Accelerated Natural Language Processing 2018

Lecture 28: Evaluation and significance

Henry S. Thompson
19 November 2018

1. The nature of evaluation

The scientific method is founded on making and testing hypotheses

Evaluation is just another name for testing

Sometimes our hypotheses are about existing linguistic objects:

- Is this text by Shakespear or Marlowe?
- Is this tweet in French or in Spanish?
- How many distinct authors can we detect in Homer, or in Genesis?
- Did the defendant actually write the document presented as his confession?
- Are Bush's inaugural speeches simpler than Clinton's?

2. In this lecture...

Not the details of how to evaluate a particular system

But the concepts, methods and materials which are drawn on to do this

3. The nature of evaluation, cont'd

Sometimes they are about the output of our systems:

- How well does this model represent the data?
  - For example, bigrams vs. trigrams
- How accurate is this segmenter/tagger/parser/disambiguator?
- How good is this information retrieval system?
- Is this segmenter/tagger/parser/disambiguator/IR system better than that one?

Sometimes they are about human beings:

- How reliable is this person's annotation?
- To what extent do these two annotators agree?

4. Measuring differences

Our hypotheses are often about *differences*

- Some experimental **condition** is manipulated
  - Or identified
- And some outcome is measured in the two conditions
  independent variable
  What we manipulate

  dependent variable
  What we measure
- And we ask: is a change in the condition reliably reflected in some measured outcome value?

We typical look at a number of **trials** (repetitions) in each condition

- different values for one or more independent variables

And ask whether the resulting **distributions** (population of values) are different or not

But what *counts* as different?

- Or, is the difference **significant**?

5. Gold standards and evaluation measures against them

In many cases we have a record of 'the truth'

That is, the best human judgement as to what the correct segmentation/tag/parse/reading is, or what the right documents are in response to a query.

Gold standards can be used both for training and for evaluation

But reliable testing *must* be done on unseen data:

> Don't use your training data for (reportable) testing!

Crucially, evaluation isn't just for public review:

- It's how you manage internal development
- That is, how systems improve themselves (see whole course on Machine Learning)
  - Often this means a division between **training** data and **development test** data
  - That's an *internal* decision
  - As opposed to an *external* holding back of test data

[Read section 4.8 of J&M (3rd edition; 5.7 in 2nd edition) for a good review of all this]

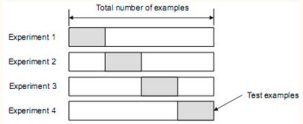
6. Making the best use of available data

Always using the same, say, 75% for training, 15% for development testing and 10% for 'real' testing isn't the best possible use of your data

- Especially if the dataset is not as large as you'd like

The answer is **cross-validation** (also sometimes referred to as **jackknife**)

- ***k*-fold** cross-validation: Partition the data into k pieces and treat them as mini develpment test sets
- Each fold is an experiment with a different held-out set, using the rest of the data for training:



- After *k* folds, every data point will have a held-out prediction

If tuning the system via cross-validation, still important to have a separate blind test set

7. The cost of gold standards

Gold standards can be *very* expensive

Because they involve lots of trained human annotators

- For reliability, you need to double- or even triple-annotate at least some of your data
- "There's no data like more data" -- to be of use for training and evaluation, you need lots.

For example, as near as I can estimate, the Penn treebank

- around 285,000 parsed sentences, 4.5 million words
- would have taken around 4 person years to produce
  - not including supervision and quality control
  - *if* the annotators had all started out fully trained and working at full capacity

Virtually all the large gold standards available today have been paid for by government agencies.

8. Gold standards without experts

It is no longer always necessary to employ more-or-less experts as annotators

- So-called **social machines** can be used instead
  - By **social machine** is meant mechanisms for exploiting the opportunity the World Wide Web offers us to recruit very large numbers of unpaid or low-paid people for a wide range of tasks
  - Examples include Wikipedia, SETI, Galaxy Zoo and reCAPTCHA
  - Another name for this is **crowd sourcing**

Amazon's **Mechanical Turk** is a social machine which has been widely used for the creation of gold standards for NLP tasks

9. Amazon's Mechanical Turk

Named after a 19th century fake automaton

- Sometimes described as "artificial artificial intelligence"
- That is, mechanising (and monetising) the deployment of large numbers of human beings to perform simple tasks
  - Simple, but not within the capacity of machines

A typical task, or HIT (**Human Intelligence Task**)

- consists of a perception and or linguistic judgement task
- takes less than a minute to perform
- has yes/no or multiple choice answers
- pays .20–.50USD == 10–20USD per hour (maybe more by now)

Amazon acts as a marketplace, managing the connection between task owners and 'turkers'

10. A real example

I used the Mechanical Turk to evaluate the results of a so-called "semantic search engine" competition in 2010.

The competition task was to provide semantic-web-sourced descriptions of people and places

- The evaluation task was to judge whether the results were in fact actually *about* the given subject

Query-result pairs packaged into batches of 12 HITs

Each HIT done by 3 workers (3 'assignments' per HIT)

- 10 real results
- 2 'fake' results: a known-good and a known-bad result

2 minutes time allotted and $0.20 per HIT

- On average, Turkers were done in 1 minute
- Netting out to $6-$12 an hour

Execution monitored online

- Time to complete
- Average performance on known-good and bad results vs. 'real' results

64 Turks in total, 4 bad apples

- Rejected assignments are not paid, redone

Total cost of (5786 / 10) * 0.20 * 3 + [admin] ≅ 400USD

11. Looking at measurements: what is significant?

So we're going to be measuring things

And comparing (distributions of) measurements

Each task we look at will have its own appropriate measurements

And thus each comparison will be in its own terms

But one issue will be present throughout: are the differences we find *significant*?

- For instance, lets look at character frequencies in the NLTK collection

```
>> from nltk.book import *
>>> f=FreqDist(x.lower() for x in [text1,text2,text3,text4,text5,text6,text7,tex
>>>  for y in x for x in y if ord(x)<128)
>>> f.items()[4:6]
 (u'n', 220713),
 (u'i', 218590),
>>> f['n']-f['i']
 2123
```

- But is the difference between 'n' and 'i' *significant*?

In general, if we can only *sample* from a distribution

- The differences we observe may not be *real*
- Or, to put it the other way around, they may be *accidental*

12. Another example

Over a period of 25 days, genders of newborns were tabulated at two hospitals

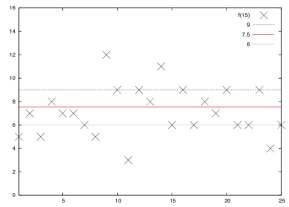
- In one hospital, 60% or more of the births were boys on 7 out of 25 days
- In the other hospital, 60% or more of the births were boys on only 2 out of 25 days

Is there something to worry about here?

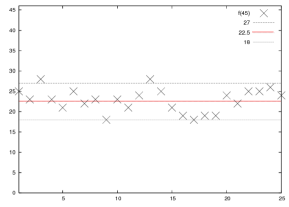
- In other words, is there a real difference?

Answer, in fact: 'No'

- Because the first hospital is much smaller than the second: 15 births a day as opposed to 45



Hospital 1: 15 births a day: number of boys per day



Hospital 2: 45 births a day: number of boys per day

Percentages can be the *wrong* basis for comparing outcomes if the population is of different sizes

13. How representative is the mean: Standard deviation

Significance measurement can be complex to understand, but the basic idea is simple (for **normal** distributions):

- Measure difference in terms of **standard deviation**

Standard deviation is essentially a measure of how representative the mean is

- The more outliers, and the further they are from the mean
  - The less representative the mean is
  - The standard deviation quantifies this

Definitions:

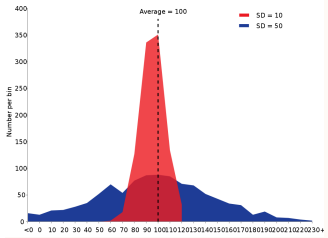
Mean of N measurements

$\frac{n_1 + n_2 + n_3 + \ldots + n_N}{N}$ Call this μ

Standard deviation of N measurements

$$\sqrt{\frac{(n_1 - \mu)^2 + (n_2 - \mu)^2 + (n_3 - \mu)^2 + \ldots + (n_N - \mu)^2}{N}}$$

Different standard deviation means different representativeness or reliability for the means



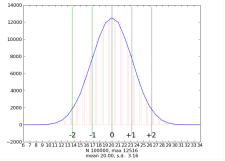
In the blue case, some items are a long way from the mean

- The mean is *less representative*

14. Normal distributions and standard deviation

- I tossed one coin 40 times: it came up heads 17 times.
  - Is it fair?
  - Probably, yes
- I toss a different coin, and it comes up heads only 13 times
  - Is it fair?
  - Probably, no

If we look at the distribution of outcomes over many coin-toss trials, it looks like this:



That's a classic **normal distribution**

The peak is at 50% heads, but there are lots of other plausible outcomes.

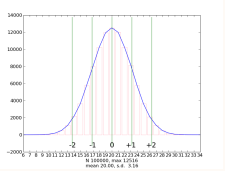
It looks like *most* of the results are within 2 standard deviations.

- In fact, about 96% of them

This is true by definition for any true normal distribution

15. Back to significance

Consider my earlier claim that 13 heads is probably a sign of an unfair coin



The graph tells us that 13 is outside the 2 standard deviation boundary

- And both our empirical observation (the graph) and the underlying maths tell us that flipping a *fair* coin 40 times will give a result *inside* the boundary about 96% of the time

So there's only about a 4% chance that a coin which gives 13 heads is fair

- That is, is drawn from a population whose measured distribution is normal with a mean of 20 and a standard deviation of 3.16

16. Reporting significance: 'p' values

"about 4%" doesn't seem a crisp way to report on an experiment

By convention, we say a result is significant if the chance is 5% or less

- What chance?
- The chance that the coin is fair after all, and we were just unlucky
- We're always reporting the chance that we're *wrong* (to conclude that the coin is bent)

So, 2 standard deviations is not quite the right boundary

- With more tosses per trial, we could see more detail
- And determine that the 95% point is 1.96 standard deviations

So a result outside the 1.96 boundary will come up once in 20 trials, even if the coin is fair.

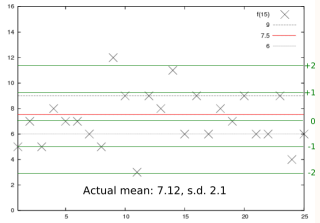
So we say that a result outside that boundary is significant "p < .05"

- That is, the probability is at most 1 in 20 that we are wrong to call such a coin bent
- Because a fair coin will only show such a result 1 time in 20

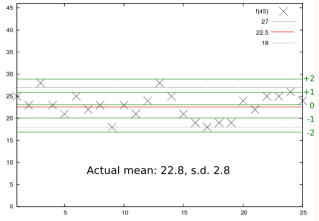
17. Back to the hospitals

We can look at the hospital data again

- Plotting SD bars this time



Hospital 1: 15 births a day: number of boys per day



Hospital 2: 45 births a day: number of boys per day

Now they don't look so different

18. Back to character frequency

Here's a tabulation of the top 6 character counts from the Project Gutenberg *Sense and Sensibility* (approx. 500,000 characters):

e 66604
t 44993
o 42015
a 40443
n 38439
i 36521

Is this ranking correct?

We can do an empirical version of the "p < .05" test

By looking at 20 samples of characters from a larger corpus (Reuters newswire)

't' vs. 'a'

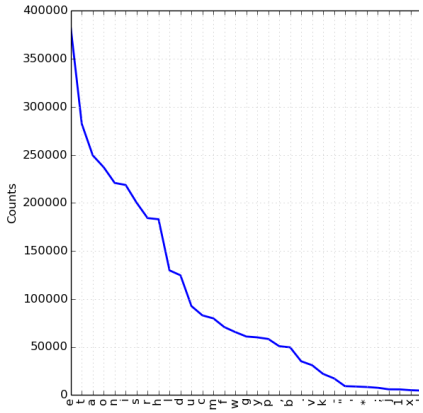
- |t|>|a| in 17 out of 20 samples of 10000 characters
- |t|>|a| in 19 out of 20 samples of 20000 characters
- |t|>|a| in 20 out of 20 samples of 40000 characters

'n' vs. 'i'

- |n|>|i| in 11 out of 20 samples of 10000 characters
- |n|>|i| in 14 out of 20 samples of 50000 characters
- |n|>|i| in 12 out of 20 samples of 150000 characters
- |n|>|i| in 15 out of 20 samples of 300000 characters
- |n|>|i| in 9 out of 10 samples of 600000 characters

So we can be pretty sure that in the underlying distribution 't' really is more frequent than 'a', but we really don't have a big enough sample to be sure about 'n' versus 'i'.

- Here's one of the big sample runs



It's also worth noting that the distribution for Austen and for the Reuters data is probably *not* the same. . .

- I *think* one measure of their agreement (Kruskal and Wallace's gamma) gives a rank correlation of 0.66 (p<.02)
- Many of the measures for comparing non-normal distributions convert the raw distribution to rank data
- J&M describe an alternative approach similar to the multiple sample approach we took above
  - Explain how to derive p-values from repeated sub-sampling

19. A common misunderstanding

What's wrong with this statement:

> There was no change in the control group's average blood pressure by the end of the trial. The intervention group showed a small improvement, but it was not statistically significant (p > .5)

The use of the word 'improvement' in the second sentence implies a particular direction of the change in average over the trial.

- But if the difference is not significant, then we don't *know* what direction the underlying change (if any) is!

The temptation is to say something like this

> Our theory predicted a speed-up in response time. Although the meaasured change was not significant, it *was* in the right direction.

Don't do that!

20. Lots of measures, lots of significance tests

Different kinds of measurements require different significance tests.

Broadly speaking, there are two classes of significance tests:

parametric When the underlying distribution is known to be normal, and the values are continuous, or at least proportionate
non-parametric Otherwise

Measures such as the t-test or z-test are the classic parametric measures of significance.

But for non-parametric distributions, for example many kinds of token frequency data, we can't use them

- Remember Zipf's Law!
- That is, many linguistic phenomena are *not* normally distributed.

21. Precision and recall one more time

Slides 29–34 of Lecture 7 introduced these two measures of classification success

- Another way to look at them

precision
Penalises false positives

- $\frac{\text{true positives}}{\text{true positives + false positives}}$
- System said *yes* when it should have said *no*

recall
Penalises false negatives

- $\frac{\text{true positives}}{\text{true positives + false negatives}}$
- System said *no* when it should have said *yes*

22. Precision and recall cont'd

Thinking still about binary classification, we can use a **contingency table** to help understand these

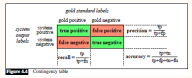


Figure 4.4 Contingency table

Figure 4.4 from https://web.stanford.edu/~jurafsky/slp3/4.pdf

Why do we need all these measures?

- It's easy to get high **recall** by always saying *yes*, but that gives low precision
  - Think of a setting the threshhold for *yes* really low
- It's easy to get high **precision** by always only saying *yes* when you're *really* sure, but that gives low recall
  - Think of a setting the threshhold for *yes* really high
- If there are very few instances, high **accuracy** is misleading, because it's easy to get lots of true negatives
  - Think of classifying for black swans: low recall and average precision would still show very high accuracy

23. Non-binary cases

Here's a *real* contingency table, in this case better named a **confusion matrix**:

	A	B	C	D	E	F	G	H	...
A	168	1	0	2	5	5	1	3	...
B	0	136	1	0	3	2	0	4	...
C	1	6	111	5	11	6	36	5	...
D	1	17	4	157	6	11	0	5	...
E	2	10	0	1	98	27	1	5	...
F	1	0	0	1	9	73	0	6	...
G	1	3	32	1	5	3	127	3	...
H	2	0	0	0	3	3	0	4	...
...	...	...	...	...	...	...	...	...	...

- Real experimental data, taken from http://obereed.net/lettersim/FisherMontyGlucksberg1969.html
- Read this as saying, e.g., "For stimulus a **C**, on 32 occasions a subject reported a **G**"
- Given there were 200 trials, we get an estimate of

$$P(\text{G} \mid \text{C}) = \frac{32}{200} = 0.16$$

We can still extract single-class precision and recall:

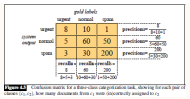


Figure 4.5 Confusion matrix for a three-class categorization task, showing for each pair of classes (c_1, c_2), how many documents from c_1 were (incorrectly) assigned to c_2

Figure 4.5 from https://web.stanford.edu/~jurafsky/slp3/4.pdf

24. Which is better, high precision or high recall?

Like so many other things we've looked at

- That depends on your application

Ever boarded an airplane, and the captain announces there's a technical problem? Ever wondered what's behind that?

- Apparently a high percentage of aircraft "technical problems" are *not* faults with the engines/wings/controls/...
- They are faults in the fault-detection system
- That is, false positives
- And that's a *good* thing
- Because the alternative is false negatives...
- That is, recall is *much* more important than precision
- So we set the threshhold for an alarm pretty low

Now consider hiring trainee pilots

- Suppose we have 20 applicants, and are only hiring 3
- It costs a lot to train a pilot
- So we really need good candidates
- It doesn't really hurt us if we can't identify all the good ones
- As long as the ones we identify *are* good
- That is, precision is much more important than recall
- So we set the threshhold pretty high

25. And finally, a relevant joke. . .

XKCD comic joke about 'statistically significant other' (Courtesy of XKCD)

Courtesy of XKCD

Another XKCD about significance. For this one you need the original, for the hovertext

> The truth about significance tests is, that p-values measure laboratory budgets -- Ron Kaplan