

CS698O: Quiz-1

Name: _____

Roll No.: _____

Please fill the square with blue ink. Check/Tick marks are NOT allowed. Each question is of 2 marks.

1. Consider sequence of R.V.s in the following order: $X_1 \rightarrow X_2, \dots, \rightarrow X_n$. Assume that these RV.s follow first order Markov assumption. Which of the following are correct? Fill all that you think are correct.

☒ $P(X_k | X_{k-1}, X_{k-2}, X_{k-3}) = P(X_k | X_{k-1})$

☐ $P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | X_{i-1})$

☒ $P(X_1, X_2, X_3) = P(X_3 | X_2) \times P(X_2, X_1)$

☐ $P(X_3 | X_2) = P(X_2 | X_1)$

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

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

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

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

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

3. Which of the following statements about Maximum Likelihood Estimate (MLE) for Language Model are correct? Fill all that you think are correct.

- ☒ MLE estimates suffer from overfitting
 - ☐ MLE underestimates probabilities for n-grams with high counts and overestimates probabilities for n-grams with low counts
 - ☐ MLE does not suffer from sparsity problems
 - ☐ MLE cannot be done for language models
-

4. Which of the following are correct MLE estimates for parameters of n(=1 or 2 or 3 or 4)-gram language models? Fill all that you think are correct.

- ☐ $\theta_{MLE}(w_i) = \frac{C(w_i)}{|\mathcal{V}|}$
 - ☐ $\theta_{MLE}(w_i | w_{i-1}) = \frac{C(w_i, w_{i-1})}{C(w_i)}$
 - ☒ $\theta_{MLE}(w_i | w_{i-1}, w_{i-2}) = \frac{C(w_i, w_{i-1}, w_{i-2})}{\sum_{w_i \in \mathcal{V}} C(w_i, w_{i-1}, w_{i-2})}$
 - ☒ $\theta_{MLE}(w_i | w_{i-1}, w_{i-2}, w_{i-3}, w_{i-4}) = \frac{C(w_i, w_{i-1}, w_{i-2}, w_{i-3}, w_{i-4})}{C(w_{i-1}, w_{i-2}, w_{i-3}, w_{i-4})}$
-

5. Which of the following are correct about perplexity measure? Fill all that you think are correct.

- ☐ The higher the perplexity of a language model the better
 - ☒ The lower the perplexity of a language model the better
 - ☐ Perplexity of a LM can never be infinite
 - ☒ Perplexity of a LM can never be zero
-

6. Which of the following are correct definition(s) of perplexity? Fill all that you think are correct.

☒ $\left(\prod_{i=1}^m p(S_i) \right)^{-\frac{1}{M}}$

☒ $2^{-\left(\frac{1}{M} \sum_{i=1}^m \log_2 p(S_i) \right)}$

☐ $\left(\frac{1}{M} \prod_{i=1}^m p(S_i) \right)^{-\frac{1}{M}}$

☐ $2^{\left(\frac{1}{M} \sum_{i=1}^m \log_2 p(S_i) \right)}$

7. Which of the following strategies can be used for overcoming MLE limitations for LM? Fill all that you think are correct.

☒ Discounting

☐ Extrapolation

☒ Class Based Clustering of Words

☐ Look Ahead Technique

8. For a trigram language model, what are the **exact** number of parameters that need to be estimated? Assume $\mathcal{V} = \{STOP, \mathcal{V}\}$. Fill all that you think are correct.

☐ $|\mathcal{V}|^3$

☒ $|\mathcal{V}|^3 - 1$

☐ $|\mathcal{V}|^2$

☐ $|\mathcal{V}|$

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

- ☐ Power law relationship
- ☐ Exponential relationship
- ☐ Square relationship
- ☒ Linear relationship

10. Consider the following corpus:

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

What is the MLE estimate of $p(STOP \mid cat, ran)$ and $p(John \mid START, START)$ respectively. Fill all that you think are correct.

- ☐ $\frac{0}{3}, \frac{0}{6}$
 - ☐ $\frac{2}{4}, \frac{2}{7}$
 - ☒ $\frac{2}{3}, \frac{2}{7}$
 - ☐ $\frac{0}{3}, \frac{1}{7}$
-