Notes on Assignment 3

Sharon Goldwater
6 November 2018

Overview of assignment

Exploration of distributional similarity

• Work with data extracted from Twitter (co-occurrence counts)
• Compare different ways to construct context vectors and compute similarities
• Analyze and discuss differences between approaches, qualitatively and quantitatively.

Remainder of the course

• Only two more lectures on specific technical topics (sentence semantics)
• Mostly focusing on broader picture: NLP in practice
  – Where does the data come from? Annotation, licensing, privacy
  – The messy world of data: user-generated text, multilingual NLP, biases
  – Issues in evaluation: reliability, human evaluation, ethics
• This assignment ties in with several of these: a step closer to real research/practice.

Dataset/Lab

• Lab 8 (next week) is a warm-up for the assignment. Feel free to start on it now if you want. Solutions will be posted after the lab sessions as usual.
• The dataset: word counts from 100 million tweets.
  – We have a license to use it for University purposes on University machines.
  – You may not download or redistribute it via your personal computer.
  – So yes, you will need to work on DICE (in person or remotely).
Qualitative and quantitative analysis

Assignment asks you to do some of each.

• Examples of qualitative analysis:
  – Using visualization to illustrate/discuss examples or trends
  – Discussing one or a few examples in more detail, by looking at
    our dataset and/or other Tweets (e.g., use the Twitter search
    page).

• Examples of quantitative analysis:
  – Often: numerical comparison to a gold standard of accuracy
  – Here: consider other options, such as correlating similarity
    measures against word frequency.

What is correlation?

• Intuitively: two random variables $X$ and $Y$ are correlated if,
  when the value of $X$ increases, the value of $Y$ also tends to
  increase (positive correlation) or decrease (negative correlation).

• Often, $X$ and $Y$ are different measurements for each data point.
  – A person’s height $X$ and weight $Y$
  – A word’s frequency $X$ and length $Y$

• Two standard ways to measure correlation:
  – Spearman (rank) correlation: roughly as above.
  – Pearson (linear) correlation: more specific.

One kind of quantitative analysis

• Assignment spec suggests you may want to consider correlation
  between similarity measures and word frequency.

• Why?
  – A good similarity measure should measure (only) similarity.
  – So presumably not be correlated with frequency.
  – Unless more frequent words really are more similar to each
    other! (Would need to test with humans... let’s assume not)

Pearson correlation

• Mathematically: the covariance of $X$ and $Y$, normalized by the
  product of their individual standard deviations.

• Intuitively: how close to a perfect linear relationship do $X$ and $Y$
  have?
  – Does not measure the slope of the line, just whether there is
    one.

• For data samples, the Pearson correlation coefficient is usually
  denoted $r$. 
Pearson correlation

Examples datasets with Pearson $r$ values shown:

1.0  0.8  0.4  0.0  -0.4 -0.8 -1.0
1.0  1.0  1.0 -1.0 -1.0 -1.0
0.0  0.0  0.0  0.0  0.0  0.0  0.0

Image source: https://commons.wikimedia.org/wiki/File:Correlation_examples.png

Spearman rank correlation

• Mathematically: compute the Pearson correlation between the rank ordering of $X$ and $Y$ values.

• Intuitively: how close to a perfectly monotonic relationship do $X$ and $Y$ have?

• For data samples, the Spearman rank correlation coefficient is usually denoted $\rho$ or $r_s$.

Spearman correlation

Data with perfect rank correlation, but not perfectly linear:

Image by Skbkekas (CC-BY-SA 3.0)
https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Which one to use?

• If correlation is roughly linear, Pearson will normally yield stronger results (larger absolute values)
  – If hypothesis testing against the possibility of no correlation, likely to have higher significance level than Spearman.
  – But if using large samples from corpora, often nearly any result is clearly “non-zero”. We may care more about the actual degree of correlation.

• If correlation is non-linear, or nothing is known, use Spearman.
But usually we do know something

Best to look at the data first! For example, word freq vs length:

Seems to follow a pattern, but not strongly linear. Indeed,

- Spearman: $\rho = -0.18$
- Pearson: $r = -0.10$

(Note: I “jittered” the data so those with same (x,y) are not right on top of each other.)

So, which one to use?

- So, Pearson can still work if there is an obvious transformation to make the correlation roughly linear.
- But if in doubt, usually fine to use Spearman.
- As with all statistics, many subtleties if using for really careful analysis (see statistics course or online tutorials), but what I’ve said is probably enough for exploratory studies (i.e., your assignment).

Log frequency

Of course, using log frequencies is often more sensible:

We now have

- Spearman: $\rho = -0.18$
- Pearson: $r = -0.21$

Notice that $\rho$ is not affected by rescaling the data. $r$ is higher, but still only a weak linear correlation.

Announcements

- You should have received an email about your assignment partner, if not please contact Maria immediately.
- Thursday’s lecture: first of several on “real-world” issues (data, ethics, evaluation).
  - These lectures mostly not from the textbook; less specific technical content for you to know, but you are expected to be familiar with the issues discussed.