Statistical Methods in Natural Language Processing (NLP)

Class 1: Introduction

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

GÖTEBORGS UNIVERSITET

Introduction

- As I speak, you are taking part in an unconscious game. You make **hypotheses** and try to **predict** what I am going to say, you might have guessed the following words, you also continuously restructure the utterances in your minds as I utter new ones. This can be easy in cases such, **went to the post-X**. (office).
- We do predictions all the time and this is not based on some kind of ocular knowledge but based on something real and really remarkable and astonishing: our innate ability to do linguistic predictions using the knowledge we have about the world.
- We will see that this basic idea is fundamental for understanding language structure, modeling language, finding information in texts and translating them. These methods and applications are some of the topics of this class.

GÖTEBORGS UNIVERSITET

Overview

- About the course.
- Introduction to Statistical Methods in NLP.
- Linguistic Data.
- Combinatorics.
- Python and R.

GÖTEBORGS UNIVERSITET

About the course

A. Understanding Probability Theory and Statistics

1. **Probability theory**: probability distributions (discrete and continuous), Randomness, Probability axioms, Bayesian probability, Conditional probability, Bayes' theorem, Moments, Central Limit Theorem.
2. **Information theory**: Claude E. Shannon's concept of information, entropy.
3. **Statistical theory**: (sampling, estimation, hypothesis testing)
4. **Machine Learning**
5. **Evaluation**: How do we know that a statistical text really accounts for the data. How do we know that a Machine Learning Algorithms learns something that is valuable?
6. **Applications**: How do we apply probability theory, information theory, machine learning to solve linguistic problems?

GÖTEBORGS UNIVERSITET

About the course

B. Language modeling: Applying Probability and Statistics in NLP

1. Python (and R)
2. Part-of-speech tagging
3. Syntactic parsing
4. Word sense disambiguation
5. Machine translation

GÖTEBORGS UNIVERSITET

About the course

1. Knowledge and understanding
   - 1.1 account for basic notions of probability theory, information theory and statistical theory
   - 1.2 give examples of how statistical methods have been applied in language technology systems
2. Skills and abilities
   - 2.1 apply statistical techniques to the development of language technology systems
   - 2.2 evaluate language technology applications using standard statistical tests
3. Judgement and approach
   - 3.1 choose the appropriate statistical method for a particular task
   - 3.2 evaluate the significance of statistical results

GÖTEBORGS UNIVERSITET

About the course: teachers and days

1. Teachers
   - 1.1 **Haris:** lectures/Tasks/Computer Assignments and course supervision.
   - 1.2 **Mehdi:** supervises/grades mandatory programming assignments.
2. Lectures
   - 2.1 all in **T346.**
   - 2.2 lab sessions/some lectures in the computer **Lab G212** (lab 4).
   - 2.3 Mondays & Thursdays at 13:15 and 10:15 respectively.

GÖTEBORGS UNIVERSITET

About the course: Assignments/Tasks/CAs

1. **Final Exams** (35%)
2. **Two (2) Assignments** (25%)
   - 2.1 Aim: Programming & Implementation of taught materials.
   - 2.2 **Deadline:** 1 week after the assignment.
3. **Weekly Tasks** (25%)
   - 3.1 Aim: Smaller assignments, which aim to help students understand and comprehend the taught materials.
   - 3.2 Finding solutions to problems.
   - 3.3 Programming tasks.
   - 3.4 **Deadline:** 3 days after the assignment.
4. **Three (3) Mandatory Programming Assignments**
   - 4.1 Aim: Programming & Implementation
   - 4.2 A. text/sound classification. (10%)
   - 4.3 B. classifier evaluation. (10%)
   - 4.4 C. PoS tagger implementation (10%)
   - 4.5 **Deadline:** 2 weeks after the assignment.

GÖTEBORGS UNIVERSITET

Course Literature: Course Materials

Materials prepared by me for the course:

1. **Course Notes**: Chapters that summarize the main ideas discussed.
2. **Supporting Materials with extra information about statistics/machine learning.**
3. **Programming Materials for Python and R.**
4. **Presentations** (i. Viewing on tablets/computers 2. Printing).
5. **Code Examples**.
6. *Visit the course's website for more information.*

GÖTEBORGS UNIVERSITET

Course Literature: Course Books

Course Books

1. Christopher Manning and Hinrich Schtze (1999) Foundations of Statistical Natural Language Processing, Cambridge, Massachusetts, USA. MIT Press.
2. Joseph K. Blitzstein, Jessica Hwang (2014). Introduction to Probability. London: CRC Press. Taylor & Francis.
3. James Gareth, Witten Daniela, Hastie Trevor and Robert Tibshirani (2013).

GÖTEBORGS UNIVERSITET

Course Literature: Complementary Textbooks

Complementary Textbooks

1. Daniel Jurafsky and James Martin (2008) An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition. Prentice Hall.
2. Russell, Stuart J.; Norvig, Peter (2009), Artificial Intelligence: A Modern Approach (3rd ed.), Upper Saddle River, New Jersey: Prentice Hall, ISBN 0-13-604259-7.

GÖTEBORGS UNIVERSITET

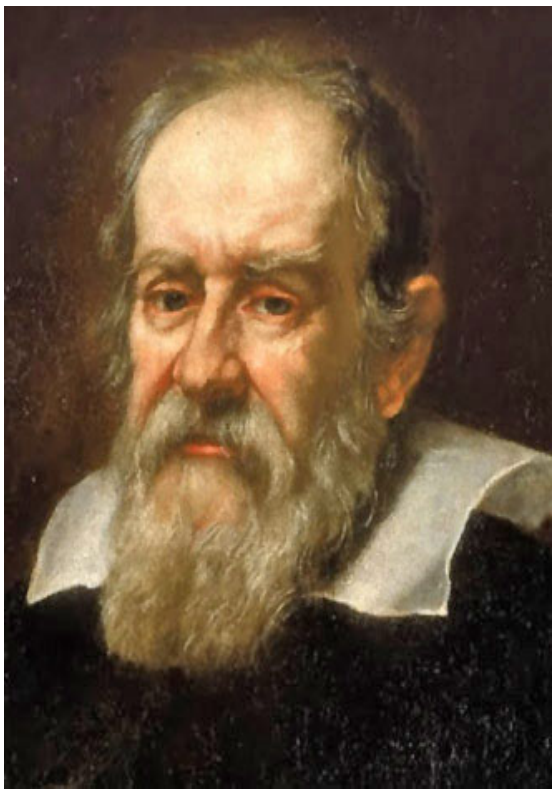
Course Literature: Resources

Resources

1. The Classification And REgression Training (caret) Package R. http://topepo.github.io/caret/index.html
2. The R Project for Statistical Computing
3. Python Programming Language
4. Scipy/Numpy Quickstart Tutorial

GÖTEBORGS UNIVERSITET

Probabilities as a window to language understanding



"La filosofia scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'universo), ma non si pu intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali scritto. Egli scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezzi impossibile a intenderne umanamente parola; senza questi un aggirarsi vanamente per un oscuro laberinto." (Galileo Galilei (1564–1642), Il Saggiatore, Cap. VI)

GÖTEBORGS UNIVERSITET

Probabilities as a window to language understanding

Some Case Studies:

1. Probabilities in Speech Perception.
2. Probabilities in Phonology.
3. Probabilistic Reasoning.

GÖTEBORGS UNIVERSITET

Probabilities and NLP applications

Probabilistic methods for understanding language are central in applications such as the following:

1. Machine Translation
2. Discourse Agents
3. Text Summarization
4. Text to Speech Systems
5. Speech to Text Systems
6. Topic segmentation
7. Information retrieval
8. Information extraction

GÖTEBORGS UNIVERSITET

Development of NLP: Early Period

1. Automata
2. Regular Expressions
3. Information Theory
4. Generative Grammar
5. Early Programming Languages

GÖTEBORGS UNIVERSITET

1960s and the Hippies

1. Chomsky and Harris
2. Marvin Minsky, McCarty, Shannon, Rochester
3. Probabilistic Models in Language study
4. The *logic theorist* and the *General Problem Solver* by Newell and Simon.
5. Heuristic approaches, rule based approaches to phonetics, features.
6. 1968-1970 SHRDLU was an early natural language understanding computer program, developed by Terry Winograd at MIT. In SHRDLU, the user carries on a conversation with the computer, moving objects, naming collections and querying the state of a simplified "blocks world".
7. Acoustic Theory of Speech Perception and the Source-filter theory by Gunnar Fant.
8. 1965– 1987 1st Generation of Automatic Speech Recognition based Heuristic Approaches
9. 1968 – 1980 2nd Generation of Automatic Speech Recognition based on LPC, FFT, DTW.

GÖTEBORGS UNIVERSITET

1970s

1. Klatt research on text to speech systems
2. HMMs
3. Prolog
4. Definite Clause Grammars
5. Natural Language Understanding
6. Network Based Semantics
7. Discourse Modelling
8. Digital Signal Processing,
9. Digital Filtering FFTs
10. Digital Communications
11. The precursor to the Internet, in 1971 ARPANET 1969 email Raymond Samuel "Ray" Tomlinson (April 23, 1941 March 5, 2016) o implemented the first email program on the ARPANET system.
12. Linear Predictive Coding LPC
13. 1975 Språkbanken University of Gothenburg

GÖTEBORGS UNIVERSITET

1980

1. Neural Networks
2. 1989- DARPA Project
3. Basic properties became clear
4. suffieciency of two layer network (one hidden layer) to model any non-linear function.
5. Map feature vectors into dimensional space where classes are more separable with hyperplanes
6. New error criterion based on Kullback-Leibler (1990)

GÖTEBORGS UNIVERSITET

1990

1. Probabilities and Statistics become mainstream in language research
2. IBM Thomas J. Watson

GÖTEBORGS UNIVERSITET

2000

1. Language Data Consortium
2. Penn Treebank
3. Praque Dependency Treebank

GÖTEBORGS UNIVERSITET

2010

The rise of deep neural networks

GÖTEBORGS UNIVERSITET

What do we study?

- **General corpora**: attempt to represent a language as a whole, e.g., the Brown Corpus, one million words from 500 American English texts of approximately 2000 words each, distributed across fifteen genres.
- **Specific corpora**: (e.g., Commerce, Finance, Food, or Law).
- Raw or **non-annotated** corpora.
- **Annotated corpora**.
- **Parallel corpora**. A parallel corpus is a corpus that consists of the same text in two or more languages. Parallel corpora can inform linguistic theory in general, enable the comparison of languages, and can have applications, such as in translation.

GÖTEBORGS UNIVERSITET

Frequency Lists

Table: Frequency table of the twenty most frequent words in the text 'Alice's Adventures in Wonderland' by Lewis Carroll.

Word	Freq(f)	Rank(r)	f.r
the	1630	1	1630
and	844	2	1688
to	721	3	2163
a	627	4	2508
she	537	5	2685
it	526	6	3156
of	508	7	3556
said	462	8	3696
i	400	9	3600
alice	385	10	3850
in	366	11	4026
you	360	12	4320
was	357	13	4641
that	276	14	3864
as	262	15	3930
her	248	16	3968
at	209	17	3553
on	193	18	3474
with	181	19	3439
all	179	20	3580
...			

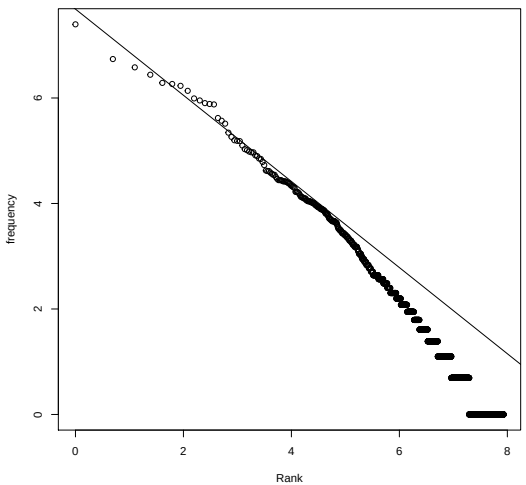
GÖTEBORGS UNIVERSITET

1. *grammatical* or
2. *function* words.
3. *content* words.

1. Most common words that appear are **function words**: e.g., the, and, to, etc.
2. 18% of words of words account for 80 % of word occurrences.
3. The others appear only once! They are called *hapax legomena*.

Zipf's law

An empirical study by the American linguist George Kingsley Zipf (1902–1950), showed that the frequency of any word is inversely proportional to its rank in the frequency table. So, the frequency



$$f = \frac{1}{r} \tag{1}$$

and according to this there is a constant k that derives from:

$$k = f \cdot r \tag{2}$$

GÖTEBORGS UNIVERSITET

Collocations

Collocations are expressions that are often found together, some of these are kind of stereotypical such as, ;

- *adjective & noun:* heavy smoker, bright idea, technical knowledge
- *Noun & noun:* sun glasses
- *verb & nouns:* take a look, make a difference, make a mess
- *verb & adverb:* talk freely
- *adverb & adjective:* strongly opposed, ridiculously easy, utterly ridiculous.
- *phrasal verbs* e.g., ask around, cheer up **Applications:** Lexicography, Machine Translation, etc.

GÖTEBORGS UNIVERSITET

Collocations

- What are the common collocates of the word **dog**?
- Can we predict the following word in a sentence?
- **Collocation Finder**: **Using Statistics** (not Astrology)!

GÖTEBORGS UNIVERSITET

Python and R

- **R** is a domain specific language. It is made for doing statistics, probabilities, and machine learning, and creating graphs. It has been also enhanced with 9813 available packages. You can download R from `https://www.r-project.org` and Rstudio from `https://www.rstudio.com`.
- **Python** is a generic language, yet it has been enhanced with added functionalities to do statistics. As a generic language it is better suited to built NLP applications.
- Python is required for the course; learning R is optional.

GÖTEBORGS UNIVERSITET

Python libraries

- **Pandas:** Facilitates working with complex data structures. Pandas Documentation: `http://pandas.pydata.org/pandas-docs/stable/`
- **NumPy:** Provides mathematical functions. NumPy Documentation:
- **Matplotlib:** Provides plotting functionality.
- **NLTK:** is a library with functionalities for natural language processing.

Quick start (proposed method): download Python and packages with Anaconda from `https://www.continuum.io/`.

GÖTEBORGS UNIVERSITET

Example: Importing Data in Python

An experimenter measured the duration of i] in three different experiments. The task is to open a simple data file, read the data, plot the data, get the means and the standard deviation.

```
# Import the data
# Create an empty list with the name "data":
data = []
```

GÖTEBORGS UNIVERSITET

Example: Importing Data in Python

Open the csv file with data and split it based on ";" the attach each part to two variables "exp" and "dur". Then append these variables to the list "data".

```
with open('duration.csv') as durdata:
 for mydata in durdata:
        exp, dur = mydata.split(";")
        data.append((exp, int(dur)) )
```

GÖTEBORGS UNIVERSITET

Example: Importing Data in Python

The list "data" now has this form: [('A', 199), ('A', 184), ...('C', 182)]. To get only the numbers, we provide the following command. Observe the position of the _. It indicates that we are not interested for the values in that position.

```
duration = [dur for _, dur in data]
# We also select the values for each experiment.
ExperimentA = [ dur for _, dur  in data if exp == 'A']
ExperimentB = [ dur for _, dur  in data if exp == 'B']
ExperimentC = [ dur for _, dur  in data if exp == 'C']
```

GÖTEBORGS UNIVERSITET

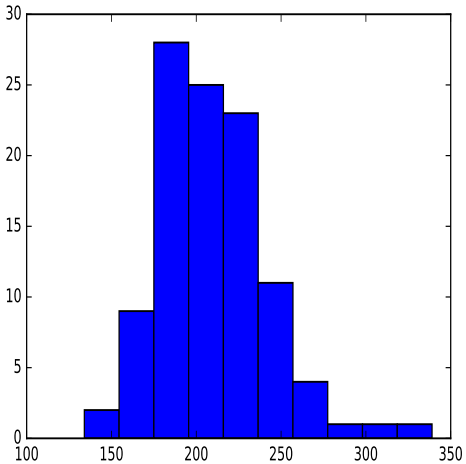
Example: Basic plots and statistics in Python

First we need to import the numpy library and the pyplot from the matplotlib. This is dones as follows:

```
import numpy as np
import matplotlib.pyplot as plt
```

GÖTEBORGS UNIVERSITET

Example: Basic plots and statistics in Python




```
plt.hist(duration)
plt.show() # This shows the plot on screen.
# plt.savefig('hist.pdf') # You may activate this to save the plot as a pdf file (do not forget to deactivate plt.show() ).
```

GÖTEBORGS UNIVERSITET

Example: Basic plots and statistics in Python

Calculate the mean of duration in ms

```
np.mean(duration)
209.86666666666667
```

Calculate the standard deviation of duration in ms

```
np.std(duration)
31.924335147394608
```

Calculate the variance of duration

```
np.var(duration)
1019.1631746031746
```

Calculate the median

```
np.median(duration)
206.0
```

GÖTEBORGS UNIVERSITET

Example: Basic plots and statistics in Python

Calculate the min value

```
np.percentile(duration,0)
134.0
```

Calculate the 25% percentile

```
np.percentile(duration,25)
189.0
```

Calculate the 50% percentile or median

```
np.percentile(duration,50)
206.0
```

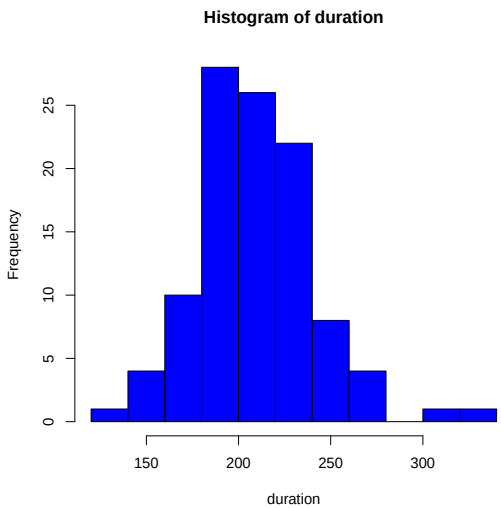
Calculate the 75% percentile

```
np.percentile(duration,75)
224.0
```

Calculate the maximum value

```
np.percentile(duration,100)
339.0
```

Example: Basic plots and statistics in R



dur $<-$ **read.csv("duration.csv", sep = ";")** # Import data
attach(dur) # Create variables using the labels of the table as headings
hist(duration, col="blue")
mean(duration)

```
[1] 209.8667
```

sd(duration)

```
[1] 32.07745
```

var(duration)

```
[1] [1] 1028.963
```

summary(duration)

```
Min.  1st Qu.  Median Mean 3rd Qu.
Max.
134.0 189.0 206.0 209.9 224.0 339.0
```

GÖTEBORGS UNIVERSITET

R vs. Python Output

R and Python provide different results for standard deviation and variance.The reason is that R, Matlab, Excel etc., calculate variance by dividing n-1, this is known as **Bessels correction**.

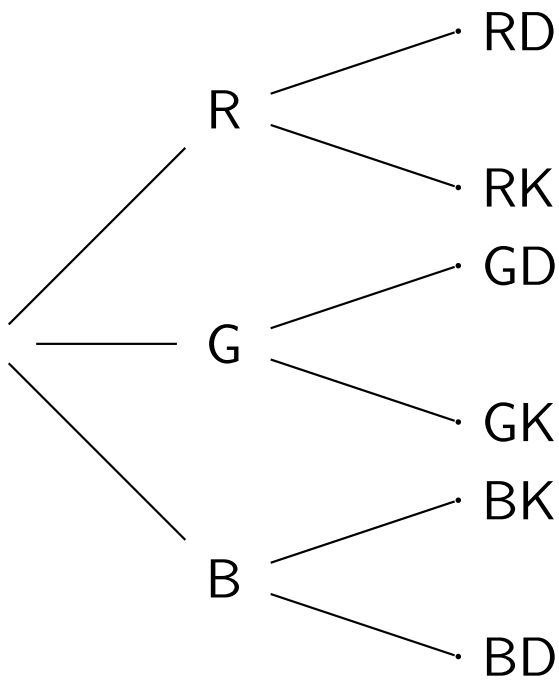
$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

In contrast, Python calculates the variance from the whole population, that it divides by N:

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

So, to get the same answers we need to modify the Python functions as follows: `np.var(duration, ddof=1)` and `np.std(duration, ddof=1)`.

Combinatorics



Problem: Imagine that we have three shirts:

- a blue shirt,
- a gray shirt, and
- a red shirt

lets us indicate these as $shirt = \{B, G, R\}$ and two pairs of hats

- a khaki and
- a dark blue

lets denote these as $hats = \{K, D\}$ and we want to see which combination is better for us. How many possible combinations are there?

The answer is 6.
$shirt \times hat = \{BK, BD, GK, GD, RK, RD\}$

GÖTEBORGS UNIVERSITET

Proof of the fundamental principle of counting

To proof the fundamental principle of counting one should enumerate all the possible outcomes of the two tests; i.e.,

$$\begin{pmatrix} (1,1) & (1,2) & \cdots & (1,n) \\ (2,1) & (2,2) & \cdots & (2,n) \\ \vdots & \vdots & \ddots & \vdots \\ (m,1) & (m,2) & \cdots & (m,n) \end{pmatrix} \tag{3}$$

so, the possible outcomes from both tests is (i,j) if the first test results in its i_{th} outcomes and the second test in its j_{th} outcomes. Hence, the set of possible outcomes consists of m rows, each containing n elements. If there is k number of tests, then there is a total of $n_1 \times n_2 \ldots n_k$ possible outcomes of the k tests.

GÖTEBORGS UNIVERSITET

Permutations

To find how many possible arrangements are possible with the three letters a, b, c then there are 6 possible permutations: $1 \times 2 \times 3 = 6$. In general to determine the number of permutations in set of n things that does not include repetitions, then this number is $n!$, i.e.,

$$n(n1)(n2)3 \times 2 \times 1 = n! \tag{4}$$

GÖTEBORGS UNIVERSITET

Next Class

We will discuss the following:

1. Probability Theory