issues. Which of the following techniques can be used to overcome these? (Fill all that you think are correct)

A: Re-distributing the probability mass among zero count n-grams.

B: By extrapolating the unknown counts.

C: By weighted product of lower order language models.

D: By backing off to lower order language models.

24. Consider the objective function of a log-linear model. What is the purpose of parameter α. (Fill all that you think are correct)

$$\mathcal{L} = \sum_{i=1}^{n} \log(p(y^{(i)} \mid x^{(i)}; \theta)) + \alpha \mid \theta \mid$$

A: As $\alpha \to 0$, the model tends to overfit

B: As $\alpha \to \infty$, the model tends to overfit

C: As $\alpha \to 0$, the model tends to underfit

D: As $\alpha \to \infty$, the model tends to underfit

25. Consider a class based language model. Suppose that vocabulary ($\mathcal{V} = 20$) is divided into 4 classes. The total number of words in the corpus $N = 100$. The number of words in each class is as follows:
$\mid c_1 \mid = 4, \mid c_2 \mid = 4, \mid c_3 \mid = 7, \mid c_4 \mid = 5$ Additionally, the frequency count for two of the words is: C(`John`) = 5, C(`Ball`) = 2. Moreover, $\{$`John, Ball` $\} \in c_3$. Which of the following are true? (Fill all that you think are correct)

A: θ_{MLE} (`John` $\mid c_3) = \frac{5}{100}$

B: θ_{MLE} (`Ball` $\mid c_3) = \frac{2}{7}$

C: θ_{MLE} (`John` $\mid c_3) = \frac{5}{7}$

D: θ_{MLE} (`Ball` $\mid c_3) = \frac{2}{100}$

26. Log-linear models are typically evaluated using cross entropy loss ($\mathcal{CE}$). Alternatively log-linear models can also be evaluated using Negative Log Likelihood ($\mathcal{NLL}$) loss. Which of the following is true? (Fill all that you think are correct)

A: $\mathcal{NLL} \leq \mathcal{CE}$

B: $\mathcal{CE}$ can never be zero

C: $\mathcal{NLL} \geq \mathcal{CE}$

D: $\mathcal{NLL} = \mathcal{CE}$

27. Distributional Hypothesis states that: (Fill all that you think are correct)

A: Each unit in a distributed representation should contribute towards the meaning of the concept.

B: A distributed representation for a concept cannot be reconstructed from its parts.

C: A distributed representation of word should capture the semantic and syntactic properties of the word.

D: A word is known by the company it keeps

28. For a trigram language model, which of the following are true? Fill all that you think are correct.

A: $\sum_{w_i \in \mathcal{V}} C(w_i, w_{i-1}, w_{i-2}) = C(w_{i-1}, w_{i-2})$

B: $\sum_{w_i \in \mathcal{V}} C(w_i, w_{i-1}, w_{i-2}) \geq C(w_{i-1}, w_{i-2})$.

C: $\sum_{w_i \in \mathcal{V}} C(w_i, w_{i-1}) = C(w_{i-1})$

D: $\sum_{w_i \in \mathcal{V}} C(w_i, w_{i-1}) \geq C(w_{i-1})$

29. In a language modeling paradigm typically how would you deal with words which are out of vocabulary? Fill all that you think are correct.

A: Discard such words.

B: Map the low frequency words in training data to special symbols and map the new words during testing to this special symbol

C: Use skip-k gram model, where k is the number of unknown words

D: Use Kneser-key smoothing

30. Consider three R.V.s: X, Y, Z. Suppose, X and Y are conditionally independent given Z. Which of the following are true? (Fill all that you think are correct)

A: $P(X \mid Z) = P(X)$

B: $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

C: $P(Y \mid X) = P(Y)$

D: $P(Y, X) = P(X)P(Y)$