Notes on Assignment 3

Sharon Goldwater

14 November 2016
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

• Discuss pros and cons of each approach, qualitatively and quantitatively.
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)
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.
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
0.0   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.)
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.
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).
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).
Announcements

• You should have received an email about your assignment partner, if not please contact Henry immediately.

• There will be someone filming tomorrow’s lecture for a promotional video.