Info 2950, Lecture 22
24 Apr 2018

Prob Set 6: due 24 Apr 2018
(see notes on Piazza page re minor edits)

tentative due dates for the rest:
ps7 due Wed 2 May
ps8 due Wed 9 May (last day classes, auto-extension to end week)
Linear Regression (Least Square Fit)

The best line $y=ax+b$ is determined by parameters $a,b$ that minimize the sum

$$\sum_{i=1}^{n} ((ax_i + b) - y_i)^2$$

over the $n$ data points $(x_i, y_i)$

The solution is

$$b = E[y] - aE[x]$$

$$a = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$$
see also http://www.tylervigen.com/spurious-correlations
Correlation coefficient only measures association. Correlation does not imply causation.

Though the correlation between the weight and the math ability of children in a school district may be positive, that does not mean that doing math makes children heavier or that putting on weight improves the children's math skills.

Age is a confounding variable: older children are both heavier and better at math than younger children, on average.
\[ \text{Cov}(X, Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \]

\[ \text{Cov}(X, X) = \text{Var}(X) \]

\[ X > \mathbb{E}[X], \ Y > \mathbb{E}[Y] \]

\[ X < \mathbb{E}[X], \ Y > \mathbb{E}[Y] \]

\[ \text{Cov}(X, Y) > 0 \]

\[ X > \mathbb{E}[X], \ Y < \mathbb{E}[Y] \]

\[ \text{Cov}(X, Y) < 0 \]
Pearson correlation coefficient

\[ \rho(x,y) = \frac{\text{Cov}(x,y)}{\sigma(x) \sigma(y)} = \frac{E[(x - E[x])(y - E[y])]}{\sigma[x]\sigma[y]} \]
Recall that if we rescale the data, $x \rightarrow x/c$, then that divides the standard deviation $\sigma[x]$ by the same $c$.

In particular, we can divide by $c = \sigma[x]$, which gives a new distribution with standard deviation normalized to 1 (as is done to calculate $z$ values).

Note that $\text{Pearsonr}(x,y) = \frac{\text{Cov}[x,y]}{\sigma[x] \sigma[y]}$ is unchanged by rescaling both $x$ and $y$.

On the other hand, if we rescale $x \rightarrow x/\sigma[x]$ and $y \rightarrow y/\sigma[y]$ , then the linear regression slope $\frac{\text{Cov}[x,y]}{\text{Var}[x]} \rightarrow \text{Cov}[x,y]$.

Therefore the $\text{Pearsonr}(x,y)$ is just the linear regression slope when the variables are rescaled to have standard deviation equal to 1.
[15, 10, 10, 10, 10, 10, 15]
1) naive bayes, 2) random exam, 3) Poisson, 4) coke/pepsi, 5) i**3, 6) poll, 7) ‘the’
[15, 10, 10, 10, 10, 10, 15]
1) naive bayes, 2) random exam, 3) Poisson, 4) coke/pepsi, 5) i**3, 6) poll, 7) ‘the’
[15, 10, 10, 10, 10, 10, 15]
1) naive bayes, 2) random exam, 3) Poisson, 4) coke/pepsi, 5) i**3, 6) poll, 7) ‘the’
1) naive bayes, 2) random exam, 3) Poisson, 4) coke/pepsi, 5) $i^3$, 6) poll, 7) ‘the’
students in info2950 spr'17

- 1110
- 2110

midterm grade ($\rho = 5$, $p = 3 \times 10^{-7}$)

grade in (later of) 1110/2110

[Graph showing a scatter plot with different grades for two classes, 1110 and 2110.]
Spearman correlation=1
Pearson correlation=0.88

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
The Pearson correlation coefficient misses non-linear relationships and is also sensitive to outliers — the Spearman correlation can sometimes find correlations that Pearson misses.

It is defined as the Pearson correlation of the rank order of the data. That means it also varies from $-1$ (perfectly anti-correlated) to $+1$ (perfectly correlated), with 0 meaning uncorrelated.

If the data has $x = [.6, .4, .2, .1, .5]$ then the ranks are $r = [5, 3, 2, 1, 4]$.

For data $y = [403, 54, 7, 2, 148]$, the ranks $s = [5, 3, 2, 1, 4]$ are the same.*

so the Spearman correlation is 1, whereas the Pearson is less than one. Both functions are available in scipy.stats (as pearsonr() and spearmanr()).

[*Actually the second was generated from the first by taking the integer part of exp(10x)*]
roughly elliptically distributed and there are no prominent outliers: same

less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples (limited to value of rank)

https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
Defined as the Pearson correlation for the ranks, the Spearman correlation is written

$$\rho = \frac{\text{Cov}[r, s]}{\sigma[r] \sigma[s]}$$, \hspace{1cm} (1)

where \( \text{Cov}[r, s] = E[(r - E[r])(s - E[s])] \)

(generalizing the \( \text{Var}[x] = E[(x - E[x])^2] \), with \( \text{Cov}[x, x] = \text{Var}[x] \)).

The formula for the Spearman correlation coefficient is given at http://en.wikipedia.org/wiki/Spearman’s_rank_correlation_coefficient in terms of the difference \( d_i = r_i - s_i \) between ranks, in this easily calculable form:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$. \hspace{1cm} (2)

It is straightforward to verify that (1) reduces to (2) (see linked notes)
\begin{align*}
\text{pearsonr}(xdata, ydata) &= 0.91 \\
\text{spearmanr}(xdata, ydata) &= 1 \\
\text{pearsonr}([1, 2, 3], [1, 2, 3]) &= 1 \\
\text{var}([1, 2, 3]) &= \frac{2}{3} \\
\text{cov}([1, 2, 3], [1, 3, 2]) &= \frac{((1-2)(1-2) + (2-2)(3-2) + (3-2)(2-2))}{3} = \frac{1}{3}
\end{align*}