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

Midterm Details

Admin

Midterm Coverage

Admin

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

3 Nearest Neighbor

Quiz

K Fold Cross Validation

Discussion

5 Fold Cross Validation Example

Discussion

Leave One Out Cross Validation

Discussion

Cross Validation

Quiz

Cross Validation 2

Quiz

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.

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

Motivation

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:

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

- For a sequence of words, independence means:

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

Maximum Likelihood Estimation

Definition

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

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

- 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.

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

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

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

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.

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

- 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.

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

Bigram Model Estimation

Definition

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

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

Unigram MLE Probability

Quiz

Bigram MLE Probability

Quiz

Unigram MLE Probability

Quiz

Bigram MLE Probability

Quiz

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 $\hat{\mathbb{P}}\left\{j' \middle| j\right\}$. 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.

Estimating Transition Matrix

Definition

Trigram Model

Definition

- The same formula can be applied to trigram: sequences of three tokens.
$$\hat{\mathbb{P}}\left\{z_t \middle| z_{t-1}, z_{t-2}\right\} = \frac{c_{z_{t-2},z_{t-1},z_t}}{c_{z_{t-2},z_{t-1}}}$$
- 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.
$$\hat{\mathbb{P}}\left\{z_t \middle| z_{t-1}, z_{t-2}\right\} = \frac{c_{z_{t-2},z_{t-1},z_t} + 1}{c_{z_{t-2},z_{t-1}} + m}$$

Laplace Smoothing

Definition

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

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

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

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

Smoothing Example

Quiz

Smoothing Example 2

Quiz

Smoothing Example 3

Quiz

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 $p_x = \hat{\mathbb{P}} \left\{ z_N = x \middle| z_{N-1}, z_{N-2}, ..., z_1 \right\}$. This process then can be repeated for on $z_2, z_3, ..., z_{N-1}, z_N$ and so on.

CDF Inversion Method Diagram

Discussion

Generating New Words 1

Quiz

Generating New Words 2

Quiz