CS540 Introduction to Artificial Intelligence
Lecture 9

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

June 15, 2020

Spam or Ham?

Admin

training data from SMS Spam Collection Dataset.

- Which one of the following messages is a spam?
- A: Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got *a*...

Socrative

Q1

- B: Ok lar... Joking wif u oni...
- C: Free entry in 2*a* wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entr...
- D: U dun say so early hor...*Uc* already then say...
- E: Nah I don't think he goes to usf, he lives around here though

grade → P1, P2 ..

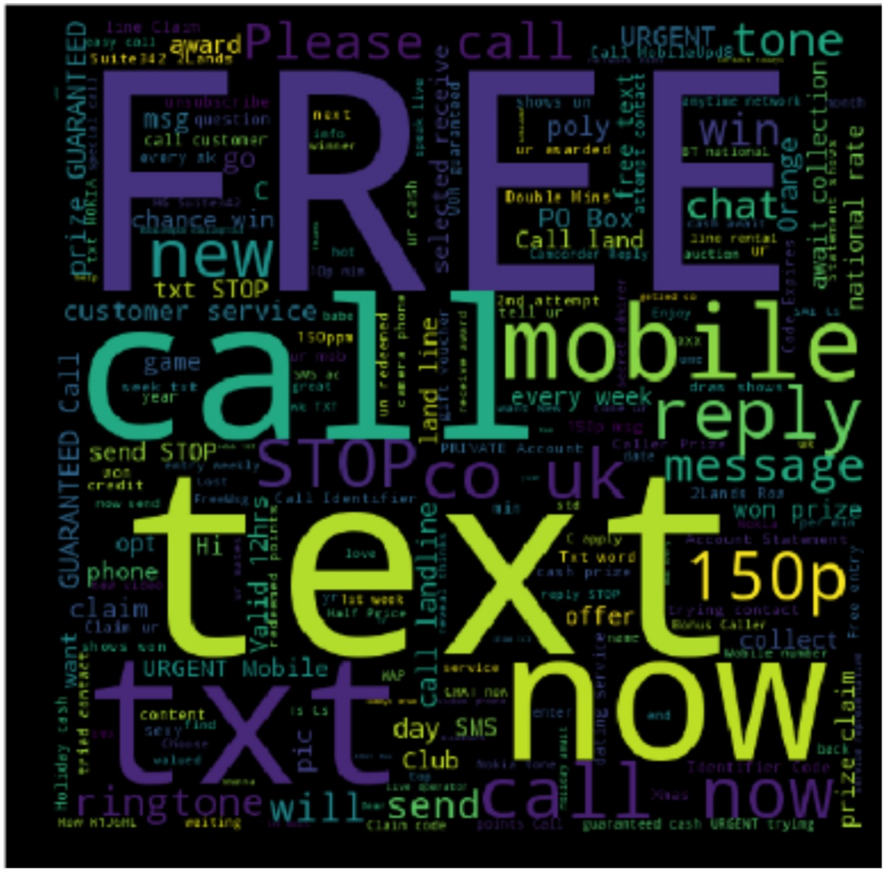
→ code, output → P15, P25 ..

Spam or Ham? Visualization

P3

Admin

Spam



ham

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 $j = 1, 2, ..., m$.
- 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.

$$x_{ij} = \dfrac{c_{ij}}{\displaystyle\sum_{j'=1}^{m} c_{ij'}}$$

Bag of Words Features Example

Definition

instance 1: [I am Groot I am Groot ... $\times 10$
We are Groot]

instance 2: I am Groot

$$X_1 = \begin{array}{c} I \\ am \\ Groot \\ we \\ are \end{array} \begin{pmatrix} 10/33 \\ 10/33 \\ 11/33 \\ 1/33 \\ 1/33 \end{pmatrix} \begin{array}{l} \leftarrow x_{11} \\ \leftarrow x_{12} \\ \leftarrow x_{13} \\ \\ \\ \end{array} \qquad X_2 = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \\ 0 \\ 0 \end{pmatrix}$$

Unigram Model

Definition

- Unigram models assume independence.

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

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

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

- For sequence of words, independence means:

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

Maximum Likelihood Estimation

Definition

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

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

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

Bigram Model

Definition

- Bigram models assume Markov property.

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

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

$$\mathbb{P}\left\{z_t | z_{t-1}, \cancel{z_{t-2}, ..., z_1}\right\} = \mathbb{P}\left\{z_t | z_{t-1}\right\}$$

Bigram Model Estimation

Definition

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

$$\hat{\mathbb{P}}\left\{z_t | z_{t-1}\right\} = \frac{c_{z_{t-1}, z_t} / \text{total}}{c_{z_{t-1}} / \text{total}}$$

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

Unigram MLE Probability 1

Quiz

- Given the training data "I am Iron Man", "I love you 3000", "I love you mom", "Tell my family I love them", 18 words in total. With the unigram model, what is the probability of observing a new sentence "I love"?

$$\Pr\{I\} \cdot \Pr\{love\}$$

$$= \frac{4}{18} \cdot \frac{3}{18}$$

Bigram MLE Probability 1

Quiz

- Given the training data "I am Iron Man", "I love you 3000", "I love you mom", "Tell my family I love them", 18 words in total. With the bigram model, what is the probability of observing $Z_2 =$ "love" given the sentence starts with $Z_1 =$ "I"?

$$\Pr\{\text{love} \mid I\} = \frac{C_{I\ \text{love}}}{C_I} = \frac{3\text{次}}{4\text{次}}$$

Unigram MLE Probability 2

Quiz

Q2

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

$$\frac{2}{5} \frac{2}{5} \frac{2}{5} = \frac{8}{125}$$

Bigram MLE Probability 2

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"?
- A: $\frac{1}{2}$
- B: $\frac{1}{4}$
- C: $\frac{1}{5}$
- D: $\frac{1}{10}$
- E: $\frac{4}{25}$

Q3

for P3

$P_C\{am \mid I\} \cdot \frac{C_{I\ am}}{C_I} = \boxed{\frac{1}{1}}$ or $\frac{1}{2}$

$P_C\{I \mid am\} = \frac{C_{am\ I}}{C_{am}} = \boxed{\frac{1}{2}}$

I after am

Bigram MLE Probability 3

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: $\frac{1}{2}$
- B: $\frac{1}{4}$
- C: $\frac{1}{5}$
- D: $\frac{1}{10}$
- E: $\frac{4}{25}$

Q4 - A

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 $\hat{\mathbb{P}}\left\{j'|j\right\}$. If there are 3 tokens $\left\{1, 2, 3\right\}$, the transition matrix is the following.

$$\begin{bmatrix} \hat{\mathbb{P}}\left\{1|1\right\} & \hat{\mathbb{P}}\left\{2|1\right\} & \hat{\mathbb{P}}\left\{3|1\right\} \\ \hat{\mathbb{P}}\left\{1|2\right\} & \hat{\mathbb{P}}\left\{2|2\right\} & \hat{\mathbb{P}}\left\{3|2\right\} \\ \hat{\mathbb{P}}\left\{1|3\right\} & \hat{\mathbb{P}}\left\{2|3\right\} & \hat{\mathbb{P}}\left\{3|3\right\} \end{bmatrix}$$

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

Trigram Model

Definition

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

$$\hat{\mathbb{P}}\left\{z_t | z_{t-1}, z_{t-2}\right\} = \frac{c_{z_{t-2}, z_{t-1}, z_t}}{c_{z_{t-2}, z_{t-1}}}$$

= 0

- In a document, it is likely that 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.

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

\# of words in vocab.

Laplace Smoothing

Definition

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

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

- 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?

$$\frac{c_z + 1}{\sum c_{z'} + m} = \frac{3+1}{10^{12} + 10^6}$$

Smoothing Example 2

Quiz

- Given a vocabulary of 5, a document with 30 words with $c_{\text{Groot}} = 10$. What is the MLE estimation of $\mathbb{P}\left\{ \text{Groot} \right\}$ with Laplace smoothing?
- A: $\dfrac{1}{2}$
- B: $\dfrac{11}{35}$
- C: $\dfrac{1}{3}$
- D: $\dfrac{11}{31}$
- E: $\dfrac{1}{4}$

Q5

$\dfrac{10+1}{30+5} \rightarrow \dfrac{c_z + 1}{\sum c_{z'} + m}$

words in vocab.

Smoothing Example 3

Quiz

- Given a vocabulary of size 5, a document with 30 words with $c_{\text{I Groot}} = 0, c_{\text{I}} = 9, c_{\text{Groot}} = 10$. What is the MLE estimation of $\mathbb{P}\{\text{Groot} \mid \text{I}\}$ with Laplace smoothing?

  $z_t \mid z_{t-1}$

- A: $\frac{1}{10}$
- B: $\frac{1}{11}$
- C: $\frac{1}{14}$
- D: $\frac{1}{15}$
- E: 0

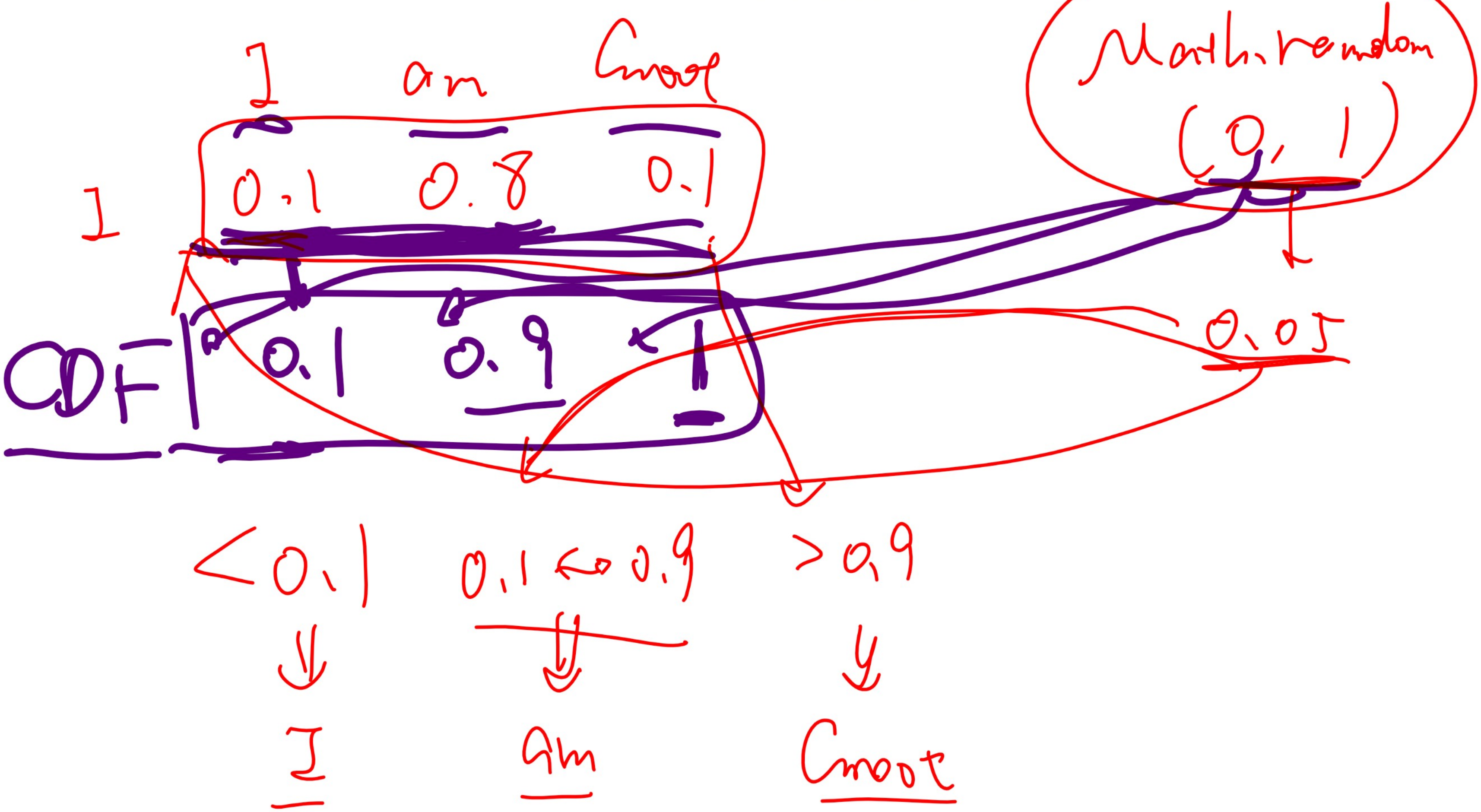
not overfit.

Q6

$$\frac{c_{\text{I Groot}} + 1}{c_{\text{I}} + m} = \frac{0+1}{9+5} = \frac{1}{14}$$

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?

next

$$
\text{from}\quad
\begin{array}{c|c}
 & \begin{matrix} I & am & Groot \end{matrix} \\
\begin{matrix} I \\ am \\ Groot \end{matrix} &
\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}
\end{array}
$$

CDF

0.1 0.6 1

am

Generating New Words 2

Quiz

back @ 6:40

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

$$\begin{array}{cc} & \begin{array}{ccc} \text{I} & \text{am} & \text{Groot} \end{array} \\ \begin{array}{c} \text{I} \\ \text{am} \\ \text{Groot} \end{array} & \begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix} \end{array}$$

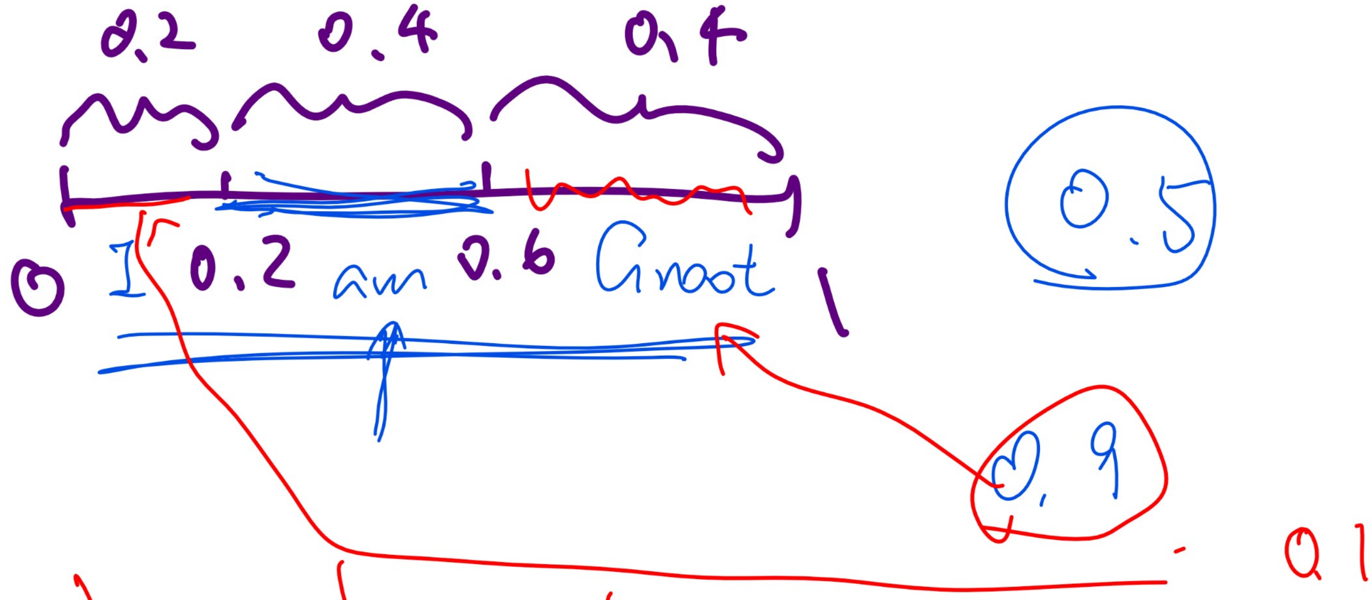
Q7

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

0.2 0.6 1

I I
I am
I Groot
am I
am am

I	am	Groot
1/3	1/3	1/3
0	0	1
⋮	⋮	⋮



$$\frac{z_{t-1}}{z_{t-3}} \qquad \frac{z_{t-2}}{z_{t-1}} \qquad \frac{z_{t-3}}{z_{t-2}}$$