CS540 Introduction to Artificial Intelligence

Lecture 7

Young Wu

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

June 20, 2022

Lecture Feedback, Additional Examples, Solutions

Admin

Q1 → anything A → E

- Thank you for the feedback on Socrative and Homework submission.
- They are addressed on Piazza (will be updated weekly on Tuesday).
- Formula and notation explanation added (will updated weekly after lecture).
- Review sessions on Wednesdays before the exams to go over past exam questions.

Midterm Details

Admin

- More midterm-related details next Monday:

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:

1. ~ 10 questions from $M2$ to $M7$ (same question different randomization).
2. ~ 10 questions from relavent questions on $X1, X2$, 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.

K Nearest Neighbor

Description

- Given a new instance, find the K instances in the training set that are the closest.
- Predict the label of the new instance by the majority of the labels of the K instances.

Distance Function

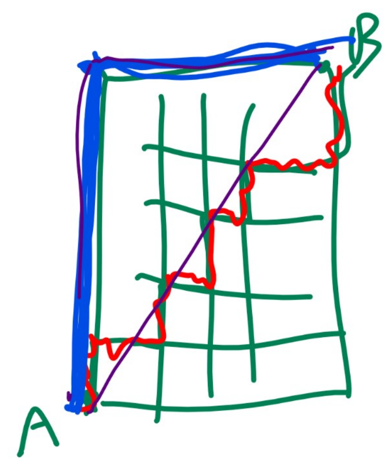
Definition

- Many distance functions can be used in place of the Euclidean distance.

$$\rho\left(x, x'\right) = \left\|x - x'\right\|_2 = \sqrt{\sum_{j=1}^{m}\left(x_j - x_j'\right)^2}$$

- An example is Manhattan distance.

$$\rho\left(x, x'\right) = \sum_{j=1}^{m}\left|x_j - x_j'\right|$$



Manhattan Distance Diagram

Definition

1 Nearest Neighbor

Quiz

x' $\Rightarrow y' = 0$

- Find the 1 Nearest Neighbor label for $\begin{bmatrix} 3 \\ 6 \end{bmatrix}$ using Manhattan distance.

x_1	1	1	3	5	2
x_2	1	7	3	4	5
y	0	1	1	0	0

training set

nearest neighbor

dist 7 3 3 4 2

3 Nearest Neighbor

Quiz

x' $y' = ?$

- Find the 3 Nearest Neighbor label for $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$ using Manhattan distance.

dist	4	6	0	3	3
x_1	1	1	3	5	2
x_2	1	7	3	4	5
y	0	1	1	0	0

- A : 0, B : Not sure, I guess it is 0.
- C : 1, D : Not sure, I guess it is 1.

K Fold Cross Validation

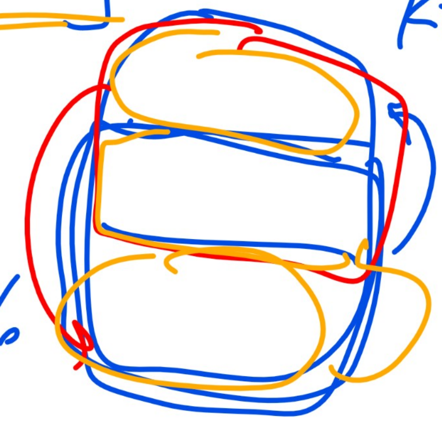
Discussion

measure accuracy on training set

↓

100%

K=3



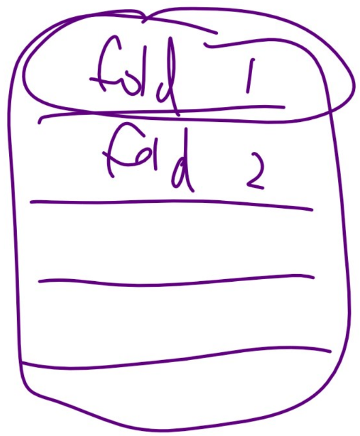
- Partition the training set into K groups.
- Pick one group as the validation set.
- Train the model on the remaining training set.
- Repeat the process for each of the K groups.
- Compare accuracy (or cost) for models with different hyperparameters and select the best one.

layers

units in layers.

5 Fold Cross Validation Example

Discussion



LOOCV

Leave One Out Cross Validation

Discussion

- If $K = n$, each time exactly one training instance is left out as the validation set. This special case is called Leave One Out Cross Validation (LOOCV).

Cross Validation

Quiz

- Given the following training data. What is the 2 fold cross-validation accuracy if 1 nearest neighbor classifier with Manhattan distance is used? The first fold is the first five data points.

x	1	1	2	2	3	3	4	4	5	5
y	1	2	3	3	2	2	3	3	2	1
$\hat{y}$	2	2	2	2	2	2	2	2	2	2

40% → acc

Cross Validation 2

Quiz

Q3

- Given the following training data. What is the 10 fold cross-validation (LOOCV) accuracy if 1 nearest neighbor classifier with Manhattan distance is used?

x	1	1	2	2	3	3	4	4	5	5
y	1	2	3	3	2	2	3	3	2	1

- A : 20 percent, B: 40, C: 60, D: 80, E: I do not understand.

2 1 3 3 2 2 3 3 1 2

1 1 1

1 2

$\frac{1}{2} \to 1$ $\frac{1}{2} \to 2$

60%

tiebreaking

default class

1

Tokenization

Motivation

- When processing language, documents (called corpus) need to be turned into a sequence of tokens.

1. Split the string by space and punctuations.
2. Remove stopwords such as "the", "of", "a", "with" ...
3. Lower case all characters.
4. Stemming or lemmatization words: make "looks", "looked", "looking" to "look".

Vocabulary

Motivation

- Word token is an occurrence of a word.
- Word type is a unique token as a dictionary entry.
- Vocabulary is the set of word types.
- Characters can be used in place of words as tokens. In this case, the types are "a", "b", ..., "z", " ", and vocabulary is the alphabet.

$\boxed{\text{P3}}$

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

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

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

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.

	I	am	Groot	We	Are
Instance 1: X_1 →	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
X_2 →	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
	...	...	...	...	...
X_9 →	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
X_{10} →	0	0	$\boxed{\frac{1}{3}}$	$\frac{1}{3}$	$\frac{1}{3}$

← 4 types

TF IDF Features

Definition

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

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

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

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

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 independent 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 a 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.

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

- 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.

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

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

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

Markov Chain Demo

Motivation

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.

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

- 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.

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

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.

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

$z_{t-1}\ z_t$

z_t

MLE

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"?

$Pr\{I\} = \frac{2}{5}$

$Pr\{am\} = \frac{2}{5}$

$Pr\{Groot\} = \frac{1}{5}$

$\underbrace{\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{2}{5}}_{}$

$\frac{8}{125}$

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"?

$$\Pr\{ am \mid I \} = \text{prob, then prev is I next is am}$$

$$= \frac{1}{2}$$

$$\Pr\{ I \mid am \} = \frac{1}{2}$$

$$\frac{1}{4}$$

$\Pr\{am \mid I\} \cdot \Pr\{I \mid am\}$

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

token EOS

Q4

$\Pr\{I\} = \Pr\{am\} = \frac{2}{5}$

$\Pr\{Groot\} = \frac{1}{5}$

$\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5}$

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 : $\frac{4}{25}$

Q5

I is followed by am
am following I

$$\Pr\{am \mid I\} = \frac{1}{2}$$

$\frac{1}{5}$ $\frac{1}{2}$

$$\Pr\{Groot \mid am\} = \frac{\#\, am\ Groot}{\#\, am} = \frac{1}{2}$$

bigram

$$\Pr\{I\ am\ Groot \mid I\}$$

$$= \Pr\{am \mid I\} \cdot \Pr\{Groot \mid am\}$$

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 $\{1, 2, 3\}$, the transition matrix is the following.

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

sum up to 1

- 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

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

$$\underline{\hat{\mathbb{P}}\left\{z_t | z_{t-1}, z_{t-2}\right\}} = \frac{c_{z_{t-2}, z_{t-1}, z_t}}{c_{z_{t-2}, z_{t-1}}} \qquad \frac{0}{0}$$

Great Great Grove

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

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

Laplace Smoothing

Definition

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

bigram

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

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

unigram

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

regularization of DT

Smoothing Example

Quiz

← # unique words + types

- 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?

$$\hat{P}\{\text{Groot}\} \; \frac{3}{10^{12}}$$

$$\hat{P}\{\text{Groot}\} = \frac{3+1}{10^{12}+10^{6}}$$

Smoothing Example 2

Quiz

Qb

- 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).
- $\boxed{B : \frac{11}{35}}$
- C : $\frac{1}{3}$
- D : $\frac{11}{31}$
- E : $\frac{1}{4}$

$$\hat{\mathbb{P}}\{\text{Groot}\} = \frac{\overset{10}{C_{\text{Groot}}} + 1}{\underset{30}{\#\text{token}} + \underset{5}{m}}$$

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?

Q7

- A : I am Groot (translation: I don't understand).
- B : $\frac{1}{10}$
- C : $\frac{1}{11}$
- D : $\frac{1}{15}$
- E : 0

$$\frac{1}{14}$$

$$\frac{\overset{0}{C}_{I\ \text{Groot}} + 1}{\underset{9}{\# I} + \underset{5}{m}}$$

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, $z_1, z_2, ..., z_{N-1}$, the distribution of next word is approximated by $p_x = \hat{\mathbb{P}}\left\{z_N = x | z_{N-1}, z_{N-2}, ..., z_1\right\}$. This process then can be repeated for on $z_2, z_3, ..., z_{N-1}, z_N$ and so on.

Unif $(0,1]$

CDF Inversion Method Diagram

Discussion

I am Groot

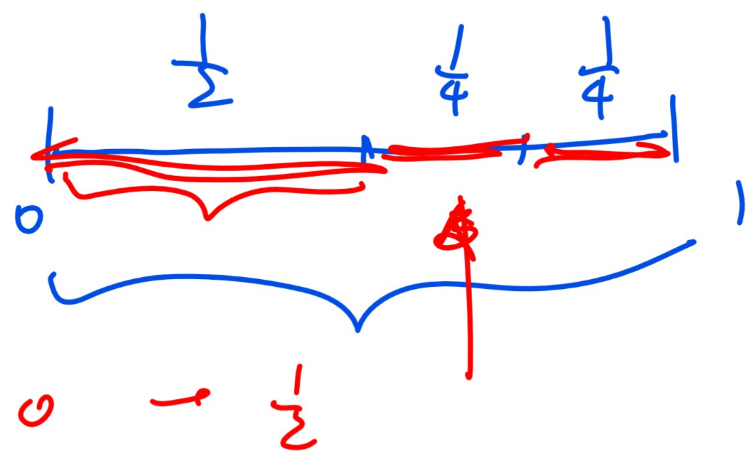
pr { $\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$ }

$u \sim \text{Unif}[0, 1]$

if $u < \frac{1}{2} \rightarrow$ I

$\frac{1}{2} < u < \frac{3}{4} \rightarrow$ am

$\frac{3}{4} < u < 1 \rightarrow$ Groot



CDF → $\frac{1}{2}$ $\frac{3}{4}$ 1

1 2 3

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?

am

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

0.5

0.1 0.5 0.4

am

0.1 0.6 1

Generating New Words 2

Quiz

Q8

- 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?

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

0.75

0.2 0.6 1 → CDF

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