A deceptive side of data mining

Jerzy Letkowski

Anil Gulati

Western New England University

Abstract

Data mining is reaching a maturity level, moving from the research labs to organizations. One of the leaders in this area, Wal-Mart, has been implementing data mining procedures for quite some time. Wal-Mart captures data from millions of transactions every day, subsequently slicing and dicing the data to enhance their business knowledge. Many of the data mining procedures are based on statistical models. In particular, when looking for associations between random variables, statistical correlation models may be used. This paper shows how a simplistic approach to correlation may lead to incorrect conclusions. It goes back to Plato, by reiterating important ingredients of knowledge, some of which are impossible or very difficult to be incorporate to or extracted by data mining. The classic case “Storks vs. Babies Born” is used to illustrate a deceptive side of data mining.

Keywords: data mining, knowledge, probability, statistics, correlation, spreadsheet, mix-up.

KNOWLEDGE AND DATA MINING

There are many definitions of knowledge; this paper is not committed to bringing another one. Instead, it will take advantage of one of the oldest definitions, provided by Plato, student of Socrates, one of the greatest philosophers of all times.

According to Plato (Wikipedia- Knowledge - 2013), knowledge has three important criteria or ingredients: truth, justification and belief. Thus a statement, contributing to someone's knowledge must be `true`, `justified` and `believable`. The first two criteria are obvious. For whatever it means, one assumes that believability has some rational aspect. If one comes up with some conclusion (information) about some situation that, based on outcomes of some statistical study, is true and justified, it will also be believable if it is consistent with a theoretical model of the situation. Moreover, the criterion of believability is stronger if such a model is commonly accepted.

Data mining, a form of knowledge discovery, is analysis of data to discover historical patterns which when converted to future trends form the basis for knowledge driven decisions. Current advances in informational technologies have led to advanced transactional systems and analytical systems. Data mining serves as a link between these two systems by which data are converted into knowledge by discovering patterns, associations and relationships. These patterns, associations and relationships are discovered by posing open ended queries. Association analyses are one form of the knowledge discovery techniques used in data mining. Sophisticated statistical methods are at the core of these fact based knowledge discovery efforts. Correlation analysis plays a significant role in Classification and Regression algorithms, among others, used in data mining.

Since data mining is expected to contribute to awareness of individuals about their operational and decision situations, it is aimed at increasing their knowledge. Thus it is reasonable to scrutinize data mining procedures and results through the lens of the knowledge definition.

WHAT IS CORRELATION?

Correlation is a domain of Probability and Statistics that provides solutions for measuring the strength of dependence between numeric variables (Wikipedia- Correlation, 2013). In an extreme case, if two numeric, random variables, (*X*, *Y*), are independent, they are said to be uncorrelated. Coefficient of correlation, ρ, is a numeric measure of the strength of dependence between the variables (Feller, 1961, p. 211), (Grinstead, 2013, p.291)[1]:

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad \text{(i)}$$

σ_X and σ_Y an are the standard deviations of variables *X* and *Y*, respectively. *Cov*(*X*,*Y*) is a covariance of the random variables. It is defined as an expected value of the joint deviation of the variables from their means:

[1] This coefficient is also known as Pearson product-moment correlation coefficient, named to honor Karl Pearson (1857-1936), (Black, 2012, p.472).

$$Cov(X,Y) = E[(X - \mu_X)](Y - \mu_Y) = E(X \cdot Y) - \mu_X \mu_Y \quad \text{(ii)}$$

If values of X above μ_X tend to occur along with values of Y above μ_Y and values of X below μ_X tend to occur along with values of Y below μ_Y (+ and +) then the covariance will assume a positive value, indicating a positive relationship (correlation) between the variables. A similar reasoning with opposite signs (- and + or + and -) leads to a negative correlation. Judging the strength of correlation between the variables based *Cov*(*X*,*Y*) is not easy. One can generally say: the higher the absolute value of the covariance, the stronger the correlation between them. The coefficient of correlation is more convenient.

The coefficient of correlation has some interesting properties. First of all, ρ is a dimensionless quantity taking values between -1 and 1, ρ [-1, 1] (Spiegel, 1975, p.82). It is also scale and shift independent (Feller, 1961, p. 211):

$$\rho(a_X X + b_X, a_Y Y + b_Y) = \rho(X,Y) = \rho \quad \text{(iii)}$$

From the definition of the covariance (ii), one can see why the coefficient of correlation is equal to zero (ρ = 0) when X and Y are independent[2]. The coefficient of correlation values closer to -1 indicate a stronger negative relationship between the variables. Its values close to 1 are reflections stronger positive relationship. It is important to note that the correlation coefficient is very sensitive to outliers (Sharpie, 2010, p.168). Wherever possible, outlier should be eliminated before the coefficient of correlation is calculated. Last but not least, the strength of the relationship between the random variables measured by the coefficient of correlation applies only to linear type of relationships. Two variables may exhibit a perfect functional association but their coefficient of correlation will be close to zero. Figure 2 shows a perfect deterministic (sinusoidal) relationship between two data sets, *X*, *Y*. Yet the coefficient of correlation for the variables is zero, ρ(*X*, *Y*) = 0.

In the knowledge based framework, assuming true data, the coefficient of correlation only contributes to justification. It misses the belief ingredient. While showing the strength of the relationship, the coefficient of correlation does not explain any causality (De Veaux, 2006, p.153). Many cases presented in introductory business statistics textbooks attempt to address the issue of causality. Some simply and correctly point out that the coefficient of correlation is unable to resolve the cause-effect problem, suggesting that more analysis is necessary (Levine, 2011, p. 130). Some others try to reason about the causality. For example, in restaurant case where customer quality-rating vs. meal price was studied, a strong positive correlation was found and followed by assertion (Anderson, 2012, p. 140):

> "However, simply increasing the meal price at a restaurant will not cause the quality rating to increase."

One can dispute this conclusion. It certainly depends on the procedure used for setting the price of the meals at the restaurant. Ideally, a subject-matter expert-belief framework (set of relevant entities and rules) would help. Analyzing correlation in a context of the belief framework may also help in detecting undiagnosed variables, sometimes referred to lurking variables (Sharpie,

[2] The expected value of joint, independent random variables, *X*, *Y*, is equal to the product of the expected values of the variables (Feller, 1961, p. 199),: $E(X \cdot Y) = E(X) \cdot E(Y) = \mu_X \cdot \mu_Y$. Thus $Cov(X,Y) = \mu_X \cdot \mu_Y - \mu_X \cdot \mu_Y = 0$.

2010, p.173). Understanding the contents and structure of the belief framework may significantly contribute to better awareness of the underlying situation in which correlation is being studied.

Arguably, causality requires more systematic and knowledge-based approach. It should be consistent with the scientific approach in which empirical findings must be subject to commonly acceptable reasoning (Wikipedia- Science, 2013).

CAPTURING CORRELATION IN A SPREADSHEET

Previous section shows the theoretical definition of the coefficient of correlation, ρ. It can also be thought as the population correlation coefficient. A sample-driven coefficient of correlation, *r*, is calculated using the following formula (McClave, 2011, p.589):

$$r = \frac{SS_{xy}}{\sqrt{S_{xx}S_{yy}}}, \qquad \text{(iv)}$$

where SS_{xy} is the sum of products of deviations between the sample values and their respective means: $SS_{xy} = \sum (x-\bar{x})(y-\bar{y})$; SS_{xx} is to sum of squared deviations between the X sample values and its sample mean: $SS_{xx} = \sum (x-\bar{x})^2$; SS_{yy} is to sum of squared deviations between the Y sample values and its sample mean: $SS_{yy} = \sum (y-\bar{y})^2$.

A spreadsheet (Excel or Google) based formula to calculate this coefficient is much simpler. Assuming that sample *X* resides in a range named as X, and sample *Y* is stored in a range named as Y, the coefficient of correlation can be calculated directly using the following formula:

=Correl(X,Y) (v)

The order of the arguments is not relevant. This formula returns the coefficient of correlation in the same way for both the population and sample. Figures 1and 2 show applications of this formula in Google spreadsheets (Letkowski- Correlation, 2013)(Letkowski- Non-linear, 2013).

CASE: STORK COUNT VS. BORN BABY COUNT

Suppose that a region was subdivided into 13 rural areas each containing the same number of households. Two random variables, V1 and V2, represent the number of storks that have nested at the households and the number of babies born in those households, respectively. The following data:

Area	1	2	3	4	5	6	7	8	9	10	11	12	13
Stork Count, V1	3	2	1	5	6	7	2	5	3	4	2	3	4
Baby Count, V2	6	5	1	6	9	10	2	8	5	6	3	7	8

Figure 1 shows this data set along with calculated coefficient of correlation, ρ = 0.8911, in a Google spreadsheet (Letkowski-Correlation, 2013) .

The coefficient of correlation between the variables, *V1* and *V2*, indicates a strong positive association between the variables. A question remains:

Which variable is independent and which one is dependent? One wants to know if $V1 = f(V2)$ or if $V2 = f(V1)$,

This case has two obvious ingredients of "knowledge": "truth" (data) and "justification" (coefficient of correlation). The third ingredient, "belief" must be established based on "rational" background knowledge. If one believed in $V2 = f(V1)$ then one could conclude that "storks cause the babies to be born". More extremely (and irrationally) one could say the storks bring the babies. Obviously the common background knowledge must invalidate such reasoning. Thus solution $V1 = f(V2)$ seems more plausible. Many sources and documents support existence of some kind of association between storks and babies. One of the best can be found at (wiseGEEK, 2013). It leads to a conclusion that storks got attracted to homes where babies were born since the homes were better heated. These smart birds just prefer better heated nests.

CONCLUSIONS

The Storks vs. Babies case is a trivial. Yet, it shows an important aspect of correlation. One should not rely exclusively on [pure] data when studying associations between random variables especially when causality is in question. Whatever associations are being examined or discovered, for example, through data mining, they must be backed up by a believable theory.

Incorporating smart techniques in information search and discovery requires that raw data be annotated with some meta data and possibly linked to the existing and relevant knowledge base. It sounds like a perfect fit for the semantic Web methodology based on the Ontology Web Language (Antoniou, 2008, p.113). OWL is possibly the most expressive language for formal knowledge representation. OWL-annotated information sources can be processed using description logic (DL) reasoners or theorem provers (Baclawski, 2006, p. 58). This methodology opens up numerous research opportunities for data mining, utilizing statistical techniques, in general, and correlation, in particular.

Data mining enhanced by automated reasoning about statistical relations is a challenging endeavor. In order to data mining outcomes to be meaningful they must be consistent with the existing knowledge base. As Plato would say, they must be true, justified and believable.

REFERENCES

Anderson, D. R., Sweeney, D. J., Williams, T. A., (2012), Essentials of Modern Business Statistics with Microsoft® Excel. Mason, OH: South-Western, Cengage Learning.

Antoniou, G., van Harmele, F.(2008) A Semantic Web Primer (second edition). Cambridge, MA, London, UK: The MIT Press.

Baclawski, K., Niu, T. (2006). Ontologies for Bioinformatics. Cambridge, MA, London, UK: The MIT Press.

Black, K. (2012) Business Statistics For Contemporary Decision Making. New York, NY: John Wiley and Sons, Inc.

De Veaux, R. D., Velleman, P. F., Bock D.E. (2006) Intro to Stats. Boston, MA: Addison Wesley, Pearson Education, Inc.

Feller, W. (1961) An Introduction to Probability Theory and Its Applications. New York, London: John Wiley and Sons Inc.

Grinstead C.M, Snell, J.L. (2013) Introduction to Probability. Retrieved through: http://doingstats.com/ref/probability/introByGrinsteadSnell.html

Levine, D. M., Stephan, D.F., Krehbiel, T.C., Berenson, M.L. (2011) Statistics for Managers Using Microsoft® Excel, Sixth Edition. Boston, MA: Prentice Hall, Pearson Education, Inc..

Letkowski, J. – Correlation (2013) Case Stork Count vs. Baby Count. Retrieved from: http://doingstats.com/ref/correlation/correlation.html

Letkowski, J. – Non-linear (2013) Case Correlation Coefficient for Non-linear Relationship. Retrieved from: http://doingstats.com/ref/correlation/uncorrelation.html

McClave, J. T., Benson, P. G., Sincich, T. (2011) Statistics for Business and Economics, 11th Edition. Boston, MA: Prentice Hall, Pearson Education, Inc.

Sharpie, N. R., De Veaux, R. D., Velleman, P. F. (2010) Business Statistics. Boston, MA: Addison Wesley, Pearson Education, Inc..

Spiegel, M. R., (1975) Probability and Statistics, Schaum's Outline Series in Mathematics. New York, NY: McGraw-Hill Book Company.

Wikipedia- Correlation (2013). Correlation and dependence. Retrieved from: http://en.wikipedia.org/wiki/Correlation_and_dependence

Wikipedia- Knowledge (2013). Knowledge. Retrieved from: http://en.wikipedia.org/wiki/Knowledge

Wikipedia- Science (2013). Scientific method. Retrieved from: http://en.wikipedia.org/wiki/Scientific_method

Wikipedia-Stork (2013). White_Stork. Retrieved from: http://en.wikipedia.org/wiki/White_Stork

wiseGEEK (2013) What Is the Connection Between Storks and Babies?. Retrieved from: http://doingstats.com/ref/correlation/storks_vs_babies.html

APPENDIX

Correlation - Stork Count vs. Baby Count

File Edit View Insert Format Data Tools Help All changes saved in Drive

=CORREL(C5:C17,D5:D17)

	Storks	Babies
Area Number	V1	V2
1	3	6
2	2	5
3	1	1
4	5	6
5	6	9
6	7	10
7	2	2
8	5	8
9	3	5
10	4	6
11	2	3
12	3	7
13	4	8
Correlation Coefficient, ρ		0.8911

Legend:

Area Number	ID of a rural area, each having approximately the same number of households.
V1	The number of households, having stork nests in the area.
V2	The number of babaies born in the area,

There exists a strong correlation between V1 and V2,
Which theory is more plausible?
V1 = f(V2)
or
V2 = f(V1)

A solution can be found at:
http://doingstats.com/ref/correlation/storks_vs_babies.html

Formula: =CORREL(C5:C17,D5:D17)

Figure 1 Correlation between the number (V1) of households with stork nests and the number (V2) of babies born at these households.

Uncorrelated

File Edit View Insert Format Data Tools Help All changes saved in Drive

=CORREL(B4:B44,C4:C44)

	A	B	C	D	E	F	G	H
1								
2		inc	0.1571		Correl	0		
3		X	Y					
4	-π/2	-1.5708	-1.0000					
5		-1.4137	-0.9877					
6		-1.2566	-0.9511					
7		-1.0996	-0.8910					
8		-0.9425	-0.8090					
9		-0.7854	-0.7071					
10		-0.6283	-0.5878					
11		-0.4712	-0.4540					
12		-0.3142	-0.3090					
13		-0.1571	-0.1564					
14	0	0.0000	0.0000					
15		0.1571	0.1564					
16		0.3142	0.3090					
17		0.4712	0.4540					
18		0.6283	0.5878					
19		0.7854	0.7071					
20		0.9425	0.8090					
21		1.0996	0.8910					
22		1.2566	0.9511					
23		1.4137	0.9877					
24	π/2	1.5708	1.0000					
25		1.7279	0.9877					
26		1.8850	0.9511					
27		2.0420	0.8910					
28		2.1991	0.8090					
29		2.3562	0.7071					
30		2.5133	0.5878					
31		2.6704	0.4540					
32		2.8274	0.3090					
33		2.9845	0.1564					
34	π	3.1416	0.0000					
35		3.2987	-0.1564					
36		3.4558	-0.3090					
37		3.6128	-0.4540					
38		3.7699	-0.5878					
39		3.9270	-0.7071					
40		4.0841	-0.8090					
41		4.2412	-0.8910					
42		4.3982	-0.9511					
43		4.5553	-0.9877					
44	3π/2	4.7124	-1.0000					

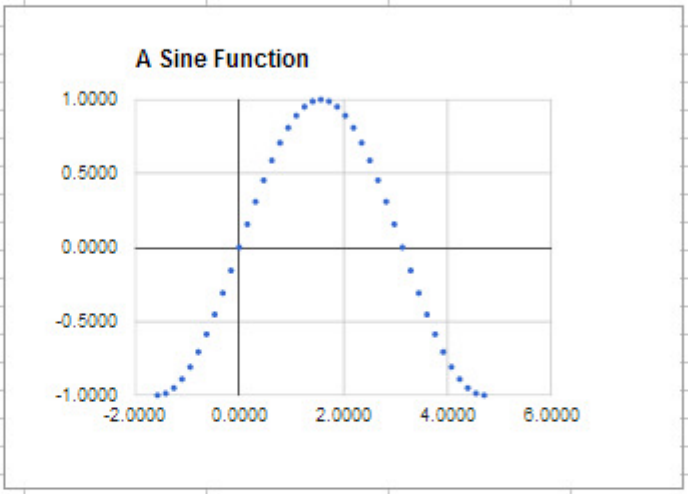

Figure 2 Correlation between independent variable X ($-\pi/2 \leq X \leq 3\pi/2$) and its sin(x) values.