

CS540 Introduction to Artificial Intelligence

Lecture 7

Young Wu

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

July 12, 2022

Exam Date

Admin

Q1

→ (● July 26 AND 27 (in person and or online) from 1 : 00 to 2 : 30.


(● July 27 (online only, with the other section) from 5 : 30 to 8 : 30. ←

- A : I will be available on July 26 AND 27 from 1 : 00 to 2 : 30.
- B : I will be available on July 27 from 5 : 30 to 8 : 30.
- C : I am not available on both dates (email me).

Midterm Details

Admin

- July 26 AND July 27 from 1 : 00 to 2 : 30:

- 
- 1 Complete the exam online at home and join by Zoom for announcements.
 - 2 Complete the exam online in person here, bring your laptop.
 - 3 Request a paper copy of the exam and submit the answer sheet (I need to know the number of exams to print).

Midterm Coverage

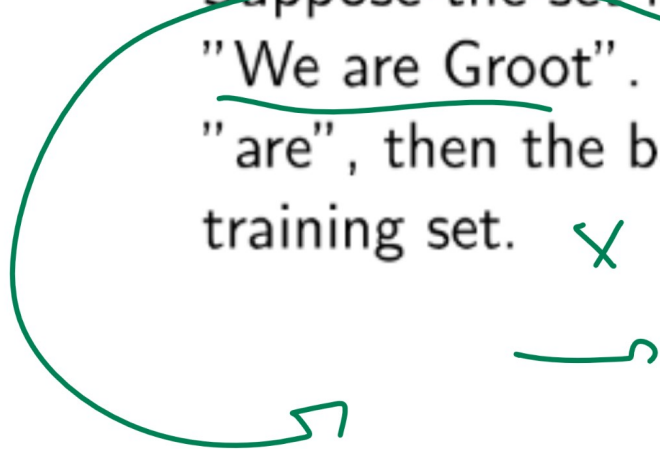
Admin

- More midterm-related details next ~~Monday~~^{Thur}:
- ① ~ 10 questions from $M2$ to $M7$ (same question different randomization).
- ② ~ 10 questions from relevant questions on $X1, X2, X3$, and in-class quizzes $Q1$ to $Q6$.
- ③ ~ 10 new questions.
- All questions have the format: enter a number, vector, matrix or select multiple options.

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.



	$x_{1,1}$	$x_{1,2}$	$x_{1,n}$		
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
...
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
	0	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
	I	am	Groot	we	are

TF IDF Features

Definition

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

$$tf_{ij} = \frac{c_{ij}}{\max_{j'} c_{ij'}}, \quad idf_j = \log \frac{n}{\sum_{i=1}^n \mathbb{1}\{c_{ij} > 0\}}$$

$$x_{ij} = tf_{ij} idf_j$$

- 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



simple not realistic

- 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:

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

Maximum Likelihood Estimation

Definition

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

$$\hat{\mathbb{P}}\{z_t\} = \frac{c_{z_t}}{\sum_{z=1}^m c_z} \Rightarrow \frac{\# \text{Grove}}{\# \text{words}}$$

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

2 = Bigram Model

Definition

1 \Rightarrow U_{ni}

- Bigram models assume Markov property.

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

Handwritten annotations: A green underline is under the entire equation. A green arrow points from the product term back to the first term. A green arrow points from the product term to the second term. A green arrow points from the product term to the third term.

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

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

Handwritten annotations: A green circle is drawn around the conditioning variables $z_{t-1}, z_{t-2}, \dots, z_1$. A green arrow points from this circle to the conditioning variable z_{t-1} in the right-hand side of the equation. A large green arrow points from the right-hand side back to the left-hand side.

Conditional Probability

Definition

- In general, the conditional probability of an event A given another event B is the probability of A and B occurring at the same time divided by the probability of event B .

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{AB\}}{\mathbb{P}\{B\}}$$

- For a sequence of words, the conditional probability of observing z_t given z_{t-1} is observed is the probability of observing both divided by the probability of observing z_{t-1} first.

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

$ch \rightarrow 0.9$
 $most \rightarrow 0.1$

$$P_r\{am|I\}$$

$$= \frac{P_r\{I am\}}{P_r\{I\}}$$

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.

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

$$\frac{\# \text{ I an}}{\# \text{ I}}$$

Unigram MLE Probability

Quiz

$$Pr\{\underline{am} | I\} = \frac{Pr\{I \text{ am}\}}{Pr\{I\}} \rightarrow \text{EoS}$$

- 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"?

uni $\hat{Pr}\{I \text{ am } I\} \stackrel{\text{unigram}}{=} \stackrel{\text{MLE}}{=} \frac{\hat{Pr}\{I\}}{2} \hat{Pr}\{am\} \cdot \hat{Pr}\{I\} = \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{2}{5} = \frac{8}{125}$

bigram $\hat{Pr}\{I \text{ am } Groot\} \stackrel{\text{bigram}}{=} \stackrel{\text{MLE}}{=} \hat{Pr}\{I\} \cdot \hat{Pr}\{am | I\} \cdot \hat{Pr}\{Groot | am\} = \frac{2}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{10}$

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 Groot"?

- A : I am Groot (translation: I don't understand).

Q4

- B : $\frac{2}{25}$

- C : $\frac{4}{25}$

- D : $\frac{4}{125}$

- E : $\frac{8}{125}$

$$\hat{\Pr}\{I \text{ am Groot}\}$$

$$\stackrel{\text{uni}}{=} \hat{\Pr}\{I\} \hat{\Pr}\{\text{am}\} \cdot \hat{\Pr}\{\text{Groot}\}$$

$$\stackrel{\text{MLE}}{=} \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5}$$

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 Groot" given the first word is "I"?
- A : I am Groot (translation: I don't understand).
- B : $\frac{1}{4}$
- C : $\frac{1}{5}$
- D : $\frac{1}{10}$
- E : $\frac{4}{25}$

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.

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} \hat{P}\{1|1\} & \hat{P}\{2|1\} & \hat{P}\{3|1\} \\ \hat{P}\{1|2\} & \hat{P}\{2|2\} & \hat{P}\{3|2\} \\ \hat{P}\{1|3\} & \hat{P}\{2|3\} & \hat{P}\{3|3\} \end{bmatrix} \end{matrix}$$

Handwritten annotations: A blue arrow points to the first column of the matrix. A blue squiggle is under the first row. A blue squiggle is under the second column. A blue squiggle is under the third column. To the right, the word "prev" is written above a blue "2", and "next" is written above a blue "3". A blue arrow points from the "2" to the "3".

- 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	$\frac{8}{9}$	0	0	$\frac{1}{9}$	0
we	0	0	0	0	1
are	0	0	1	0	0

Trigram Model

Definition

$\hat{P}_r \{z_{t-2}, z_{t-1}, z_t\}$

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

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

MLE

$\frac{C_{z_{t-2}, z_{t-1}, z_t}}{C_{z_{t-2}, z_{t-1}}}$
 $\frac{0}{0}$

- 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.

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

regularization

size of vocab.

$C_{z_{t-2}, z_{t-1}, z_t} = 0$

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.

Handwritten notes:
 $\frac{c_{x_{t-1}, y_t}}{c_{y=1}}$
 $\frac{P}{\sum P}$

Smoothing Example

Quiz

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

with Laplace

$$\hat{P}_r\{\text{Groot}\} = \frac{c_{\text{Groot}} + 1}{\# \text{ words} + m} = \frac{3 + 1}{10^{12} + 10^6}$$

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 : I am Groot (translation: I don't understand).

- **B** : $\frac{11}{35}$

- C : $\frac{1}{3}$

- D : $\frac{11}{31}$

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

$$\frac{10 + 1}{30 + 5}$$

Sampling from Discrete Distribution

Discussion

$(0, 1)$
Uniform

- 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.

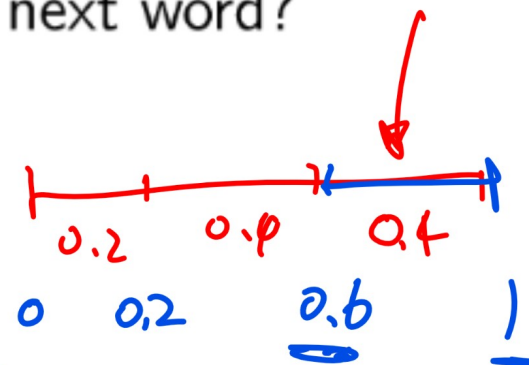
Generating New Words 2

Quiz

Q2

- 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?

	I	am	Groot
I	0.1	0.5	0.4
am	0.2	0.4	0.4
Groot	0.3	0.2	0.5



- A : I, B: am, C: Groot, D: I don't understand

p3 → don't use package