Statistics for NLP

May 27, 2013

Statistics for NLP

< D > < P > < P > < P >

э

э

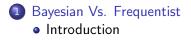
Outline



- Measuring the goodness of a language model
 Measuring the quality of a model for a language
- Measuring the quality of a system
 Introduction
 - Significance tests

Introduction

Outline



2 Measuring the goodness of a language model• Measuring the quality of a model for a language

Measuring the quality of a system
 Introduction
 Significance tests

Statistics for NLP

Introduction

Bayesian Vs. Frequentist

• Two approaches to probability reasoning

Frequentist

- Data are *repeatable* random samples from an underlying distribution
- This distribution is *fixed*
- The probability of an event is its frequency (in the limit)

- Data is what we observe from the sample
- The underlying parameters may vary
- And are treated as distributions themselves

Introduction

Bayesian Vs. Frequentist

• Two approaches to probability reasoning

Frequentist

- Data are repeatable random samples from an underlying distribution
- This distribution is *fixed*
- The probability of an event is its frequency (in the limit)

- Data is what we observe from the sample
- The underlying parameters may vary
- And are treated as distributions themselves

Introduction

Bayesian Vs. Frequentist

• Two approaches to probability reasoning

Frequentist

- Data are repeatable random samples from an underlying distribution
- This distribution is *fixed*
- The probability of an event is its frequency (in the limit)

- Data is what we observe from the sample
- The underlying parameters may vary
- And are treated as distributions themselves

Introduction

An example

• We want to know the average height h of adult males in Rome

Frequentist

- There is an underlying distribution *H*
- With sufficient sampling we can estimate *h*
- *h* is just an unknown number!
 - It doesn't make sense asking what P(160 < h < 190) is!
 - Only P(160 < H < 190)

- We can use a prior probability on what *h* may be
- Without any information we could say *h* is uniform in [130,250]
- We *can* ask *P*(160 < *h* < 190)

Introduction

An example

• We want to know the average height h of adult males in Rome

Frequentist

- There is an underlying distribution *H*
- With sufficient sampling we can estimate *h*
- *h* is just an unknown number!
 - It doesn't make sense asking what P(160 < h < 190) is!
 - Only *P*(160 < *H* < 190)

- We can use a prior probability on what *h* may be
- Without any information we could say *h* is uniform in [130,250]
- We *can* ask *P*(160 < *h* < 190)

Introduction

An example

• We want to know the average height h of adult males in Rome

Frequentist

- There is an underlying distribution *H*
- With sufficient sampling we can estimate *h*
- *h* is just an unknown number!
 - It doesn't make sense asking what P(160 < h < 190) is!
 - Only P(160 < H < 190)

- We can use a prior probability on what *h* may be
- Without any information we could say *h* is uniform in [130,250]
- We *can* ask *P*(160 < *h* < 190)

- A Bayesian approach allows us to reason about events for which there is no associated frequency
 - What is the probability that a suspect committed a crime, given the evidence?
- It permits to include prior information and subjective belief in calculations
 - What is the probability that the sun will rise tomorrow?

- A Bayesian approach allows us to reason about events for which there is no associated frequency
 - What is the probability that a suspect committed a crime, given the evidence?
- It permits to include prior information and subjective belief in calculations
 - What is the probability that the sun will rise tomorrow?

Model

Outline



Measuring the goodness of a language model
 Measuring the quality of a model for a language

Measuring the quality of a system
 Introduction
 Significance tests

Statistics for NLP

Model

Model for a language.

- A model for a language is a way that we human use to represent language
 - For example the representation of a sentence by means of a syntactic tree
- Models do *not* exist in nature!
 - Saying that a certain word is an adjective is a human construction!
- A model may or may not be a good representation of the language itself

Model

Model for a language.

- A model for a language is a way that we human use to represent language
 - For example the representation of a sentence by means of a syntactic tree
- Models do not exist in nature!
 - Saying that a certain word is an adjective is a human construction!
- A model may or may not be a good representation of the language itself

Model

Model for a language.

- A model for a language is a way that we human use to represent language
 - For example the representation of a sentence by means of a syntactic tree
- Models do not exist in nature!
 - Saying that a certain word is an adjective is a human construction!
- A model may or may not be a good representation of the language itself

Model

How to measure the quality of a model?

Idea:

If two (or more) persons agree in doing a specific task that tests the model, it *may* mean that the model is near to the language representation that people have in mind

- Take two or more persons (annotators) and teach them a specific (linguistic) task
 - for example POS tagging on a small set of sentences
- These annotators execute the task indipendently on the same set
- Measure how much do they agree!
- High agreement will also guarantee that the corpus will be annotated in a reliable manner

Image: Image:

Model

How to measure the quality of a model?

Idea:

If two (or more) persons agree in doing a specific task that tests the model, it *may* mean that the model is near to the language representation that people have in mind

- Take two or more persons (annotators) and teach them a specific (linguistic) task
 - for example POS tagging on a small set of sentences
- These annotators execute the task indipendently on the same set
- Measure how much do they agree!
- High agreement will also guarantee that the corpus will be annotated in a reliable manner

Model

How to measure the quality of a model?

Idea:

If two (or more) persons agree in doing a specific task that tests the model, it *may* mean that the model is near to the language representation that people have in mind

- Take two or more persons (annotators) and teach them a specific (linguistic) task
 - for example POS tagging on a small set of sentences
- These annotators execute the task indipendently on the same set
- Measure how much do they agree!
- High agreement will also guarantee that the corpus will be annotated in a reliable manner

A D b 4 A

Model

How to measure the quality of a model?

Idea:

If two (or more) persons agree in doing a specific task that tests the model, it *may* mean that the model is near to the language representation that people have in mind

- Take two or more persons (annotators) and teach them a specific (linguistic) task
 - for example POS tagging on a small set of sentences
- These annotators execute the task indipendently on the same set
- Measure how much do they agree!
- High agreement will also guarantee that the corpus will be annotated in a reliable manner

Model

How to measure the quality of a model?

Idea:

If two (or more) persons agree in doing a specific task that tests the model, it *may* mean that the model is near to the language representation that people have in mind

- Take two or more persons (annotators) and teach them a specific (linguistic) task
 - for example POS tagging on a small set of sentences
- These annotators execute the task indipendently on the same set
- Measure how much do they agree!
- High agreement will also guarantee that the corpus will be annotated in a reliable manner

Inter-rater agreement

- Not really a good measure!
- We are not taking into account the possibility that the annotators may agree only by chance

Inter-rater agreement

$$\frac{A_o - A_e}{1 - A_e}$$

- A_o is the fraction of times the annotators agree
- A_e is the probability the annotators agree by chance on the set:
 - Let C_1 and C_2 be the annotators, and k the categories they have to choose from
 - $A_e = \sum_k P(k|C_1) \cdot P(k|C_2)$
 - Different assumptions on $P(k|C_i)$ lead to different statistics
- Works only with 2 annotators!

Inter-rater agreement

- Not really a good measure!
- We are not taking into account the possibility that the annotators may agree only by chance

Inter-rater agreement

$$\frac{A_o - A_e}{1 - A_e}$$

- A_o is the fraction of times the annotators agree
- A_e is the probability the annotators agree by chance on the set:
 - Let C_1 and C_2 be the annotators, and k the categories they have to choose from
 - $A_e = \sum_k P(k|C_1) \cdot P(k|C_2)$
 - Different assumptions on $P(k|C_i)$ lead to different statistics
- Works only with 2 annotators!

Model

κ -statistics

κ -statistics (Cohen, 1960)

- Assumes each annotator has its own bias, in general $P(k|C_1) \neq P(k|C_2)$
- Estimates $P(k|C_j)$ with $\hat{P}(k|C_j) = \frac{n_{C_j,k}}{i}$
 - $n_{C_j,k}$ is the number of times category k get chosen by annotator C_j
 - *i* is the total number of choice made by each annotator

•
$$A_e = \sum_k \hat{P}(k|C_1) \cdot \hat{P}(k|C_2) = \sum_k \frac{n_{C_1,k}}{i} \frac{n_{C_2,k}}{i} = \frac{1}{i^2} \sum_k n_{C_1,k} n_{C_2,k}$$

Model

κ -statistics (more than 2 annotators)

Multi- κ (Fleiss, 1971)

A different definition is needed if we have c > 2 annotators:

- We define A_o in terms of pairwise agreement:
 - Let $n_{i,k}$ the number of annotators that put element i in category k
 - For each item *i* we count how many pairs of annotators agree on it out of all pairs:

$$\operatorname{agr}_{i} = \frac{1}{\binom{c}{2}} \sum_{k} \binom{n_{i,k}}{2}$$

• We obtain A_o by averaging:

$$A_o = \frac{1}{i} \sum_i \operatorname{agr}_i$$

Model

κ -statistics (more than 2 annotators)

Multi- κ (Fleiss, 1971)

For the expected agreement we have:

$$A_{e} = \sum_{k} \frac{1}{\binom{c}{2}} \sum_{(I,m),l < m} (\frac{1}{i} n_{C_{l},k}) (\frac{1}{i} n_{C_{m},k})$$

This can be shown to be equal to the average of the two-annotator version of A_e over all pairs of annotators

Model

Guidelines

Guidelines (Landis & Koch, 1977)

There is *no absolute standard* on the values of κ -statistics, Landis & Koch suggest:

κ	Agreement
< 0	no agreement
$0 \sim 0.2$	slight
$0.2 \sim 0.4$	fair
$0.4 \sim 0.6$	moderate
$0.6\sim0.8$	substantial
$0.8 \sim 1$	almost perfect

• κ values tend to be higher when there are fewer categories

If $\kappa < 0.6$ we probably don't have a good model!

< - 12 →

Introduction Significance tests

Outline

- Bayesian Vs. Frequentist
 Introduction
- Measuring the goodness of a language model
 Measuring the quality of a model for a language
- Measuring the quality of a system
 Introduction
 - Significance tests

Introduction Significance tests

- We have two systems S_1 and S_2
 - For example two algorithms for recognizing textual entailment
- We want to decide wheter these two systems are really different or are in fact the same
- Systems receive an input and produce a *result*:
 - The result may be a measure like precision, recall, f-measure,...
- The input set C is only a small subset of all the possible inputs.
 - Similar results on *C* implies similar in general?
 - If they are different, how significant is the difference?

- We have two systems S_1 and S_2
 - For example two algorithms for recognizing textual entailment
- We want to decide wheter these two systems are really different or are in fact the same
- Systems receive an input and produce a *result*:
 - The result may be a measure like precision, recall, f-measure,...
- The input set C is only a small subset of all the possible inputs.
 - Similar results on *C* implies similar in general?
 - If they are different, how significant is the difference?

- We have two systems S_1 and S_2
 - For example two algorithms for recognizing textual entailment
- We want to decide wheter these two systems are really different or are in fact the same
- Systems receive an input and produce a *result*:
 - The result may be a measure like precision, recall, f-measure,...
- The input set C is only a small subset of all the possible inputs.
 - Similar results on *C* implies similar in general?
 - If they are different, how significant is the difference?

- We have two systems S_1 and S_2
 - For example two algorithms for recognizing textual entailment
- We want to decide wheter these two systems are really different or are in fact the same
- Systems receive an input and produce a *result*:
 - The result may be a measure like precision, recall, f-measure,...
- The input set C is only a small subset of all the possible inputs.
 - Similar results on *C* implies similar in general?
 - If they are different, how significant is the difference?

Introduction Significance tests

Outline

- Bayesian Vs. Frequentist
 Introduction
- Measuring the goodness of a language model
 Measuring the quality of a model for a language
- Measuring the quality of a system
 Introduction
 - Significance tests

Significance test

Introduction Significance tests

Probabilities:

We think of S_1 and S_2 as two random variable with unknown distributions.

- $P(\operatorname{res}(S_1)|C)$
- $P(\operatorname{res}(S_2)|C)$

Null Hypothesis H_0

- We assume that two distributions are the same
- We want to quantify the probability that H_0 is false

< ロ > < 同 > < 回 > <

Introduction Significance tests

Significance test

Probabilities:

We think of S_1 and S_2 as two random variable with unknown distributions.

- $P(\operatorname{res}(S_1)|C)$
- $P(\operatorname{res}(S_2)|C)$

Null Hypothesis H_0

- We assume that two distributions are the same
- We want to quantify the probability that H_0 is false

< ロ > < 同 > < 回 > <

Significance test

Introduction Significance tests

Significance test:

Any procedure that gives a probability that the null hypothesis is rejected

• There are a number of significance test

• Sign test, Wilcoxon test, student *t*-test, χ^2 test,...

- Each test makes some assumption on the distributions
- Each work by calculating a *test statistic q* that, if H₀ is true, follows a given distribution

Significance test

Introduction Significance tests

Significance test:

Any procedure that gives a probability that the null hypothesis is rejected

- There are a number of significance test
 - Sign test, Wilcoxon test, student *t*-test, χ^2 test,...
- Each test makes some assumption on the distributions
- Each work by calculating a *test statistic q* that, if H₀ is true, follows a given distribution

Significance test

Introduction Significance tests

Significance test:

Any procedure that gives a probability that the null hypothesis is rejected

- There are a number of significance test
 - Sign test, Wilcoxon test, student *t*-test, χ^2 test,...
- Each test makes some assumption on the distributions
- Each work by calculating a *test statistic q* that, if H₀ is true, follows a given distribution

Introduction Significance tests

Significance test

• Significance can be expressed via *p*-value or *z*-score

p-value

The probability that, assuming H_0 true, the test statistic assume a value as extreme as q

z-score

Assuming H_0 true, the number of standard deviations that q differs from the mean

• • • • • • • • • • •

Introduction Significance tests

Significance test

• Significance can be expressed via *p*-value or *z*-score

p-value

The probability that, assuming H_0 true, the test statistic assume a value as extreme as q

z-score

Assuming H_0 true, the number of standard deviations that q differs from the mean

Introduction Significance tests

Sign test

- One of the simplest
- Few assumptions

Given two random variables X and Y, we want to test the null hypothesis H_0 :

P(X > Y) = 0.5

Statistics for NLP

Sign test

Algorithm

- Randomly sample *n* pairs of points (x_i, y_i)
- If there are pairs in which $x_i = y_i$
 - Discard those pairs and let m be the new number of points
- Let w be the number of times that $y_i > x_i$
- If H_0 is true we expect $w \sim B(m, 0.5)$
 - That is: $P(w = k) = {m \choose k} \frac{1}{2}^m$
- Let $p = P(W \leq w)$
- If $p < p_{critical}$: reject H_0

Wilcoxon sign test

An improved version of the sign test

- Sample *m* pairs (x_i, y_i) such that $x_i \neq y_i \ \forall i$
- For each pair compute $|x_i y_i|$ and rank them from smallest (R = 1) to largest:
 - If some pairs are tied, give them a rank that is the average of their ranks
- Compute:

$$W = \sum_{i=1}^{m} \operatorname{sign}(x_i - y_i) \cdot R_i$$

• Let
$$\sigma = \sqrt{rac{m(m+1)(2m+1)}{6}}$$
 and $z = rac{W-0.5}{\sigma}$

• If $z > z_{critical}$: Reject H_0

Introduction Significance tests

Student's t-test

• Any test for which, if H_0 is true, we obtain a Student *t*-distribution

Student's *t*-distribution

The Student's *t*-distribution with v degree of freedoms has probability density function given by:

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})} (1 + \frac{t^2}{v})^{-\frac{v+1}{2}}$$

Statistics for NLP

Student's *t*-test

Introduction Significance tests

Student's test

Given two sample X and Y, test if they have the same mean Assumption:

- X and Y have the same size n
- We can assume the distributions have the same variance

Introduction Significance tests

Algorithm

- Compute the sample mean and variance: \overline{X} , \overline{Y} , S_X^2 , S_Y^2
- Compute:

$$S_{XY} = \sqrt{rac{1}{2}(S_X^2 + S_Y^2)}$$

• Compute:

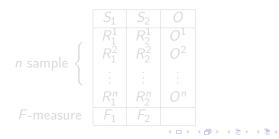
$$t = \frac{\overline{X} - \overline{Y}}{S_{XY}\sqrt{\frac{2}{n}}}$$

• Compare t with a t-distribution with 2n-2 degrees of freedom

Image: Image:

Introduction Significance tests

- All the previous tests requires a sufficient amount of data to be sampled
- Annotated corpus for NLP are in generale expensive and laborious to obtain
- Suppose that we have a test set T of n elements:
 - For both systems we can calculate the *F*-measure on *T* (based on the *oracle O*)



Introduction Significance tests

- All the previous tests requires a sufficient amount of data to be sampled
- Annotated corpus for NLP are in generale expensive and laborious to obtain
- Suppose that we have a test set *T* of *n* elements:
 - For both systems we can calculate the *F*-measure on *T* (based on the *oracle O*)



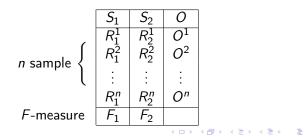
Introduction Significance tests

- All the previous tests requires a sufficient amount of data to be sampled
- Annotated corpus for NLP are in generale expensive and laborious to obtain
- Suppose that we have a test set T of n elements:
 - For both systems we can calculate the *F*-measure on *T* (based on the *oracle O*)



Introduction Significance tests

- All the previous tests requires a sufficient amount of data to be sampled
- Annotated corpus for NLP are in generale expensive and laborious to obtain
- Suppose that we have a test set T of n elements:
 - For both systems we can calculate the *F*-measure on *T* (based on the *oracle O*)



Introduction Significance tests

A test for NLP tasks

- Let's say that $F_1 > F_2$, and call $d = F_1 F_2$
 - Is S_1 really a better system or we were lucky?
 - We want to test the null hypothesys that S_1 and S_2 are equivalent

(A. Yeh, 2000)

- We generate a number *m* of *fictitious* systems S₁ⁱ and S₂ⁱ obtained by swapping an element in S₁ⁱ⁻¹ with one in S₂ⁱ⁻¹
- We compute $d^i = F(S_1^i) F(S_2^i)$ for each system
- Let k be the number of time $d^i > d$
- Compute $p = \frac{k}{m}$
- p is the probability of obtaining d under H_0

(日) (同) (三) (