MATH V1201: CORRELATION

ROBERT LIPSHITZ

Suppose I have some data; for example:

Student	Midterm score	Final exam score
A	56	100
B	42	74
C	63	97
D	39	59
E	52	72

I want to know if these data are correlated—is a student's midterm exam score a good predictor of his or her final exam score? We will compute a number, called the correlation, which (partly) answers this question. The correlation has the following properties:

- The correlation is always between -1 and 1.
- Values close to 1 indicate that the data are *positively correlated.* In our example, this would mean that if a student does well on the midterm then he or she is likely to do well on the final exam.
- Values close to 0 indicate that the data are *uncorrelated.* In our example, this would mean the midterm score doesn't predict the final exam score.
- Values close to -1 indicate the data are *negatively correlated.* In our example, this would mean that if a student does well on the midterm then he or she is likely to do poorly on the final exam.

To define the correlation, view the two lists of data as vectors:

$$v = \langle 56, 42, 63, 39, 52\rangle \qquad \text{and} \qquad w = \langle 100, 74, 97, 59, 72\rangle.$$

Roughly, the correlation is the angle θ between v and w. This makes sense as a definition: if θ is small then v and w point in roughly the same direction, so v is a good predictor for w. To get the actual definition of correlation, we make two adjustments:

Adjustment 1. To get a number between -1 and 1, take $\cos(\theta)$ instead of θ itself. Notice that $\cos(\theta) \sim 1$ when $\theta \sim 0$, i.e., when v and w point in roughly the same direction, and $\cos(\theta) \sim 0$ when $\theta \sim \pi/2$, i.e., when v and w are roughly orthogonal.[1] Finally, $\cos(\theta) \sim -1$ when $\theta \sim \pi$, i.e., when v and w point in roughly opposite directions.

Adjustment 2. Before computing $\cos(\theta)$ we need to recenter v and w so that they each have mean (average) 0. That is, let a be the mean of the entries of w, and b the

Date: January 28, 2015.

[1] Strange fact: randomly chosen vectors with lots of components tend to be almost orthogonal. Another way of saying this is that in big-dimensional spheres, most of the volume is concentrated near the equator.

mean of the entries of v. Instead of computing the angle between v and w, compute the (cosine of the) angle between

$$v' = v - \langle a, a, \dots, a \rangle \qquad \text{and} \qquad w' = w - \langle b, b, \dots, b \rangle.$$

In our example, the averages are

$$\begin{aligned} a &= \frac{1}{5}(56 + 42 + 63 + 39 + 52) = \frac{262}{5} = 50.4 \\ b &= \frac{1}{5}(100 + 74 + 97 + 59 + 72) = \frac{402}{5} = 80.4. \end{aligned}$$

So,

$$\begin{aligned} v' &= \langle 56 - 50.4, 42 - 50.4, 63 - 50.4, 39 - 50.4, 52 - 50.4 \rangle \\ &= \langle 5.6, -8.4, 12.6, -11.4, 1.6 \rangle \\ w' &= \langle 100 - 80.4, 74 - 80.4, 97 - 80.4, 59 - 80.4, 72 - 80.4 \rangle \\ &= \langle 19.6, -6.4, 16.6, -21.4, -8.4 \rangle. \end{aligned}$$

To see why we have to make adjustment 2, think about the exam scores. Exam scores are (almost) always positive. So, the angle between v and w is always going to be less than $\pi/2$ (think about vectors in the first quadrant). Without making adjustment 2, we will always see a positive correlation—which doesn't make sense.

(Another way of thinking about adjustment 2 is that v and w each have a systematic bias—their means. We are subtracting off the biases before computing the correlation.)

Finally, how to compute $\cos(\theta)$? The answer, of course, is the dot product:

$$\cos(\theta) = \frac{(v') \cdot (w')}{\|v'\| \|w'\|}.$$

In our example:

$$\begin{aligned} v' \cdot w' &= (5.6)(19.6) + (-8.4)(-6.4) + (12.6)(16.6) + (-11.4)(-21.4) + (1.6)(-8.4) \\ &= 603.2 \\ \|v'\| &= \sqrt{(5.6)^2 + (-8.4)^2 + (12.6)^2 + (-11.4)^2 + (1.6)^2} \simeq 19.8 \\ \|w'\| &= \sqrt{(19.6)^2 + (-6.4)^2 + (16.6)^2 + (-21.4)^2 + (-8.4)^2} \simeq 35.1 \end{aligned}$$

$$\text{correlation} = \frac{603.2}{(19.8)(35.1)} = 0.87.$$

So, in our example, the scores are fairly strongly correlated. (By the way, these are actual exam scores.)

This definition of correlation has a few other nice (if obvious) properties:

- If you add the same number to all of the entries of v (or w) then the correlation doesn't change.
- It is *scale invariant*: if you multiply all of the entries of v by the same number then the correlation doesn't change.
- It is *symmetric*: the correlation of v and w is the same as the correlation of w and v.

Summary: computing correlation.

Given vectors v and w (of data), to compute the correlation between v and w:

(1) Compute the average a of the entries of v and the average b of the entries of w.

(2) Compute the *recentered vectors* $v' = v - \langle a, a, \dots, a\rangle$ and $w' = w - \langle b, b, \dots, b\rangle$.

(3) The correlation is then given by

$$\text{correlation}(v, w) = \frac{(v')\cdot(w')}{\|v'\|\|w'\|}.$$

(In practice, of course, one uses a computer. In Excel, the function for correlation is "correl". It does exactly what we said; try it.)

Practice problems.

(1) Compute the correlation between $\langle 1, 3\rangle$ and $\langle 4, 0\rangle$.

(2) Compute the correlation between $\langle 1, 3, 2\rangle$ and $\langle 0, 2, 2\rangle$.

(3) Does it make sense to compute the correlation between $\langle 1, 2, 3\rangle$ and $\langle 3, 4, 1, 2\rangle$?

(4) Let $v = \langle 1, 2, 3\rangle$.

   (a) Find a vector which is positively correlated with v.

   (b) Find a vector which is negatively correlated with v.

   (c) Find a vector which is perfectly uncorrelated with v (i.e., where the correlation is 0).

E-mail address: `lipshitz@math.columbia.edu`