AI and Social Science – Brendan O'Connor
cognition, language, social systems;
statistics, visualization, computation

Cosine similarity, Pearson correlation, and OLS coefficients

Posted on March 13, 2012

Cosine similarity, Pearson correlations, and OLS coefficients can all be viewed as vari
tweaked in different ways for centering and magnitude (i.e. location and scale, or som

Details:

You have two vectors x and y and want to measure similarity between them. A basic s
inner product

$$Inner(x, y) = \sum_i x_i y_i = \langle x, y \rangle$$

If x tends to be high where y is also high, and low where y is low, the inner product wi
more similar.

The inner product is unbounded. One way to make it bounded between -1 and 1 is to d
norms, giving the **cosine similarity**

$$CosSim(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\langle x, y \rangle}{||x||\ ||y||}$$

This is actually bounded between 0 and 1 if x and y are non-negative. Cosine similarit
cosine of the angle between the two vectors; you can illustrate this for vectors in $\mathbb{R}^2$ (e

Cosine similarity is not invariant to shifts. If x was shifted to x+1, the cosine similarity
invariant, though, is the **Pearson correlation**. Let $\bar{x}$ and $\bar{y}$ be the respective means

$$\begin{aligned} Corr(x,y) &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \\ &= \frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{||x - \bar{x}||\ ||y - \bar{y}||} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

Correlation is the cosine similarity between centered versions of x and y, again bound
usually talk about cosine similarity in terms of vector angles, but it can be loosely thou
you think of the vectors as paired samples. Unlike the cosine, the correlation is invari
changes of x and y.

This isn't the usual way to derive the Pearson correlation; usually it's presented as a n
covariance, which is a centered average inner product (no normalization)

$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{n}$$

Finally, these are all related to the coefficient in a **one-variable linear regression**
with Gaussian noise, whose MLE is the least-squares problem $\arg\min_a \sum (y_i - a x_i)^2$,
a is

$$OLSCoef(x,y) = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\langle x, y \rangle}{||x||^2}$$

This looks like another normalized inner product. But unlike cosine similarity, we are
— instead we only use x's norm (and use it twice): denominator of $||x||\ ||y||$ versus $||x$

Not normalizing for y is what you want for the linear regression: if y was stretched to
would need to increase a to match, to get your predictions spread out too.

Often it's desirable to do the OLS model with an intercept term: $\min_{a,b} \sum (y - a x_i - b)^2$

$$\begin{aligned} OLSCoefWithIntercept(x,y) &= \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\langle x - \bar{x},}{||x - \bar{x}} \\ &= OLSCoef(x - \bar{x}, y) \end{aligned}$$

It's different because the intercept term picks up the slack associated with where x's c

OLSCoefWithIntercept is invariant to shifts of x. It's still different than cosine similar normalizing at all for y. Though, subtly, it does actually control for shifts of y. This isn but with a little arithmetic it's easy to derive that $\langle x - \bar{x},\ y \rangle = \langle x - \bar{x},\ y + c \rangle$ for any nice geometric interpretation of this.)

Finally, what if x and y are standardized: both centered and normalized to unit standa coefficient for that is the same as the Pearson correlation between the original vectors means or if it's a useful fact, but:

$$OLSCoef\left(\sqrt{n}\frac{x-\bar{x}}{||x-\bar{x}||}, \sqrt{n}\frac{y-\bar{y}}{||y-\bar{y}||}\right) = Corr(x,y)$$

Summarizing: Cosine similarity is normalized inner product. Pearson correlation is ce one-variable OLS coefficient is like cosine but with one-sided normalization. With an

Of course we need a summary table. "Symmetric" means, if you swap the inputs, do y "Invariant to shift in input" means, if you add an arbitrary constant to either input, do

Function	Equation	Symmetric?	Output range	Invariant to shift in input
Inner(x,y)	$\langle x, y \rangle$	Yes	$\mathbb{R}$	No
CosSim(x,y)	$\frac{\langle x, y \rangle}{\|\|x\|\| \ \|\|y\|\|}$	Yes	[-1,1] or [0,1] if inputs non-neg	No
Corr(x,y)	$\frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{\|\|x - \bar{x}\|\| \ \|\|y - \bar{y}\|\|}$	Yes	[-1,1]	Yes
Cov(x,y)	$\frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{n}$	Yes	$\mathbb{R}$	Yes

OLSCoefNoIntcpt(x,y)	$\frac{\langle x, y \rangle}{\|\|x\|\|^2}$	No	$\mathbb{R}$	No
OLSCoefWithIntcpt(x,y)	$\frac{\langle x - \bar{x},\ y \rangle}{\|\|x - \bar{x}\|\|^2}$	No	$\mathbb{R}$	Yes

Are there any implications? I've been wondering for a while why cosine similarity ten language processing applications. Maybe this has something to do with it. Or not. On product stuff is computational strategies to make it faster when there's high-dimensic Friedman et al. 2010 glmnet paper talks about this in the context of coordinate desce Dhillon et al., NIPS 2011 applies LSH in a similar setting (but haven't read it yet). An LSH for cosine similarity; e.g. van Durme and Lall 2010 [slides].

Any other cool identities? Any corrections to the above?

References: I use Hastie et al 2009, chapter 3 to look up linear regression, but it's cov places. I linked to a nice chapter in Tufte's little 1974 book that he wrote before he we visualization stuff. (He calls it "two-variable regression", but I think "one-variable reg "one-feature" or "one-covariate" might be most accurate.) In my experience, cosine si often in text processing or machine learning contexts.

This entry was posted in Uncategorized. Bookmark the permalink.

23 Responses to *Cosine similarity, Pearson correlation, and OLS coefficie*

VC

Victor Chahuneau *says:*
March 15, 2012 at 1:21 am

I think your OLSCoefWithIntercept is wrong unless y is centered: the right part of the dot product should
Then the invariance by translation is obvious...
Otherwise you would get <x-, y+c> = <x-,y> + c(n-1)
See Wikipedia for the equation

VC

Victor Chahuneau *says:*

March 15, 2012 at 1:24 am

... but of course WordPress doesn't like my brackets...
Line 1:$(y-\bar y)$
Line 3: $ = + c(n-1)\bar x$

brendano *says:*
March 15, 2012 at 5:05 am

Nope, you don't need to center y if you're centering x. The Wikipedia equation isn't as correct as Hastie :) was writing the post, but if you write out the arithmetic like I said you can derive it.

Example:

$ R
> x=c(1,2,3); y=c(5,6,10)
> inner_and_xnorm=function(x,y) sum(x*y) / sum(x**2)
> inner_and_xnorm(x-mean(x),y)
[1] 2.5
> inner_and_xnorm(x-mean(x),y+5)
[1] 2.5

... if you don't center x, then shifting y matters.

VC

Victor Chahuneau *says:*
March 15, 2012 at 4:15 pm

Oops... I was wrong about the invariance!
It turns out that we were both right on the formula for the coefficient... thanks to this same invariance.
Here is the full derivation:
http://dl.dropbox.com/u/2803234/ols.pdf

Wikipedia & Hastie can be reconciled now...

Mike *says:*
March 26, 2012 at 8:40 am

Nice breakdown Brendan.

I've been working recently with high-dimensional sparse data. The covariance/correlation matrices can b

after rearranging some terms.

http://stackoverflow.com/a/9626089/1257542

Mike *says:*
March 26, 2012 at 12:17 pm

for instance, with two sparse vectors, you can get the correlation and covariance without subtra

cov(x,y) = (inner(x,y) – n mean(x) mean(y)) / (n-1)
cor(x,y) = (inner(x,y) – n mean(x) mean(y)) / (sd(x) sd(y) (n-1))

Brendan O'Connor *says:*
March 26, 2012 at 1:18 pm

Oh awesome, thanks!

Kat *says:*
April 24, 2012 at 11:12 pm

Hey Brendan! Maybe you are the right person to ask this to – if I want to figure out how similar two sets magnitude) how would I do that? I originally started by looking at cosine similarity (well, I started them was correlation?) but of course that doesn't look at magnitude at all. Is there a way that people usually we that arbitrary?

Brendan O'Connor *says:*
April 24, 2012 at 11:25 pm

Why not inner product?

Kat *says:*
April 24, 2012 at 11:43 pm

I would like and to be more similar than and , for example

Kat *says:*
April 24, 2012 at 11:44 pm

ok no tags this time – 1,1 and 1,1 to be more similar than 1,1 and 5,5

Pingback: *Triangle problem – finding height with given area and angles. « Math World – etidhor*

Adam *says:*
February 1, 2013 at 5:57 pm

This is one of the best technical summary blog posts that I can remember seeing. I've just started in NLP appear as the de facto relatedness measure—this really helped me mentally reconcile it with the alternati

Paul Moore *says:*
March 18, 2013 at 5:14 pm

A very helpful discussion – thanks.

Have you seen – 'Thirteen Ways to Look at the Correlation Coefficient' by Joseph Lee Rodgers; W. Alan N Statistician, Vol. 42, No. 1. (Feb., 1988), pp. 59-66. It covers a related discussion.

brendano *says:*
March 18, 2013 at 5:17 pm

Great tip — I remember seeing that once but totally forgot about it.

Here's a link, http://data.psych.udel.edu/laurenceau/PSYC861Regression%20Spring%202012/READIN ways.pdf

Pingback: *Correlation picture | AI and Social Science – Brendan O'Connor*

Peter *says:*
March 29, 2013 at 3:24 am

Useful info:

I have a few questions (i am pretty new to that field). You say correlation is invariant of shifts.

i guess you just mean if the x-axis is not 1 2 3 4 but 10 20 30 or 30 20 10.. then it doesn't change anythin

but you doesn't mean that if i shift the signal i will get the same correlation right?

ex: [1 2 1 2 1] and [1 2 1 2 1], corr = 1

but if i cyclically shift [1 2 1 2 1] and [2 1 2 1 2], corr = -1
or if i just shift by padding zeros [1 2 1 2 1 0] and [0 1 2 1 2 1] then corr = -0.0588

Please elaborate on that.

Also could we say that distance correlation (1-correlation) can be considered as norm_1 or norm_2 dista want to minimize the squared errors, usually we need to use euclidean distance, but could pearson's corr

Ans last, OLSCoef(x,y) can be considered as scale invariant? is very correlated to cosine similarity which correlation is right?). Look at: "Patterns of Temporal Variation in Online Media" and "Fast time-series se That confuses me.. but maybe i am missing something.

Brendan O'Connor *says:*
April 1, 2013 at 10:27 pm

Hi Peter –

By "invariant to shift in input", I mean, if you *add* to the input. That is,
f(x, y) = f(x+a, y) for any scalar 'a'.

By "scale invariant", I mean, if you *multiply* the input by something.

For (1-corr), the problem is negative correlations. I think maximizing the squared correlation is error .. that's why it's called R^2, the explained variance ratio.

I don't understand your question about OLSCoef and have not seen the papers you're talking ab

Pingback: *Machine learning literary genres from 19th century seafaring, horror and western novels | Sub-Sub Algorithm*

Pingback: *Machine learning literary genres from 19th century seafaring, horror and western novels | Sub-Subroutine*

Waylon Flinn *says:*
December 11, 2013 at 4:51 am

Wonderful post. The more I investigate it the more it looks like every relatedness measure around is just product.
Similar analyses reveal that Lift, Jaccard Index and even the standard Euclidean metric can be viewed as product. It's not a viewpoint I've seen a lot of. It was this post that started my investigation of this pheno

The fact that the basic dot product can be seen to underlie all these similarity measures turns out to be cc in your space on top of each other to create a matrix, you can produce all the inner products simply by m Furthermore, the extra ingredient in every similarity measure I've looked at so far involves the magnitud individual vectors. These drop out of this matrix multiplication as well. Just extract the diagonal.

Because of it's exceptional utility, I've dubbed the symmetric matrix that results from this product the ba
able to find many other references which formulate these metrics in terms of this matrix, or the inner pro
mathematics is both broad and deep, so it seems likely that I'm stumbling upon something that's already

Do you know of other work that explores this underlying structure of similarity measures? Is the constru
standard technique in the calculation of these measures? Does it have a common name?

Thanks again for sharing your explorations of this topic.

P.S. Here's the other reference I've found that does similar work:
http://arxiv.org/pdf/1308.3740.pdf

Pingback: *Building the connection between cosine similarity and correlation in R | Question and Answer*

Pingback: 相似性度量 - *CSer*之声

AI and Social Science – Brendan O'Connor
Proudly powered by WordPress.