An interdisciplinary glossary (DM–Statistics)

W. Hämäläinen

Statisticians and data miners have often difficulties to understand each other, because they have given different meanings for similar or even identic terms. The situation is even more confusing for researchers in other disciplines, who have certain knowledge in statistics but would like to use also data mining tools. In the worst case, this ambiguity can lead to wrong interpretations of results or at least cumbersome expressions like "The support of the support is $\mathbb{Z}^{+}$" or "The 99% confidence interval of the confidence is $[a, b]$".

The aim of this glossary is to list some of these ambiguous terms. The definitions are simplified so that one can get the main idea without any detailed background knowledge. Note that these definitions try to give only the most common meaning(s) of a term. For some terms, there can be other meanings (or nuances) even inside a discipline.

Association: (Stat.) Generally, statistical dependence between two (or more) random variables (see Dependence). In a narrower sense, it refers to statistical dependence between categorical variables, while the word "correlation" is used for numerical variables. The following is one of the oldest definitions for the association between two binary variables:

"*Having given the number of instances respectively in which things are thus and so, in which they are thus and not so, in which they are so and not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering to the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.*" (M.H. Doolittle, 1887)

Note that association does not necessarily mean causation.

Association rule: (DM) Traditionally, an association rule $X \rightarrow Y$ merely means sufficiently frequent co-occurrence of two attribute sets, X and Y. The sufficient frequency is defined by a user-specified threshold min_{fr}. In the traditional definition [AIS93], it has also been required that the "association" should be sufficiently strong, measured by confidence, $P(Y|X)$. As such, an association rule does not necessarily express any statistical dependence or the dependence may be negative, instead of the assumed positive dependence. Therefore, it has become more common to require that the rule expresses statistical dependence, i.e. $P(Y|X) > P(Y)$, and use statistical dependence measures like lift or leverage instead of confidence. (See confidence.)

Confidence: (DM) A traditional measure for the strengh of an association rule defined as $cf(X \rightarrow A) = \frac{fr(XA)}{fr(X)}$. In the frequentist interpretation of probability, this is the same as $P(A|X)$, an empirical estimate for the conditional probability of A given X.

Confidence interval: (Stat.) An interval where the true value of a parameter lies with a prespecified probability (e.g. 99% or 95% probability).

Confidence level (confidence coefficient): (Stat.) Probability that the confidence interval captures the true population parameter given a distribution of samples. The desired level of confidence (e.g. 99% or 95%) is set by the researcher (not determined by data).

Correlation: (Stat.) In a narrow sense, correlation refers to statistical dependence between two numerical variables. In a wider sense, it can refer to any departure of two or more random variables from independence. The most familiar measure of correlation is Pearson product-moment correlation coefficient (commonly ”the correlation coefficient”), which detects linear dependence between variables. Note that correlation does not necessarily mean causation.

Correlated (item) set (dependent set): (DM) At least two meanings: 1) A set of binary attributes, $X = \{A_1, \ldots, A_m\}$, is a correlated set if $P(A_1 = a_i, \ldots, A_m = a_m) \neq P(A_1 = a_1) \ldots P(A_m = a_m)$ for some value combination $(a_1, \ldots, a_m)$, $a_i \in \{0, 1\}$ (e.g. [GLW00]). This kind of sets have been called also “correlation rules” [BMS97] or “dependence rules” [SBM98], because each set expresses (implicitly) at least one statistical dependency rule (as defined below). Several types of correlated sets have been introduced with different names and extra requirements. 2) Sometimes, the condition part X of a classification rule $X \rightarrow C$, has been called a “correlated itemset” [NGR09].

Correlation rule (dependence rule): (DM) see correlated set.

Dependence: (Stat.) Any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. Note that statistical dependence does not necessarily mean causation. (See Association and Correlation.)

Dependency rule: (DM) Rule $X \rightarrow A$ expressing a statistical dependency between a set of binary attributes X and a single binary attribute A. Like an association rule, but expresses always a statistical dependency and has no minimum frequency requirements. (See also correlated set.)

Item: (DM) a synonym for a binary attribute, especially in occurrence data. Originally, used (logically) for products in the market basket analysis, but later on, it has become to mean any binary attribute, like *colour=red* or *age=(20–30).*

Head and **Tail**: (Prob.) Head is the range of values where the probability density function is relatively high. Tail is the complement of the head within the support; the large set of values where the density function is relatively low.

Head and **Tail**: (DM) In attribute enumeration algorithms, the head refers to the current set of attributes and the tail to attributes which can be added to it. Head can also refer to the consequence of an association rule, while the condition part is called the body of the rule.

Support: (DM) A term used in frequent itemset and association rule mining. The support of a pattern X can mean either absolute frequency, $fr(X)$ or relative frequency, $P(X)$. In the case of association rules, support($X \rightarrow A$) usually means $P(XA)$ or $fr(XA)$ but sometimes it can refer to $P(X)$ or $fr(X)$. The latter are also called the “coverage” of the rule, although coverage can sometimes refer to $P(XA)$ or $fr(XA)$.

Support: (Math.) The support of a function is the set of points where the function is not zero-valued or the closure of that set. In the probability theory and statistics, support of a random variable X, given a probability density function f, is the smallest closed set S such that $f(x) = 0$ for all $x \notin S$. (The same word may be used with other statistical measures, too, like the logarithm of the likelihood of a probability density function.)

References

[AIS93] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International*

Conference on Management of Data, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM Press, 1997.

[GLW00] G. Grahne, L.V.S. Lakshmanan, and X. Wang. Efficient mining of constrained correlated sets. In *Proceedings of the 16th International Conference on Data Engineering*, pages 512–521, 2000.

[NGR09] S. Nijssen, T. Guns, and L. De Raedt. Correlated itemset mining in ROC space: a constraint programming approach. In *Proceedings the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 647–656. ACM Press, 2009.

[SBM98] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, 1998.