Lecture Questions

1 Language Model 1

1.1 Question (time: 6:17)

Say we have a vocabulary $\mathcal{V} = \{\text{the}\}$ and a constant $N \geq 1$.

For any $x_1 \dots x_n$ such that $x_i \in \mathcal{V}$ for $i = 1 \dots (n-1)$ and $x_n = \text{STOP}$, we define $p(x_1, \dots, x_n) = \begin{cases} \frac{1}{N} & \text{if } n \leq N \\ 0 & \text{otherwise} \end{cases}$

Is this a valid language model?

(a) True

(b) False

1.2 Question (time: 6:17)

Say we have a vocabulary $\mathcal{V} = \{\text{the, dog}\}$.

For any $x_1 \dots x_n$ such that $x_i \in \mathcal{V}$ for $i = 1 \dots (n-1)$ and $x_n = \text{STOP}$, we define $p(x_1, \dots, x_n) = \begin{cases} \frac{1}{2} & \text{if } n = 2 \\ 0 & \text{otherwise} \end{cases}$

Is this a valid language model?

(a) True

(b) False

2 Markov Process 1

2.1 Question (time: 2:47)

Consider a Markov process with states $\mathcal{V} = \{0, 1, 2\}$ and length $n = 10$.

How many different sequences can be generated by this process?

(a) 2^{10}

(b) 10^2

(c) 3^{10}

(d) 10^3

3 Trigram

3.1 Question (time: 5:12)

Say we have a language model with $\mathcal{V} = \{\text{the, dog, runs}\}$, and the following parameters:

- $q(\text{the}|*, *) = 1$
- $q(\text{dog}|*, \text{the}) = 0.5$
- $q(\text{STOP}|*, \text{the}) = 0.5$
- $q(\text{runs}|\text{the}, \text{dog}) = 0.5$
- $q(\text{STOP}|\text{the}, \text{dog}) = 0.5$
- $q(\text{STOP}|\text{dog}, \text{runs}) = 1$

How many sentences have non-zero probability under this model?

3.2 Question (time: 7:01)

Consider the following corpus of sentences:

- the dog walks STOP
- walks the dog STOP
- dog walks fast STOP

Let q_{ML} be the maximum-likelihood parameters of a trigram language model trained on this corpus. Which of the following parameters have a value that is both well-defined and non zero?

(a) $q_{\text{ML}}(\text{walks}|\text{dog}, \text{the})$

(b) $q_{\text{ML}}(\text{fast}|\text{dog}, \text{the})$

(c) $q_{\text{ML}}(\text{walks}|*, \text{dog})$

(d) $q_{\text{ML}}(\text{STOP}|\text{walks}, \text{dog})$

(e) $q_{\text{ML}}(\text{dog}|\text{walks}, \text{the})$

(f) $q_{\text{ML}}(\text{walks}|\text{the}, \text{dog})$

4 Perplexity

4.1 Question (time: 6:37)

Define a trigram language model with the following parameters:

- $q(\text{the}|*, *) = 1$, $q(\text{dog}|*, \text{the}) = 0.5$
- $q(\text{cat}|*, \text{the}) = 0.5$, $q(\text{walks}|\text{the}, \text{cat}) = 1$
- $q(\text{STOP}|\text{cat}, \text{walks}) = 1$, $q(\text{runs}|\text{the}, \text{dog}) = 1$
- $q(\text{STOP}|\text{dog}, \text{runs}) = 1$

Now consider a test corpus with the following sentences:

- the dog runs STOP, the cat walks STOP, the dog runs STOP

What is the perplexity of the language model on this test corpus to three decimal places? (Note: use log_2 for your calculations. Note that the number of words in this corpus, M, is equal to 12)

5 Linear Interpolation 2

5.1 Question (time: 2:21)

We are given the following corpus:

- the green book STOP
- my blue book STOP
- his green house STOP
- book STOP

Assume we compute a language model based on this corpus using linear interpolation with $\lambda_i = 1/3$ for all $i \in \{1, 2, 3\}$.

What is the value of the parameter $q_{\text{LI}}(\text{book}|\text{the, green})$ in this model to three decimal places?

(Note: please include STOP words in your unigram model.)

5.2 Question (time: 5:07)

Say that we train a language model using linear interpolation with $\lambda_1 = -0.5$, $\lambda_2 = 0.5$, and $\lambda_3 = 1.0$. Note that these values satisfy the constraint $\sum_i \lambda_i = 1$ but violate the constraint $\lambda_i \geq 0$.

What problems might occur in the resulting language model? Check all that apply.

(a) we may have a bigram u, v such that $\sum_{w \in \mathcal{V}} q(w|u,v) \neq 1$

(b) we may have a trigram u, v, w such that $q(w|u,v) < 0$

(c) we may have a trigram u, v, w such that $q(w|u,v) > 1$

6 Discounting Methods 1

6.1 Question (time: 6:09)

Assume that we are given a corpus with the following properties:

- Count(the) = 70
- $|\{w : c(\text{the}, w) > 0\}| = 15$, i.e. there are 15 different words that follow "the".

Furthermore assume that discounted counts are defined as $c^*(\text{the}, w) = c(\text{the}, w) - 0.3$.

Under this corpus, what is the missing probability mass, $\alpha(\text{the})$, to three decimal places?

6.2 Question (time: 9:27)

Let's return to a smaller version of our corpus.

- the book STOP
- his house STOP

This time we compute a bigram language model using Katz back-off with $c^*(v, w) = c(v, w) - 0.5$.

What is the value of $q_{\text{BO}}(\text{book}|\text{his})$ estimated from this corpus?

A Answers

- (1.1) a
- (1.2) a
- (2.1) c
- (3.1) 3
- (3.2) c e f
- (4.1) 1.189
- (5.1) 0.571
- (5.2) b c
- (6.1) 0.064
- (6.2) 0.1