

CS540 Introduction to Artificial Intelligence

Lecture 9

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 6, 2021

Coordination Game

Admin

- Q, M, P Grades are updated.
- The ordering of the lectures are updated.
- Last year's exams are fixed.
- Anonymous feedback on Socrative, Room CS540A.

$$\begin{array}{r} 0.5 \times 3 = 1.5 \\ + 0.5 \\ \hline 2.0 \end{array}$$

Q, pick A.

Remind Me to Start Recording

Admin

- Change your Zoom name to your favorite movie or TV show character (add a random number at the end to avoid repetition).
 - For the next question, you will not be allowed to communicate in chat.

Coordination Game

Admin



- You are not allowed to discuss anything about this question in chat. There will be around 10 new questions on the final exam. I will post n of them before the exam (before next Tuesday):
 - A: $n = 0$.
 - B: $n = 1$ if more than 50 percent of you choose B.
 - C: $n = 2$ if more than 75 percent of you choose C.
 - D: $n = 3$ if more than 95 percent of you choose D.
 - E: $n = 0$.
- I will repeat this question a second time. If you fail to coordinate both times, I will not post any of the new questions.

Coordination Game Repeat

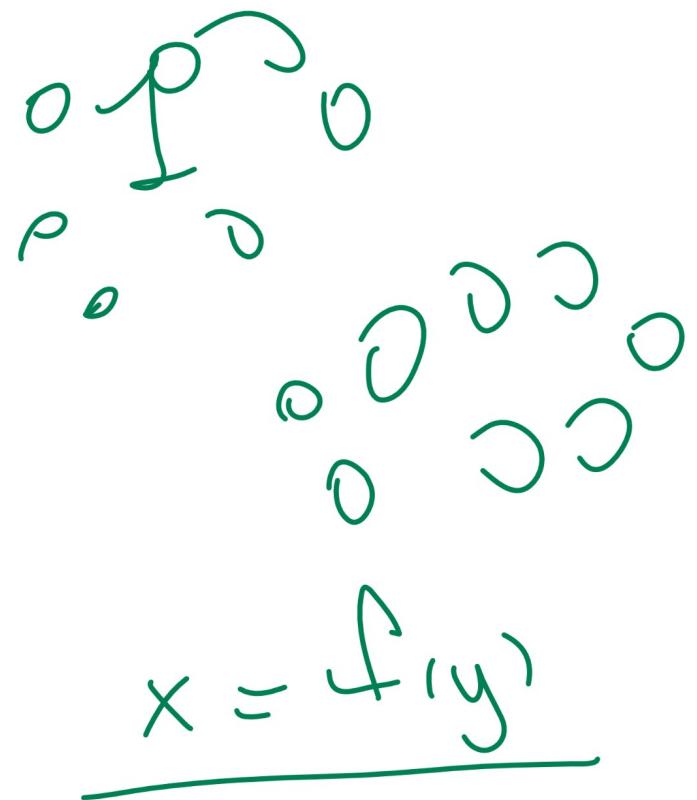
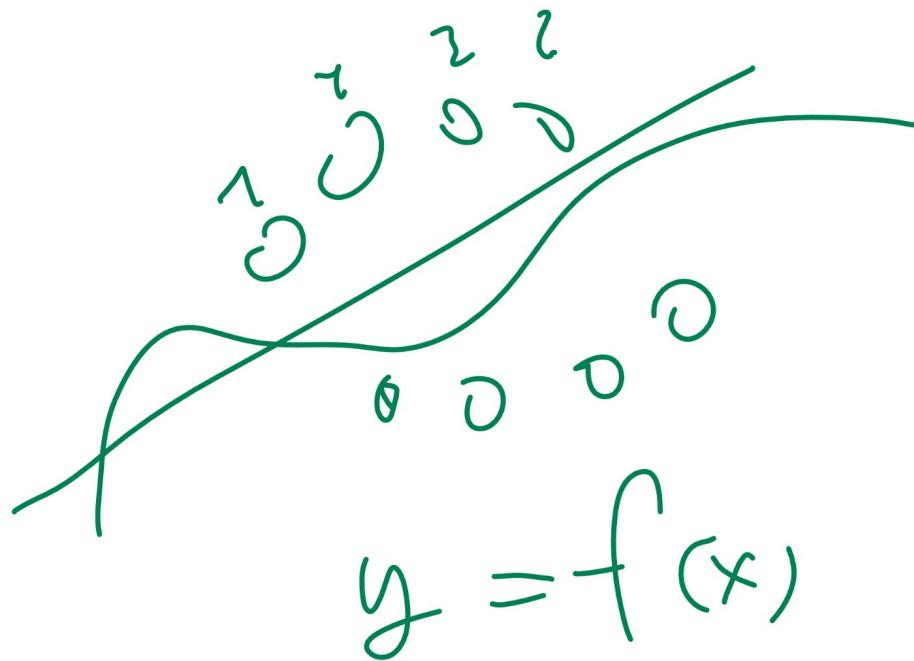
Admin

Q4

- You are not allowed to discuss anything about this question in chat. There will be around 10 new questions on the ~~final~~ ^{midterm} exam. I will post n of them before the exam:
 - A: $n = 0$.
 - B: $n = 1$ if more than 50 percent of you choose B.
 - C: $n = 2$ if more than 75 percent of you choose C.
 - D: $n = 3$ if more than 95 percent of you choose D.
 - E: $n = 0$.

Discriminative Model vs Generative Model

Motivation



Generative Models

Motivation

- In probability terms, discriminative models are estimating $\mathbb{P}\{Y|X\}$, the conditional distribution. For example, $a_i \approx \mathbb{P}\{y_i = 1|x_i\}$ and $1 - a_i \approx \mathbb{P}\{y_i = 0|x_i\}$. $f(x | w)$
- Generative models are estimating $\mathbb{P}\{Y, X\}$, the joint distribution.
- Bayes rule is used to perform classification tasks.

$$\mathbb{P}\{Y|X\} = \frac{\mathbb{P}\{Y, X\}}{\mathbb{P}\{X\}} = \frac{\mathbb{P}\{X|Y\}\mathbb{P}\{Y\}}{\mathbb{P}\{X\}}$$

Bag of Words Features

Definition

- Given a document i and vocabulary with size m , let c_{ij} be the count of the word j in the document i for $j = 1, 2, \dots, m$.
 - Bag of words representation of a document has features that are the count of each word divided by the total number of words in the document.
(word token)

the document. word token.

$$x_{ij} = \frac{c_{ij}}{m}$$

$$= \frac{\sum_{j'=1}^m c_{ij'}}{m}$$

instance item i feature j

word types

Bag of Words Features Example

Motivation

- Given a training set, the set of documents is called a corpus. Suppose the set is "I am Groot", "I am Groot", ... (9 times), "We are Groot". The vocabulary is "I" "am" "Groot" "we" "are", then the bag of words features will have the following training set.

The diagram shows a Bag of Words feature matrix. The columns represent words in the vocabulary: "I", "am", "Groot", "we", and "are". The rows represent documents. The matrix has 9 rows, corresponding to the 9 occurrences of the corpus. Handwritten annotations include:

- A bracket above the first three rows labeled "I am Groot".
- A bracket above the last two rows labeled "We are".
- A bracket on the right side of the matrix labeled "9 times".
- A bracket on the far left labeled "1 3".
- An arrow pointing to the first row labeled x_1 .
- Handwritten values in the matrix cells: the first three rows have "1" in the "I" column and "3" in the "am" column; the last two rows have "1" in the "we" column and "3" in the "are" column; the other cells are "0".

	I	am	Groot	we	are
1	1	1	1	0	0
2	1	1	1	0	0
3	1	1	1	0	0
...
7	1	1	1	0	0
8	1	1	1	0	0
9	0	0	1	1	1

image

pixel
intensity

NLP

bag of
words

TF IDF Features

Definition

- Another feature representation is called tf-idf, which stands for normalized term frequency, inverse document frequency.

SIFT
HOG
Haar

TFIDF

$$\text{tf}_{ij} = \frac{c_{ij}}{\max_{j'} c_{ij'}}, \quad \text{idf}_j = \log \frac{n}{\sum_{i=1}^n \mathbb{1}_{\{c_{ij}>0\}}}$$

CRNN RNN

$$x_{ij} = \text{tf}_{ij} \text{idf}_j$$

- n is the total number of documents and $\sum_{i=1}^n \mathbb{1}_{\{c_{ij}>0\}}$ is the number of documents containing word j .

Unigram Model Definition

↑
|
N gram

- Unigram models assume independence.

$$\mathbb{P}\{z_1, z_2, \dots, z_d\} = \prod_{t=1}^d \mathbb{P}\{z_t\}$$

- In general, two events A and B are independent if:

$$\mathbb{P}\{A|B\} = \mathbb{P}\{A\} \text{ or } \mathbb{P}\{A, B\} = \mathbb{P}\{A\} \mathbb{P}\{B\}$$

- For a sequence of words, independence means:

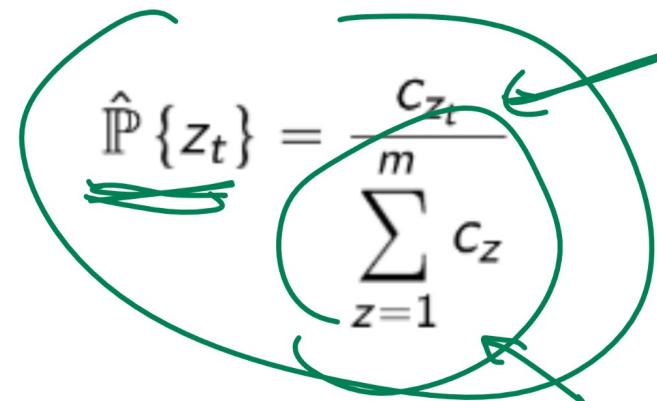
$$\mathbb{P}\{z_t | z_{t-1}, z_{t-2}, \dots, z_1\} = \mathbb{P}\{z_t\}$$

word at
position t
in a sentence.

Maximum Likelihood Estimation

Definition

- $\mathbb{P}\{z_t\}$ can be estimated by the count of the word z_t .

$$\hat{\mathbb{P}}\{z_t\} = \frac{c_{z_t}}{m} \sum_{z=1} c_z$$


- This is called the maximum likelihood estimator because it maximizes the probability of observing the sentences in the training set.

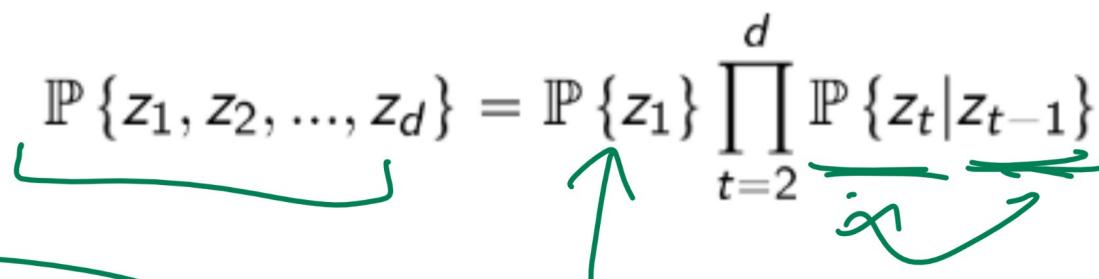
MLE Derivation

Definition

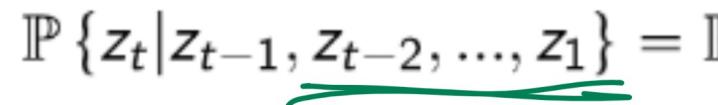
Bigram Model

Definition

- Bigram models assume Markov property.

$$\mathbb{P}\{z_1, z_2, \dots, z_d\} = \mathbb{P}\{z_1\} \prod_{t=2}^d \mathbb{P}\{z_t | z_{t-1}\}$$


- Markov property means the distribution of an element in the sequence only depends on the previous element.

$$\mathbb{P}\{z_t | z_{t-1}, z_{t-2}, \dots, z_1\} = \mathbb{P}\{z_t | z_{t-1}\}$$




Bigram Model Estimation

Definition

- Using the conditional probability formula, $\mathbb{P}\{z_t|z_{t-1}\}$, called transition probabilities, can be estimated by counting all bigrams and unigrams.

$$\hat{\mathbb{P}}\{z_t|z_{t-1}\} = \frac{c_{z_{t-1}, z_t}}{c_{z_{t-1}}}$$

The equation is annotated with handwritten green text and arrows. Above the first $\hat{\mathbb{P}}$, there is a circled 'Qm' and a circled 'I'. Above the second $\hat{\mathbb{P}}$, there is a circled 'Qm'. To the right of the fraction, there is a circled 'Qm'. Below the denominator, there is a circled 'P'. A green arrow points from the circled 'Qm' above the first $\hat{\mathbb{P}}$ to the circled 'Qm' above the second $\hat{\mathbb{P}}\}$. Another green arrow points from the circled 'I' to the circled 'P'.

Unigram MLE Probability

Quiz

- Given the training data "I am Groot am I", with the unigram model, what is the probability of observing a new sentence "I am I"?

- A: $\frac{2}{5}$
- B: $\frac{2}{25}$
- C: $\frac{4}{25}$
- D: $\frac{4}{125}$
- E: $\frac{8}{125}$

$$\Pr\{I\} \Pr\{am\} \Pr\{I\}$$

MLE

$$\frac{C_I}{\# \text{tokens}}$$

$$\frac{C_{am}}{\# \text{tokens}}$$

$$\frac{C_{am}}{\# \text{tokens} \text{ in train}}$$

$$\left(\frac{2}{5}\right)^2$$

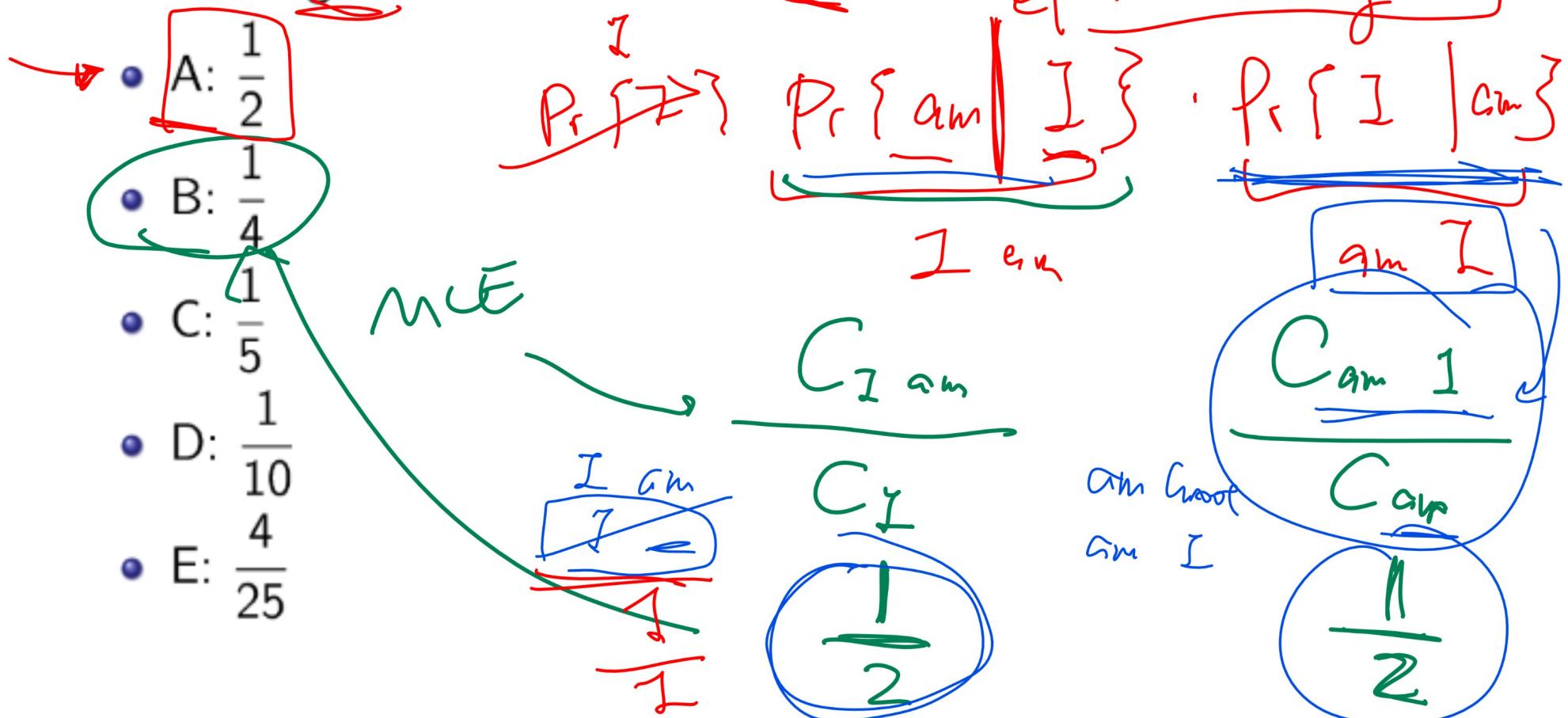
$$\frac{2}{5}$$

$$\rightarrow \frac{2}{125}$$

Bigram MLE Probability

Quiz

- Given the training data "I am Groot am I" with the bigram model, what is the probability of observing a new sentence "I am I" given the first word is "I"?



Unigram MLE Probability

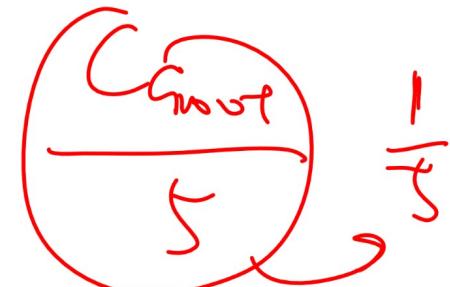
Quiz

- Q5
- Given the training data "I am Groot am I", with the unigram model, what is the probability of observing a new sentence "I am Groot"?

- A: $\frac{2}{5}$
- B: $\frac{2}{25}$
- C: $\frac{4}{25}$
- D: $\frac{4}{125}$
- E: $\frac{8}{125}$

$$P_{\text{G}}\{I\} \quad P_{\text{G}}\{a_n\} \quad P_{\text{G}}\{\text{Groot}\}$$

$$\frac{2}{5} \quad \frac{2}{5}$$



Bigram MLE Probability

Quiz

- Q6
- Given the training data "I am Groot am I" with the bigram model, what is the probability of observing a new sentence "I am Groot" given the first word is "I"?

- A: $\frac{1}{2}$

- B: $\frac{1}{4}$

- C: $\frac{1}{5}$

- D: $\frac{1}{10}$

- E: $\frac{4}{25} = \frac{1}{2}$

$$\Pr(I, I) = \Pr\{am | I\} \cdot \Pr\{Groot | am\}$$

$$\frac{1}{2} \cdot \frac{C_{2am}}{C_I} \rightarrow \frac{1}{2}$$

$$\frac{C_{sh\ Groot}}{C_{am}} = 1$$

$$\frac{C_{am}}{C_I} = 2$$

Transition Matrix

Definition

- These probabilities can be stored in a matrix called transition matrix of a Markov Chain. The number on row j column j' is the estimated probability $\hat{P}\{j'|j\}$. If there are 3 tokens $\{1, 2, 3\}$, the transition matrix is the following.

$$\xrightarrow{\text{Row}} \begin{matrix} & \overset{\Sigma}{\text{1}} & \overset{\alpha_m}{\text{2}} & \overset{\text{Cross}}{\text{3}} \\ \underset{\text{Col}}{\text{1}} & \left[\begin{array}{ccc} \hat{P}\{1|1\} & \hat{P}\{2|1\} & \hat{P}\{3|1\} \\ \hat{P}\{1|2\} & \hat{P}\{2|2\} & \hat{P}\{3|2\} \\ \hat{P}\{1|3\} & \hat{P}\{2|3\} & \hat{P}\{3|3\} \end{array} \right] \\ \xrightarrow{\text{Cross}} & \quad & \quad & \end{matrix}$$

- Given the initial distribution of tokens, the distribution of the next token can be found by multiplying it by the transition probabilities.

Estimating Transition Matrix

Definition

Suppose the vocabulary is "I", "am", "Groot", "we", "are", and the training set contains 9 "I am Groot" then 1 "We are Groot". Then the transition matrix is:

-	I	am	Groot	we	are
I	0	1	0	0	0
am	0	0	1	0	0
Groot	8 — 9	0	0	1 — 9	0
we	0	0	0	0	1
are	0	0	1	0	0

MCF

Trigram Model

Definition

- The same formula can be applied to trigram: sequences of three tokens.

$$\hat{P}\{z_t | z_{t-1}, z_{t-2}\} = \frac{c_{z_{t-2}, z_{t-1}, z_t}}{c_{z_{t-2}, z_{t-1}}} \quad \text{MLE}$$

C I am Great
C I am

- In a document, likely, these longer sequences of tokens never appear. In those cases, the probabilities are $\frac{0}{0}$. Because of this, Laplace smoothing adds 1 to all counts.

regularization

$$\hat{P}\{z_t | z_{t-1}, z_{t-2}\} = \frac{c_{z_{t-2}, z_{t-1}, z_t} + 1}{c_{z_{t-2}, z_{t-1}} + m}$$

$$\sum_e (c_{z_{t-2}, z_{t-1}, e} + 1)$$

Laplace Smoothing

Definition

- Laplace smoothing should be used for bigram and unigram models too.

$$\hat{\mathbb{P}}\{z_t | z_{t-1}\} = \frac{c_{z_{t-1}, z_t} + 1}{c_{z_{t-1}} + m}$$
$$\hat{\mathbb{P}}\{z_t\} = \frac{c_{z_t} + 1}{\sum_{z=1}^m c_z + m}$$

- Aside: Laplace smoothing can also be used in decision tree training to compute entropy.

Smoothing Example

Quiz

- Fall 2018 Midterm Q12.
- Given a vocabulary of 10^6 , a document with 10^{12} tokens with $c_{\text{zoodles}} = 3$. What is the MLE estimation of $\mathbb{P}\{\text{zoodles}\}$ with and without Laplace smoothing?

Smoothing Example 2

Quiz

- Given the training instance with 9 "I am Groot" followed by 1 "We are Groot", what is the MLE estimation of $\mathbb{P}\{\text{Groot}\}$ with Laplace smoothing?
- A: $\frac{1}{2}$
- B: $\frac{11}{35}$
- C: $\frac{1}{3}$
- D: $\frac{11}{31}$
- E: $\frac{1}{4}$

Smoothing Example 3

Quiz

- Given the training instance with 9 "I am Groot" followed by 1 "We are Groot", what is the MLE estimation of $\mathbb{P}\{\text{Groot} \mid \text{I}\}$ with Laplace smoothing?
- A: $\frac{1}{10}$
- B: $\frac{1}{11}$
- C: $\frac{1}{14}$
- D: $\frac{1}{15}$
- E: 0

Sampling from Discrete Distribution

Discussion

- To generate new sentences given an N gram model, random realizations need to be generated given the conditional probability distribution.
- Given the first $N - 1$ words, z_1, z_2, \dots, z_{N-1} , the distribution of next word is approximated by $p_x = \hat{\mathbb{P}}\{z_N = x | z_{N-1}, z_{N-2}, \dots, z_1\}$. This process then can be repeated for on $z_2, z_3, \dots, z_{N-1}, z_N$ and so on.

CDF Inversion Method Diagram

Discussion

Generating New Words 1

Quiz

- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "I" and a uniform random variable $u = 0.5$ is produced. What is the next word?

$$\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$$

Generating New Words 2

Quiz

- Given the transition matrix for words "I" "am" "Groot", starting a sentence with the "I am" and a uniform random variable $u = 0.75$ is produced. What is the next word?

$$\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$$

- A: I, B: am, C: Groot