Research School of Computer Science, Australian National University

COMP3420, Advanced Databases and Data Mining

Tutorial 3

1. Calculating data correlations – Pearson + χ^2

   Both the Pearson's correlation coefficient and the χ^2 test statistic are used to determine whether two groups of data are (positively or negatively) correlated.

   (a) There are two definitions of the standard deviation, one is called *standard deviation of the sample*, the other is called *sample standard deviation*. What are the differences between them?

   (b) What kind of data are Pearson's correlation and χ^2 value applied to, respectively?

   (c) What are the definitions of Pearson's correlation and χ^2 value?

   (d) Look at the Adult data set (link on wattle), and take the first five rows. If you're using a spreadsheet, import the csv file and look at the first 5 data rows. If in **Rattle**, on the **Data** page you can look at the data set using the **View** button.

   Taking the first 5 rows in this data set, calculate the Pearson's product moment between attributes *age* and *hours-per-week*.

   (e) In the same data set, calculate the χ^2 value between attributes *sex* and *income* for the first ten rows.

   Answer:

   (a) For general random variable X, the standard deviation is defined as

   $$\sigma = \sqrt{\mathbb{E}\left[(X-\mu)^2\right]} = \sqrt{\int_{\mathbb{R}} (X-\mu)^2 p(X) \mathrm{d}X} \ ,$$

   where $\mu = \mathbb{E}[X] = \int_{\mathbb{R}} X p(X) \mathrm{d}X$ is the mean, $p(X)$ is the probability distribution for X. In practice, we usually use N samples from the distribution to approximate the standard deviation, so the definition of *standard deviation of the sample* is given by

   $$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

   while that of *sample standard deviation* is

   $$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \ .$$

The difference between these two versions is that s^2 is an unbiased estimator of σ^2, while s_N^2 is biased. This means that

$$\mathbb{E}\left[s^2 - \sigma^2\right] = 0, \text{ but } \mathbb{E}\left[s_N^2 - \sigma^2\right] \neq 0.$$

However, both s and s_N are biased estimators of σ:

$$\mathbb{E}\left[s - \sigma\right] \neq 0, \text{ and } \mathbb{E}\left[s_N - \sigma\right] \neq 0$$

In fact, it turns out that there is no single unbiased estimator of σ that works across all distributions.

(b) Pearson's product moment deals with numeric data, whereas χ^2 value deals with nominal data.

(c) For Pearson's product moment,

$$r_{A.B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

where n is the number of tuples, a_i and b_i are the respective values of attributes A and B in tuple i, $\bar{A}$ and $\bar{B}$ are the respective mean values of A and B, σ_A and σ_B are the respective standard deviations of A and B.

For χ^2 value, suppose attribute A has c distinct values $a_1, \cdots, a_c$, attribute B has r distinct values $b_1, \cdots, b_r$, then

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where o_{ij} is the *observed frequency* (*i.e.*, actual count) of the joint event (a_i, b_j) and e_{ij} is the *expected frequency* of (a_i, b_j), which can be computed as $e_{ij} = \frac{\text{count}(A=a_i)\times\text{count}(B=b_j)}{n}$.

(d) Using the formula from part c, we can compute the Pearson product moment between *age* and *hours-per-week*:

$$\bar{A} = \frac{39 + 50 + 38 + 53 + 28}{5} = 41.6,$$

$$\bar{B} = \frac{40 + 13 + 40 + 40 + 40}{5} = 34.6,$$

$$\sigma_A = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{A})^2} = 9.002,$$

$$\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{B})^2} = 10.8,$$

$$r_{A,B} = \frac{(39 \times 40) + (50 \times 13) + (38 \times 40) + (53 \times 40) + (28 \times 40) - 5 \times 41.6 \times 34.6}{5 \times 9.002 \times 10.8}$$
$$= -0.466 \ .$$

(e) Let $a_1 =$ 'Male', $a_2 =$ 'Female', $b_1 =$ '$\leq 50K$', $b_2 =$ '$> 50K$', then

$$o_{11} = 4, o_{12} = 2, o_{21} = 3, o_{22} = 1,$$
$$e_{11} = 6 \times 7/10 = 4.2, e_{12} = 6 \times 3/10 = 1.8,$$
$$e_{21} = 4 \times 7/10 = 2.8, e_{22} = 4 \times 3/10 = 1.2,$$
$$\chi^2 = \tfrac{(o_{11}-e_{11})^2}{e_{11}} + \tfrac{(o_{12}-e_{12})^2}{e_{12}} + \tfrac{(o_{21}-e_{21})^2}{e_{21}} + \tfrac{(o_{22}-e_{22})^2}{e_{22}} = 0.0794 \ .$$

For a 2 by 2 contingency table, the degrees of freedom are $(2-1)(2-1) = 1$. The cut-off point for a χ^2 distribution with 1 degree of freedom at the 5% significance level is 3.841. Since our χ^2 test statistic is much less than 3.841, there is ***not*** enough evidence to reject the null hypothesis that `gender` and `income` are independent.

2. **Independence and Correlation** Two random variables X and Y are independent if the following formula holds:

$$p(X, Y) = p(X) \times p(Y) \ ,$$

where $p(a)$ represents the distribution function of the random variable a. Two random variables X and Y are uncorrelated if $\text{cov}(X, Y) = 0$.

(a) Motivating example: For a city, there seems to be more number of hospitals when there are more car accidents incidences. Think, for instance, Melbourne vs Canberra. Is the quantity `#hospitals` *positively or negatively correlated* with `#accidents`? Does this mean that `#accidents` would *cause* `#hospitals` to increase, or vice versa?

(b) Let X be uniformly distributed on the interval $[-1, 1]$. If $X \leq 0$, then $Y = -X$, while if X is positive, then $Y = X$. What would be E(X), E(Y), E(XY), and cov(X, Y)? Can you try plotting the joint distribution P(X, Y), how does it look like ?

(c) (*) Show that if two random variables X and Y are independent, then they are uncorrelated.

(d) (*) Show that if the two normally distributed variables are uncorrelated, then they are independent. *i.e.*, we know x and y follow normal distribution

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad p(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

and that $\text{cov}(X, Y) = 0$, show that it leads to $p(x, y) = p(x)p(y)$.

(e) (*) Name another distribution such that uncorrelation implies independence, *i.e.*, $\text{cov}(X, Y) = 0 \Rightarrow X, Y$ are independent.

Answer:

(a) Correlation does not imply causation. The increase of `#accidents` and `#hospitals` could both be caused by some other factor, such as the population in the city.

(b) We have: $p(x) = 1/2$, for $-1 \leq x \leq 1$.

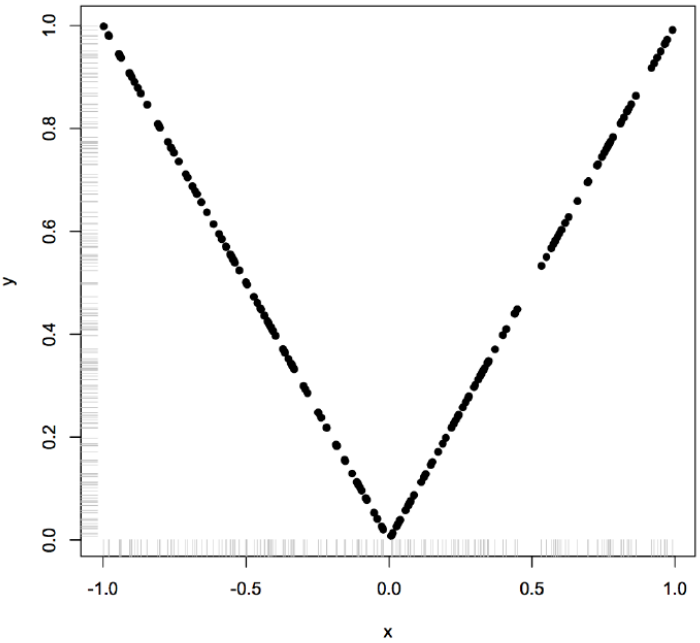
i. $\mathbb{E}[X] = \int_{-1}^{1} x\, p(x) dx = 0$
when $X \leq 0, Y = -X$, $p(y|x \leq 0) = 1, 0 \leq y \leq 1$,
$\mathbb{E}[Y|X \leq 0] = \int_{-1}^{0} -x dx = 1/2$;
when $X > 0, Y = X$, $p(y|x > 0) = 1, 0 \leq y \leq 1$,
$\mathbb{E}[Y|X > 0] = \int_{0}^{1} x dx = 1/2$;
$\mathbb{E}[Y] = \mathbb{E}[Y|X > 0]P(X > 0) + \mathbb{E}[Y|X \leq 0]P(X \leq 0) = 1/2$

ii. $$\begin{cases} X \leq 0 & \mathbb{E}[XY|X \leq 0] = \int_{-1}^{0} -x^2 dx = -1/3 \\ X > 0 & \mathbb{E}[XY|X > 0] = \int_{0}^{1} x^2 dx = 1/3 \end{cases} \Rightarrow \mathbb{E}[XY] = 0$$

iii. $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \times 1/2 = 0$

The joint distribution of X and Y is not uniform on the rectangle $[1, 1] \times [0, 1]$, as it would be if X and Y were independent (Figure 1).

Figure 1:



(c) We know from definition of covariance that $\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, hence,

$$\text{cov}(X, Y) = \int\int XYp(X, Y)dYdX - \int Xp(X)dX \int Yp(Y)dY \tag{1}$$

Since from the definition of independence $p(X, Y) = p(X) \times p(Y)$, then,

$$\begin{aligned} \text{cov}(X, Y) &= \int\int XYp(X)p(Y)dYdX - \int Xp(X)dX \int Yp(Y)dY \\ &= \int\int Xp(X)Yp(Y)dYdX - \int Xp(X)dX \int Yp(Y)dY \\ &= \int Xp(X)dX \int Yp(Y)dY - \int Xp(X)dX \int Yp(Y)dY \\ &= 0\ . \end{aligned} \tag{2}$$

Therefore these two variables are uncorrelated.

(d) To compute the joint probability we need to first build the covariance matrix which is equal to identity matrix in this case (because the variance of each variable is one and their covariance is zero). Therefore, the joint probability of X and Y can be written as[1]:

$$p(X, Y) = \frac{1}{(2\pi)^{2/2}} \exp(-\frac{1}{2} \begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix}) \tag{3}$$

computing this product yields,

$$\begin{aligned} p(X, Y) &= \frac{1}{2\pi} \exp(-\frac{x^2 + y^2}{2}) \\ &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \times \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2}) = p(X)p(Y) \end{aligned} \tag{4}$$

(e) For example binomial distribution.

3. **Practical question**: Analyzing stock prices using *Google Spreadsheet*. Historical stock data obtained from www.dailyfinance.com.

First download data from http://bit.ly/comp3420-tut2. Then open docs.google.com in a browser window and create a blank spreadsheet. Copy-paste the data into the spreadsheet, it should have four columns: (A) Data – all 2011 trading days before March 16. (B-D) The closing price on each day for Google (GOOGLE) Apple(APPL) and Cisco (CSCO).

[1] The determinant of the covariance matrix here is 1 thus omitted in the denominator.

Note that this exercise is designed to provide some hands-on experience in data cleaning, and pre-processing. It is perfectly fine if you'd like to do it in R, Matlab or other numerical programming packages you know.

(a) Data cleaning. There are a number of rows in this data file that are not correct. Can you find them and delete them? – use any sensible method you can think of.

(b) Data statistics. Compute the minimum, maximum, median and quartile values for each of the cleaned columns. Compute the mean and standard deviation.

(c) Histogram. Graph the histogram of each stock price for this period, in ten equal-width bins.
Note: Use the "FREQUENCY" function in Google Spreadsheet, the stepsize is stepsize = (max - min) / 10.

(d) Data normalization. Perform min-max normalization and z-score (zero-mean, unit deviation) normalization on these traces, compare (by visually plotting) the scores produced from the two normalizations.

(e) (*) Plot and compare the three proce traces on the same graph.
Note: How to plot them in order to make the visualization make sense? What visualization works, or seems useful? Why?

(f) Are these price traces correlated? Can you quantify this by computing the covariances between two stocks? Do their covariance change when the data is normalized?

(g) (*) If *risk* is literally defined as the expected fluctuation per unit currency in a given time. Which one of these three stocks is the most risky, why?

Answer:

(a) Here are a few ways: (i) Look at the dates. Notice that there are two repeated entries for 2/18, furthermore 2/19 and 2/20 are weekends and there should be no trading. (ii) Notice the closing prices in row 18 is the same with row 17. (iii) Look at the data values. Plot each of the GOOG, AAPL and CSCO traces over time, spot that there are upwards or downward spikes in rows 19 and 20. Double check with online finance data (google/yahoo finance) that there should be no spikes.
Delete rows 18, 19, 20. Done.

(b)-(e) For this and the rest of this problem, see worked answers at https://spreadsheets.google.com/... [2].

(f) The covariance changes but correlation coefficient does not.

(g)

$$\begin{aligned} \text{risk}_{GOOG} &= 0.024, \\ \text{risk}_{AAPL} &= 0.026, \\ \text{risk}_{CSCO} &= 0.074, \end{aligned}$$

so *CSCO* is the most risky one.

4. **Similarity Metrics comparison.**

(a) What are the definitions of Euclidean distance and cosine similarity?

(b) What is the relation between them (consider in a two dimensional space)? What is the form when the data vectors are unit vectors ($\|x\| = 1$)? Does this relation holds for high (>2) dimension spaces?

(c) In what situations the cosine similarity is used?

Answer:

(a) We use E and C to denote the Euclidean distance and cosine similarity,respectively, let two data points in a n-dimensional space be $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$, then

$$E = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2},$$

$$C = \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}}.$$

(b) Consider two points $\vec{x} = (x_1, x_2)$, $\vec{y} = (y_1, y_2)$ in this space, then the relationship between Euclidean distance and cosine similarity is shown in Figure 2.

To formulate this, consider the following equations:

$$\begin{aligned} E &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\ &= \sqrt{\|\vec{x}\|^2 + \|\vec{y}\|^2 - 2(x_1y_1 + x_2y_2)} \\ &= \sqrt{\|\vec{x}\|^2 + \|\vec{y}\|^2 - 2\|\vec{x}\|\|\vec{y}\|\frac{x_1y_1 + x_2y_2}{\|\vec{x}\|\|\vec{y}\|}} \\ &= \sqrt{\|\vec{x}\|^2 + \|\vec{y}\|^2 - 2\|\vec{x}\|\|\vec{y}\|C} \end{aligned}$$

[2] https://docs.google.com/spreadsheet/ccc?key=0Aq9ItKKywzgadHMxeW9FSHlaMEE3Y3FiYmVoRzk2LXc&authkey=CK6RyvgF&hlēn&authkey=CK6RyvgF#gid=0.

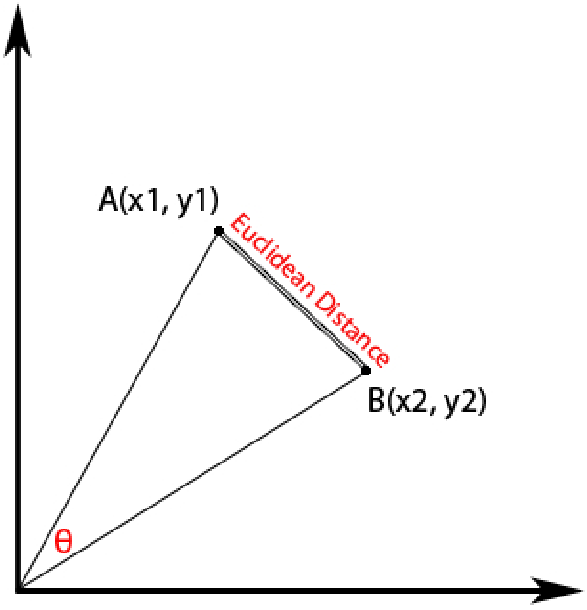


Figure 2: The relationship between Euclidean distance and cosine similarity in 2D space

If $\|\vec{x}\| = \|\vec{y}\| = 1$, $E = \sqrt{2(1 - C)}$.

This relation still holds in arbitrary n-dimensional space. The derivations are the same as above except in the first equation it is $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ instead of $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.

In general (which means for any dimensions), Euclidean distance of two points is the line segment connecting them, whereas cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them. Euclidean distance is a measurement of difference between two records, cosine similarity measures how similar they are.

(c) Cosine similarity is often used when the data is in high dimensional spaces. For example, to compare documents (keywords) in text mining. Also it is used to measure the similarity between two DNA sequences.

5. Suppose that a patient record table (see below) contains the attributes *name, gender, fever, cough, test-1, test-2, test-3,* and *test-4*, where *name* is an object identifier, *gender* is a symmetric attribute, and the remaining attributes are asymmetric binary.

For asymmetric attribute values, let the values Y (*yes*) and P (*positive*) be set to 1, and the value N (*no* or *negative*) be set to 0. Suppose that the distance between objects (patients) is computed based only on the asymmetric attributes[3].

(a) Please compute the distance between each pair of the three patients (Jack, Jim, and Mary).

(b) Which pair of them is more related (That is, they are more likely to have a similar disease)?

name	**gender**	**fever**	**cough**	**test-1**	**test-2**	**test-3**	**test-4**
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

Answer:

The asymmetric distance is defined as:

$$d(i,j) = \frac{r+s}{q+r+s} \qquad (5)$$

where

<table>
<tr><td colspan="2" rowspan="2">attributes</td><td colspan="2">obj_i</td></tr>
<tr><td>1</td><td>0</td></tr>
<tr><td rowspan="2">obj_j</td><td>1</td><td><i>q</i></td><td><i>r</i></td></tr>
<tr><td>0</td><td><i>s</i></td><td><i>t</i></td></tr>
</table>

(a) $d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$
$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$
$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$

[3] This question is taken from the textbook, on page 72, Example 2.18.

(b) Jim and Mary are unlikely to have a similar disease, Jack and Mary are the most likely to have a similar disease.

6. Find the edit distances (using only insertions and deletions) between the following pairs of strings.

   (a) **abcdef** and **bdaefc**.

   (b) **abccdabc** and **acbdcab**.

   (c) **abcdef** and **baedfc**.

   Answer:
   In the following we let A denote the first string, B denote the second string, C denote the *longest common subsequence* (LCS) of A and B, and use $|\cdot|$ to denote the length of a string.

   (a) The *longest common subsequence* (LCS) of A and B is: $C =$ **bdef**. Furthermore we have:

$$|A| = 6$$

$$|B| = 6$$

$$|C| = 4$$

   thus the edit distance $= |A| + |B| - 2 * |C| = 6 + 6 - 2 \times 4 = 4$.

   (b) Similarly, we have

$$C = \textbf{abdab} \text{ or } \textbf{acdab} \text{ or } \textbf{accab}$$

$$|A| = 8$$

$$|B| = 7$$

$$|C| = 5$$

   thus the edit distance $= |A| + |B| - 2 * |C| = 8 + 7 - 2 \times 5 = 5$.

   (c) Again, we have

$$C = \textbf{aef} \text{ or } \textbf{adf} \text{ or } \textbf{bdf}$$

$$|A| = 6$$

$$|B| = 6$$

$$|C| = 3$$

   thus the edit distance $= |A| + |B| - 2 * |C| = 6 + 6 - 2 \times 3 = 6$.