Statistics and Natural Language Processing

Daifeng Wang

daifeng.wang@wisc.edu

University of Wisconsin, Madison

Based on slides from Xiaojin Zhu and Yingyu Liang (http://pages.cs.wisc.edu/~jerryzhu/cs540.html), modified by Daifeng Wang

Outline

- Probability
  - Independence
  - Conditional independence
  - Expectation
- Natural Language Processing
  - Preprocessing
  - Statistics
  - Language models

Independence

- Two events A, B are independent, if (the following are equivalent)
  - $P(A, B) = P(A) * P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- For a 4-sided die, let
  - A=outcome is small
  - B=outcome is even
  - Are A and B independent?
- How about a 6-sided die?

Independence

- Independence is a domain knowledge
- If A, B are independent, the joint probability table between A, B is simple:
  - it has k^2 cells, but only 2k-2 parameters. This is good news – more on this later…
- Example: P(burglary)=0.001, P(earthquake)=0.002. Let's say they are independent. The full joint probability table=?

Conditional independence

- Random variables can be dependent, but conditionally independent
- Your house has an alarm
  - Neighbor John will call when he hears the alarm
  - Neighbor Mary will call when she hears the alarm
  - Assume John and Mary don't talk to each other
- JohnCall independent of MaryCall?
  - No – If John called, likely the alarm went off, which increases the probability of Mary calling
  - $P(MaryCall \mid JohnCall) \neq P(MaryCall)$

Conditional independence

- If we know the status of the alarm, JohnCall won't affect Mary at all

  P(MaryCall | Alarm, JohnCall) = P(MaryCall | Alarm)

- We say JohnCall and MaryCall are conditionally independent, given Alarm
- In general A, B are conditionally independent given C
  - if P(A | B, C) = P(A | C), or
  - P(B | A, C) = P(B | C), or
  - P(A, B | C) = P(A | C) * P(B | C)

Independence example #1

x, y	$P(X=x, Y=y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

x	$P(X=x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y=y)$
on-time	0.45
late	0.55

Are X and Y independent here? NO.

Independence example #2

x, y	$P(X=x, Y=y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Delta	0.03
rain, fly-Delta	0.05
snow, fly-Delta	0.02

x	$P(X=x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y=y)$
fly-United	0.9
fly-Delta	0.1

Are X and Y independent here? YES.

Expected values

- The *expected value* of a random variable that takes on numerical values is defined as:

$$\mathbf{E}[X] = \sum_x x P(x)$$

This is the same thing as the *mean*

- We can also talk about the expected value of a function of a random variable

$$\mathbf{E}[g(X)] = \sum_x g(x) P(x)$$

Expected value examples

- Shoesize

$$\begin{aligned} & \mathbf{E}[Shoesize] \\ = \quad & 5 \times P(Shoesize = 5) + \cdots + 14 \times P(Shoesize = 14) \end{aligned}$$

- Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

What is the expectation of the gain?

Expected value examples

- Shoesize

$$
\begin{aligned}
& E[Shoesize] \\
= \quad & 5 \times P(Shoesize = 5) + \cdots + 14 \times P(Shoesize = 14)
\end{aligned}
$$

- Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

$$
\begin{aligned}
& \mathbf{E}[gain(Lottery)] \\
= \quad & gain(winning)P(winning) + gain(losing)P(losing) \\
= \quad & (\$100 - \$1) \times 0.001 - \$1 \times 0.999 \\
= \quad & -\$0.9
\end{aligned}
$$

Summary

- Axioms of probability and related properties
- Joint/marginal/conditional probabilities
- Bayes' rule for reasoning
- Independence and conditional independence
- Expectation

Natural language Processing (NLP)

- The processing of the human languages by computers

- One of the oldest AI tasks
- One of the most important AI tasks
- One of the hottest AI tasks nowadays

Difficulty

- Difficulty 1: ambiguous, typically no formal description
- Example: "*We saw her duck.*"

How many different meanings?

Difficulty

- Difficulty 1: ambiguous, typically no formal description
- Example: "*We saw her duck.*"
- 1. We looked at a duck that belonged to her.
- 2. We looked at her quickly squat down to avoid something.
- 3. We use a saw to cut her duck.

Difficulty

- Difficulty 2: computers do not have human concepts

- Example: "*She like little animals. For example, yesterday we saw her duck.*"
- 1. We looked at a duck that belonged to her.
- 2. We looked at her quickly squat down to avoid something.
- 3. We use a saw to cut her duck.

Preprocess

Zipf's Law

WORDS

Preprocess

- Corpus: often a set of text documents
- Tokenization or text normalization: turn corpus into sequence(s) of tokens

1. Remove unwanted stuff: HTML tags, encoding tags
2. Determine word boundaries: usually white space and punctuations
   - Sometimes can be tricky, like Ph.D.

Preprocess

- Tokenization or text normalization: turn data into sequence(s) of tokens

1. Remove unwanted stuff: HTML tags, encoding tags
2. Determine word boundaries: usually white space and punctuations
   - Sometimes can be tricky, like Ph.D.
3. Remove stopwords: the, of, a, with, …

Preprocess

- Tokenization or text normalization: turn data into sequence(s) of tokens

1. Remove unwanted stuff: HTML tags, encoding tags
2. Determine word boundaries: usually white space and punctuations
   - Sometimes can be tricky, like Ph.D.
3. Remove stopwords: the, of, a, with, …
4. Case folding: lower-case all characters.
   - Sometimes can be tricky, like US and us
5. Stemming/Lemmatization (optional): looks, looked, looking → look

Vocabulary

Given the preprocessed text

- Word token: occurrences of a word
- Word type: unique word as a dictionary entry (i.e., unique tokens)

- Vocabulary: the set of word types
  - Often 10k to 1 million on different corpora
  - Often remove too rare words

Zipf's Law

- Word count f, word rank r
- Zipf's law: $f * r \approx \text{constant}$

Word	Count f	rank r	fr
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
two	104	100	10400
turned	51	200	10200
comes	16	500	8000
family	8	1000	8000
brushed	4	2000	8000
Could	2	4000	8000
Applausive	1	8000	8000

Zipf's law on the corpus *Tom Sawyer*

Bag-of-Words

tf-idf

TEXT: BAG-OF-WORDS REPRESENTATION

Bag-of-Words

How to represent a piece of text (sentence/document) as numbers?

- Let m denote the size of the vocabulary
- Given a document d, let $c(w, d)$ denote the #occurrence of w in d
- Bag-of-Words representation of the document

$$v_d = [c(w_1, d), c(w_2, d), \dots, c(w_m, d)]/Z_d$$

- Often $Z_d = \sum_w c(w, d)$

Example

- Preprocessed text: *this is a good sentence this is another good sentence*

- BoW representation:
  $[c('a', d)/Z_d, c('is', d)/Z_d, \ldots, c('example', d)/Z_d]$
- What is Z_d?
- What is $c('a', d)/Z_d$?
- What is $c('example', d)/Z_d$?

tf-idf

- tf: normalized term frequency

$$tf_w = \frac{c(w,d)}{\max_v c(v,d)}$$

- idf: inverse document frequency

$$idf_w = \log \frac{\text{total \#doucments}}{\text{\#documents containing } w}$$

- tf-idf: $tf\text{-}idf_w = tf_w * idf_w$
- Representation of the document

$$v_d = [tf{-}idf_{w_1}, tf{-}idf_{w_2}, \ldots, tf{-}idf_{w_m}]$$

Cosine Similarity

How to measure similarities between pieces of text?

- Given the document vectors, can use any similarity notion on vectors
- Commonly used in NLP: cosine of the angle between the two vectors

$$sim(x, y) = \frac{x^\top y}{\sqrt{x^\top x}\sqrt{y^\top y}}$$

Statistical language model

N-gram

Smoothing

TEXT: STATISTICAL LANGUAGE MODEL

Probabilistic view

- Use probabilistic distribution to model the language
- Dates back to Shannon (information theory; bits in the message)

Statistical language model

- Language model: probability distribution over sequences of tokens
- Typically, tokens are words, and distribution is discrete

- Tokens can also be characters or even bytes

- Sentence: "*the quick brown fox jumps over the lazy dog*"

Tokens: x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9

Statistical language model

- For simplification, consider fixed length sequence of tokens (sentence)

$$(x_1, x_2, x_3, \dots, x_{\tau-1}, x_\tau)$$

- Probabilistic model:

$$\mathrm{P}\,[x_1, x_2, x_3, \dots, x_{\tau-1}, x_\tau]$$

Unigram model

- Unigram model: define the probability of the sequence as the product of the probabilities of the tokens in the sequence

$$\mathrm{P}[x_1, x_2, \ldots, x_\tau] = \prod_{t=1}^{\tau} \mathrm{P}[x_t]$$

- Independence!

A simple unigram example

- Sentence: "*the dog ran away*"

  $$\hat{\mathrm{P}}[the\ dog\ ran\ away] = \hat{\mathrm{P}}[the]\,\hat{\mathrm{P}}[dog]\,\hat{\mathrm{P}}[ran]\,\hat{\mathrm{P}}[away]$$

- How to estimate $\hat{\mathrm{P}}[the]$ on the training corpus?

A simple unigram example

- Sentence: "*the dog ran away*"

$$\hat{\mathrm{P}}[the\ dog\ ran\ away] = \hat{\mathrm{P}}[the]\,\hat{\mathrm{P}}[dog]\,\hat{\mathrm{P}}[ran]\,\hat{\mathrm{P}}[away]$$

- How to estimate $\hat{\mathrm{P}}[the]$ on the training corpus?

Word	Count f
the	3332
and	2972
a	1775
he	877
but	410
be	294
there	222
one	172

n-gram model

- n-gram: sequence of n tokens

- n-gram model: define the conditional probability of the n-th token given the preceding $n-1$ tokens

$$\mathrm{P}[x_1, x_2, \dots, x_\tau] = \mathrm{P}[x_1, \dots, x_{n-1}] \prod_{t=n}^{\tau} \mathrm{P}[x_t | x_{t-n+1}, \dots, x_{t-1}]$$

n-gram model

- n-gram: sequence of n tokens

- n-gram model: define the conditional probability of the n-th token given the preceding $n-1$ tokens

$$\mathrm{P}[x_1, x_2, \dots, x_\tau] = \mathrm{P}[x_1, \dots, x_{n-1}] \prod_{t=n}^{\tau} \mathrm{P}[x_t | x_{t-n+1}, \dots, x_{t-1}]$$

Markovian assumptions

Typical n-gram model

- $n = 1$: unigram
- $n = 2$: bigram
- $n = 3$: trigram

Training n-gram model

- Straightforward counting: counting the co-occurrence of the grams

For all grams $(x_{t-n+1}, \dots, x_{t-1}, x_t)$

1. count and estimate $\widehat{\mathrm{P}}[x_{t-n+1}, \dots, x_{t-1}, x_t]$
2. count and estimate $\widehat{\mathrm{P}}[x_{t-n+1}, \dots, x_{t-1}]$
3. compute

$$\hat{\mathrm{P}}[x_t | x_{t-n+1}, \dots, x_{t-1}] = \frac{\hat{\mathrm{P}}[x_{t-n+1}, \dots, x_{t-1}, x_t]}{\hat{\mathrm{P}}[x_{t-n+1}, \dots, x_{t-1}]}$$

A simple trigram example

- Sentence: "*the dog ran away*"

$$\hat{\mathrm{P}}[the\ dog\ ran\ away] = \hat{\mathrm{P}}[the\ dog\ ran]\ \hat{\mathrm{P}}[away|dog\ ran]$$

$$\hat{\mathrm{P}}[the\ dog\ ran\ away] = \hat{\mathrm{P}}[the\ dog\ ran]\frac{\hat{\mathrm{P}}[dog\ ran\ away]}{\hat{\mathrm{P}}[dog\ ran]}$$

Drawback

- Sparsity issue: $\widehat{\mathrm{P}}[\ldots]$ most likely to be 0

- Bad case: *"dog ran away"* never appear in the training corpus, so $\widehat{\mathrm{P}}[dog\ ran\ away] = 0$
- Even worse: *"dog ran"* never appear in the training corpus, so $\widehat{\mathrm{P}}[dog\ ran] = 0$

Rectify: smoothing

- Basic method: adding non-zero probability mass to zero entries

- Example: Laplace smoothing that adds one count to all n-grams

$$\text{pseudocount}[dog] = \text{actualcount}[dog] + 1$$

Rectify: smoothing

- Basic method: adding non-zero probability mass to zero entries
- Example: Laplace smoothing that adds one count to all n-grams

$$\text{pseudocount}[dog] = \text{actualcount}[dog] + 1$$

$$\hat{\mathrm{P}}[dog] = \frac{\text{pseudocount}[dog]}{\text{pseudo length of the corpus}} = \frac{\text{pseudocount}[dog]}{\text{actual length of the corpus } + |V|}$$

Rectify: smoothing

- Basic method: adding non-zero probability mass to zero entries

- Example: Laplace smoothing that adds one count to all n-grams

pseudocount$[dog\ ran\ away] = \text{actualcount}$
$[dog\ ran\ away] + 1$

pseudocount$[dog\ ran] = ?$

Rectify: smoothing

- Basic method: adding non-zero probability mass to zero entries

- Example: Laplace smoothing that adds one count to all n-grams

pseudocount$[dog\ ran\ away] =$ actualcount $[dog\ ran\ away] + 1$

pseudocount$[dog\ ran] \approx$ actualcount$[dog\ ran] + |V|$

$$\hat{\mathrm{P}}[away|dog\ ran] = \frac{\text{pseudocount}[dog\ ran\ away]}{\text{pseudocount}\ [dog\ ran]}$$

since #bigrams ≈#trigrams on the corpus

Example

- Preprocessed text: *this is a good sentence this is another good sentence*

- How many unigrams?
- How many bigrams?
- Estimate $\widehat{\mathrm{P}}[is|this]$ without using Laplace smoothing
- Estimate $\widehat{\mathrm{P}}[is|this]$ using Laplace smoothing (|V| = 10000)

Rectify: smoothing

- Basic method: adding non-zero probability mass to zero entries
  - Example: Laplace smoothing

- Back-off methods: restore to lower order statistics
  - Example: if $\widehat{\mathrm{P}}[away|dog\ ran]$ does not work, use $\widehat{\mathrm{P}}[away|ran]$ as replacement

- Mixture methods: use a linear combination of $\widehat{\mathrm{P}}[away|ran]$ and $\widehat{\mathrm{P}}[away|dog\ ran]$

Another drawback

- High dimesion: # of grams too large

- Vocabulary size: about 10k=2^14
- #trigram: about 2^42

Rectify: clustering

- Class-based language models: cluster tokens into classes; replace each token with its class
- Significantly reduces the vocabulary size; also address sparsity issue

- Combinations of smoothing and clustering are also possible