Lecture 5: Probabilities

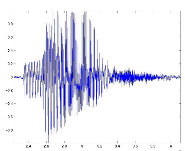
PCL II, CL, UZH
March 23, 2016

Universität Zürich[UZH]

Contents

Probabilities
Language Model

Universität Zürich[UZH]

- Language Model is the distribution of sequence of words
- Applications
  - Handwriting recognition
  - Tagging
    *p(*Adj Noun Verb Det Verb | "fruit flies like a banana"*)*
  - speech recognition
    p("wreck a nice beach" |  *)*
  - Machine Translation
    p("All your base are belong to us." | "君達の基地は、全てCATSがいただいた。"*)*
  - Language Identification
    - Based on “Character-N-Grams” [Dunning, 1994]

Probabilities
Language Model

How do we get the probabilities?

Contents

Probability Theory Background

- *Probability theory* to predict how likely is that something will happen
- *Experiment (or trial)* observations are made

  e.g. Tossing a coin

- An experiment is a collection of *basic outcomes*
- *Sample space Ω* all possible outcomes
  - Discrete: countable number of basic outcomes
  - Continuous: uncountable number of basic outcomes
- *Event A* a subset of Ω

Probability Theory Background

Probability distribution: probability 1 is distributed throughout the sample space Ω

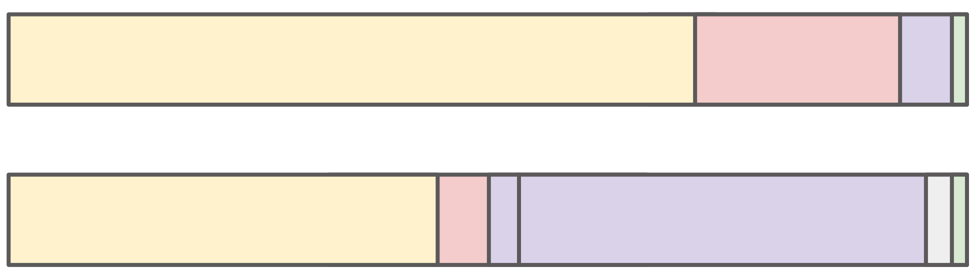
- Each basic outcome has a probability
  e.g. P(heads) = 0.5; P(tails) = 0.5;

- Restrictions:
  - all *p*'s have to be between 0 and 1
  - the sum of all *p*'s over all possible outcomes of the same event has to be 1

Probability Theory Background

- Uniform distribution: all parameters equal *1/N*

- Skewed distribution and others



Probability Theory Background

Universität Zürich[UZH]

- is 0.003 "probable"?

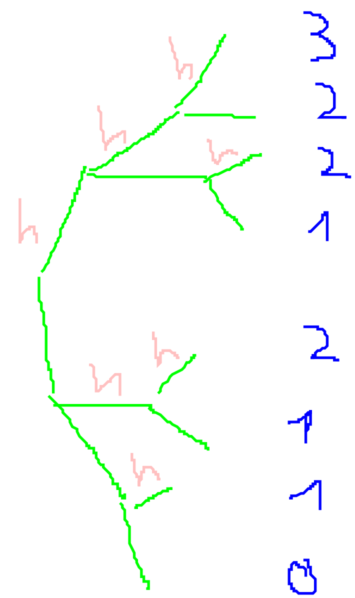
Probability Theory Background

- is 0.003 "probable"?
    - for 3 possible outcomes?
    - for 1000 possible outcomes
        - where all other probabilities are 0.000997?
        - where another outcome has a probability of 0.1?

Probability Theory Background
Example

A fair coin is tossed ***3 times***.

What is the chance of ***2 heads***?



Experiment = ?

Ω = ?

a basic outcome = ?

Probability distribution = ?

A = ?

P(A) = ?

Probability Theory Background
(In)dependent Events

- Two events A, B are *independent* if
  - $P(A \cap B) = P(A)P(B)$ or
  - $P(A) = P(A|B)$ when $P(B) > 0$

  e.g. landing on heads after tossing a coin
- Otherwise A, B are *dependent*
  - $P(A|B) = P(A \cap B) / P(B)$ when $P(B) > 0$

    The multiplication rule:
  - $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$ also when $P(B) = 0$

- Words in a sentence, text, … are *dependent*.

Contents

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

der Wagen	10
der schnelle	2
der Mensch	3
schnelle Wagen	5

Q: What is the probability that *Wagen* follows *der*?

$p(\text{Wagen}|\text{der}) = ?$

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

- What probability estimates should we use for estimating the next word?
  - Absolute frequency: $f(\mathrm{x})$
  - Relative frequency: $f(\mathrm{x})/\mathrm{N}$

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

- What probability estimates should we use for estimating the next word?
  - Absolute frequency: $f(\mathrm{x})$
  - **Relative frequency: $f(\mathrm{x})/\mathrm{N}$**

- Maximum Likelihood Estimate (MLE)

$$P_{\mathrm{MLE}}(\mathrm{w}_1 \ldots \mathrm{w}_n) = \mathrm{C}(\mathrm{w}_1 \ldots \mathrm{w}_n)/\mathrm{N}$$

$$P_{\mathrm{MLE}}(\mathrm{w}_n | \mathrm{w}_1 \ldots \mathrm{w}_{n-1}) = \mathrm{C}(\mathrm{w}_1 \ldots \mathrm{w}_n)/\mathrm{C}(\mathrm{w}_1 \ldots \mathrm{w}_{n-1})$$

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

der Wagen	10
der schnelle	2
der Mensch	3
schnelle Wagen	5

Q: What is the probability that *Wagen* follows *der*?

$p(\text{Wagen}|\text{der}) = ?$

$P(\text{Wagen}|\text{der}) = f(\text{der, Wagen})/f(\text{der})$ =10/15

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

Conditional Probabilities in Python:
Nested dictionaries

```
f = {"der": {"Wagen": 10, "schnelle": 2, "Mensch": 3},
  "schnelle": {"Wagen": 5, "Mensch": 2}}

total = sum(f["der"].values())
f["der"]["Wagen"]/float(total)
```

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

Conditional Probabilities in Python:
nltk.ConditionalFreqDist

```
import nltk

b_a = [("der", "Wagen"), ("der", "Wagen"), ("der", "Wagen"),
("der", "Wagen"), ("der", "Wagen"), ("der", "Wagen"),
("der", "Wagen"), ("der", "Wagen"), ("der", "Wagen"),
("der", "Wagen"), ("der", "schnelle"), ("der", "schnelle"),
("der", "Mensch"), ("der", "Mensch"), ("der", "Mensch"),
("schnelle", "Wagen"), ("schnelle", "Wagen"), ("schnelle", "Wagen"),
("schnelle", "Wagen"), ("schnelle", "Wagen"), ("schnelle", "Mensch"),
("schnelle", "Mensch")]

cfd = nltk.ConditionalFreqDist(b_a)
```

Conditional Probabilities
Maximum Likelihood Estimate (MLE)

Conditional Probabilities in Python:
nltk.ConditionalFreqDist

```
print cfd["der"]
print cfd["schnelle"]
print cfd["der"]["Wagen"]
print sum(cfd["der"].values())
print cfd["der"]["Wagen"]/float(sum(cfd["der"].values()))

cfd.tabulate()
cfd.plot()
```

Conditional Probabilities

Probability of a text?

How can we generalize it to several events?

Contents

Chain Rule

- Conditional Probability
  $P(A|B) = P(A \cap B) / P(B)$ when $P(B) > 0$
- The multiplication rule:

  $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$ also when $P(B) = 0$

Chain Rule

- Conditional Probability
  $P(A|B) = P(A \cap B) / P(B)$ when $P(B) > 0$
- The multiplication rule:

  $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$ also when $P(B) = 0$

<u>Chain Rule</u> generalization of multiplication rule:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\ldots P(A_n|A_1 \cap \ldots \cap A_{n-1})$$

Chain Rule

Chain rule for Texts:

$$P(w_1, \ldots, w_l) = \prod_{n=1}^{l} P(w_n | w_1, \ldots, w_{n-1})$$

Chain Rule

Chain rule for Texts:

$$P(w_1, \ldots, w_l) = \prod_{n=1}^{l} P(w_n | w_1, \ldots, w_{n-1})$$

Problem: Data sparseness $\longrightarrow$ Zero probabilities

Solution: only N-Grams - Markov assumption

Contents

Markov Model

Markov Assumption

- A word w_n depends only on k previous words w_{n-k}, ..., w_{n-1}, where $0 <= k < n$

Markov Model

Markov Assumption

- A word w_n depends only on k previous words w_{n-k}, …, w_{n-1}, where $0 <= k < n$

Markov Models

Unigram (k = 0) Normal word frequency

Bigram (k = 1) w_n depends on w_{n-1}

Trigram (k = 2) w_n depends on w_{n-2}, w_{n-1}

Markov Model
Example: Unigram

Universität Zürich[UZH]

Was sich seit Monaten angedeutet hatte , was nach dem letzten , ernüchternden Vorbereitungstreffen der Unterhändler **in** Barcelona keine Überraschung mehr ist , es reift nun zur sicheren Prognose : Das **in** diversen Konferenzen beschworene Ziel , **in** diesem Jahr noch den soliden Rahmen für ein globales Klimaschutz-Abkommen **in** der Nachfolge des Kyoto-Protokolls zu zimmern , ist illusorisch geworden . Daran ist nicht mehr zu zweifeln , wenn nicht einmal der dänische Ministerpräsident Lars Løkke Rasmussen als Repräsentant des traditionell optimistisch auftretenden Gastgebers noch an einen Erfolg glaubt . Kopenhagen wird eine Station von vielen werden auf dem schwierigen Weg der Staatengemeinschaft , die Treibhausgase **in** diesem Jahrhundert gerade noch auf ein halbwegs erträgliches Ausmaß zu begrenzen .

$$P(\text{in}) = ?$$

Markov Model
Example: Unigram

Universität Zürich[UZH]

Was sich seit Monaten angedeutet hatte , was nach dem letzten , ernüchternden Vorbereitungstreffen der Unterhändler **in** Barcelona keine Überraschung mehr ist , es reift nun zur sicheren Prognose : Das **in** diversen Konferenzen beschworene Ziel , **in** diesem Jahr noch den soliden Rahmen für ein globales Klimaschutz-Abkommen **in** der Nachfolge des Kyoto-Protokolls zu zimmern , ist illusorisch geworden . Daran ist nicht mehr zu zweifeln , wenn nicht einmal der dänische Ministerpräsident Lars Løkke Rasmussen als Repräsentant des traditionell optimistisch auftretenden Gastgebers noch an einen Erfolg glaubt . Kopenhagen wird eine Station von vielen werden auf dem schwierigen Weg der Staatengemeinschaft , die Treibhausgase **in** diesem Jahrhundert gerade noch auf ein halbwegs erträgliches Ausmaß zu begrenzen .

$$P(\text{in}) = \text{count}(\text{in}) \,/\, \text{count}(\text{all words})$$

$$P(\text{in}) = 5 \,/\, 117 \approx 0.042$$

Markov Model Example: Bigram

Universität Zürich[UZH]

Was sich seit Monaten angedeutet hatte , was nach dem letzten , ernüchternden Vorbereitungstreffen der Unterhändler in Barcelona keine Überraschung mehr ist , es reift nun zur sicheren Prognose : Das in diversen Konferenzen beschworene Ziel , **in diesem** Jahr noch den soliden Rahmen für ein globales Klimaschutz-Abkommen in der Nachfolge des Kyoto-Protokolls zu zimmern , ist illusorisch geworden . Daran ist nicht mehr zu zweifeln , wenn nicht einmal der dänische Ministerpräsident Lars Løkke Rasmussen als Repräsentant des traditionell optimistisch auftretenden Gastgebers noch an einen Erfolg glaubt . Kopenhagen wird eine Station von vielen werden auf dem schwierigen Weg der Staatengemeinschaft , die Treibhausgase **in diesem** Jahrhundert gerade noch auf ein halbwegs erträgliches Ausmaß zu begrenzen .

$$P(\text{diesem}|\text{in}) = ?$$

Markov Model
Example: Bigram

Was sich seit Monaten angedeutet hatte , was nach dem letzten , ernüchternden Vorbereitungstreffen der Unterhändler in Barcelona keine Überraschung mehr ist , es reift nun zur sicheren Prognose : Das in diversen Konferenzen beschworene Ziel , **in diesem** Jahr noch den soliden Rahmen für ein globales Klimaschutz-Abkommen in der Nachfolge des Kyoto-Protokolls zu zimmern , ist illusorisch geworden . Daran ist nicht mehr zu zweifeln , wenn nicht einmal der dänische Ministerpräsident Lars Løkke Rasmussen als Repräsentant des traditionell optimistisch auftretenden Gastgebers noch an einen Erfolg glaubt . Kopenhagen wird eine Station von vielen werden auf dem schwierigen Weg der Staatengemeinschaft , die Treibhausgase **in diesem** Jahrhundert gerade noch auf ein halbwegs erträgliches Ausmaß zu begrenzen .

$$P(\text{diesem}|\text{in}) = P(\text{in}, \text{diesem})/\text{P}(\text{in})$$

Markov Model
Example: Bigram

$$P(\text{diesem}|\text{in}) = P(\text{in, diesem})/P(\text{in})$$

$$P(\text{in diesem}) = P(\text{in diesem})/\ \#\text{bigrams}$$

$$P(\text{in diesem}) = 2/116$$

$$P(\text{diesem}|\text{in}) = \frac{\frac{2}{116}}{\frac{5}{117}} \approx 0.411$$

Markov Model
Example: Sentence probability

Sentence: This sentence has 5 tokens

➔ Bigram model?

➔ P(This sentence has 5 tokens)

$P(\text{This}|..) \times P(\text{sentence}|\text{this}) \times P(\text{has}|\text{sentence}) \times P(5|\text{has}) \times P(\text{tokens}|5)$

Markov Model
Python - defaultdict

```
def count1(text):
    absfreqs = {}
    for word in text.split():
        absfreqs[word] = absfreqs.setdefault(word, 0) + 1
    return absfreqs
```

Markov Model
Python - defaultdict

```
def count1(text):
    absfreqs = {}
    for word in text.split():
        absfreqs[word] = absfreqs.setdefault(word, 0) + 1
    return absfreqs
```

or using defaultdict:

```
from collections import defaultdict

def count2(text):
  freqs = defaultdict(int)
  for word in text.split():
    freqs[word] += 1
  return freqs
```

Markov Model
Python - defaultdict

This is a sentence . Each sentence has a number of words.
The number of sentences is 3.

Using nested dictionaries (Slide 19)

```
{'a': {'sample': 1, 'number': 1}, 'sentence': {'has': 1, '.': 1},
None: {'This': 1}, 'of': {'words': 1, 'sentences': 1},
'is': {'a': 1, '3': 1}, 'sentences': {'is': 1}, 'number': {'of': 2},
'.': {'The': 1, 'Each': 1}, 'sample': {'sentence': 1},
'This': {'is': 1}, '3': {'.': 1}, 'words': {'.': 1},
'Each': {'sentence': 1}, 'The': {'number': 1}, 'has': {'a': 1}}
```

Markov Model
Python - defaultdict

Using defaultdict:

```
def count_bigrams(text):
    f = defaultdict(lambda: defaultdict(int))
    hist = None
    for word in text:
        f[hist][word] += 1
        hist = word
    return f
```

Contents

1. Probabilities: Language Model
2. Probability Theory Background
3. Conditional Probabilities and MLE
4. Chain Rule
5. Markov Model
6. Examples Language Modelling

Examples Language Modelling
Statistical Machine Translation (SMT)

- Essential component of any SMT system

- LM ensures fluent sentences on the target side
  - It helps to decide about word order:

    $p_{\mathrm{LM}}(\text{the house is small}) > p_{\mathrm{LM}}(\text{small the is house})$
  - and word translation:

    Haus has multiple translations (*house*, *home*, …)

    $p_{\mathrm{LM}}(\text{I am going } \textit{home}) > p_{\mathrm{LM}}(\text{I am going } \textit{house})$

Examples Language Modelling
Optical Character Recognition (OCR)

- task: recognize the characters in a picture and generate the text, corresponding to them
- challenges:
  - formatting, location, picture-text mix
  - specific lexicon
  - font, italics/etc.
  - problems with old paper
  - sub-script/super-script
  - etc.

http://kitt.cl.uzh.ch/kitt/kokos/

Lecture 5: Probabilities

PCL II, CL, UZH
March 23, 2016

Universität Zürich[UZH]