Title: Category utility

URL: https://en.wikipedia.org/wiki/Category_utility

PageID: 8964665

Categories: Category:Cognitive science, Category:Machine learning

-----

Category utility is a measure of "category goodness" defined in Gluck & Corter (1985) and Corter & Gluck (1992) . It attempts to maximize both the probability that two objects in the same category have attribute values in common, and the probability that objects from different categories have different attribute values. It was intended to supersede more limited measures of category goodness such as " cue validity " (Reed 1972 ; Rosch & Mervis 1975) and "collocation index" (Jones 1983). It provides a normative information-theoretic measure of the predictive advantage gained by the observer who possesses knowledge of the given category structure (i.e., the class labels of instances) over the observer who does not possess knowledge of the category structure. In this sense the motivation for the category utility measure is similar to the information gain metric used in decision tree learning. In certain presentations, it is also formally equivalent to the mutual information , as discussed below. A review of category utility in its probabilistic incarnation, with applications to machine learning , is provided in Witten & Frank (2005 , pp. 260–262).

Probability-theoretic definition of category utility

The probability-theoretic definition of category utility given in Fisher (1987) and Witten & Frank (2005) is as follows:

where F = { f i } , i = 1 … n {\displaystyle F=\{f_{i}\},\ i=1\ldots n} is a size- n {\displaystyle n\ } set of m {\displaystyle m\ } -ary features, and C = { c j } j = 1 … p {\displaystyle C=\{c_{j}\}\ j=1\ldots p} is a set of p {\displaystyle p\ } categories. The term p (f i k) {\displaystyle p(f_{ik})\ } designates the marginal probability that feature f i {\displaystyle f_{i}\ } takes on value k {\displaystyle k\ } , and the term p (f i k | c j) {\displaystyle p(f_{ik}|c_{j})\ } designates the category- conditional probability that feature f i {\displaystyle f_{i}\ } takes on value k {\displaystyle k\ } given that the object in question belongs to category c j {\displaystyle c_{j}\ } .

The motivation and development of this expression for category utility, and the role of the multiplicand 1 p {\displaystyle \textstyle {\tfrac {1}{p}}} as a crude overfitting control, is given in the above sources. Loosely (Fisher 1987), the term p (c j) ∑ f i ∈ F ∑ k = 1 m p (f i k | c j) 2 {\displaystyle \textstyle p(c_{j})\sum _{f_{i}\in F}\sum _{k=1}^{m}p(f_{ik}|c_{j})^{2}} is the expected number of attribute values that can be correctly guessed by an observer using a probability-matching strategy together with knowledge of the category labels, while p (c j) ∑ f i ∈ F ∑ k = 1 m p (f i k) 2 {\displaystyle \textstyle p(c_{j})\sum _{f_{i}\in F}\sum _{k=1}^{m}p(f_{ik})^{2}} is the expected number of attribute values that can be correctly guessed by an observer the same strategy but without any knowledge of the category labels. Their difference therefore reflects the relative advantage accruing to the observer by having knowledge of the category structure.

Information-theoretic definition of category utility

The information-theoretic definition of category utility for a set of entities with size- n {\displaystyle n\ } binary feature set F = { f i } , i = 1 … n {\displaystyle F=\{f_{i}\},\ i=1\ldots n} , and a binary category C = { c , c ¯ } {\displaystyle C=\{c,{\bar {c}}\}} is given in Gluck & Corter (1985) as follows:

where p (c) {\displaystyle p(c)\ } is the prior probability of an entity belonging to the positive category c {\displaystyle c\ } (in the absence of any feature information), p (f i | c) {\displaystyle p(f_{i}|c)\ } is the conditional probability of an entity having feature f i {\displaystyle f_{i}\ } given that the entity belongs to category c {\displaystyle c\ } , p (f i | c ¯) {\displaystyle p(f_{i}|{\bar {c}})} is likewise the conditional probability of an entity having feature f i {\displaystyle f_{i}\ } given that the entity belongs to category c ¯ {\displaystyle {\bar {c}}} , and p (f i) {\displaystyle p(f_{i})\ } is the prior probability of an entity possessing feature f i {\displaystyle f_{i}\ } (in the absence of any category information).

The intuition behind the above expression is as follows: The term p (c) ∑ i = 1 n p (f i | c) log ■ p (f i | c) {\displaystyle p(c)\textstyle \sum _{i=1}^{n}p(f_{i}|c)\log p(f_{i}|c)} represents the cost (in bits) of optimally encoding (or transmitting) feature information when it is known that the objects to be described belong to category c {\displaystyle c\ } . Similarly, the term p (c ¯) ∑ i = 1 n p (f i | c ¯) log ■ p (f i | c ¯) {\displaystyle p({\bar {c}})\textstyle \sum _{i=1}^{n}p(f_{i}|{\bar {c}})\log p(f_{i}|{\bar {c}})} represents the cost (in bits) of optimally encoding (or transmitting) feature information when it is known that the objects to be described belong to category c ¯ {\displaystyle {\bar {c}}} . The sum of these two terms in the brackets is therefore the weighted average of these two costs. The final term, ∑ i = 1 n p (f i) log ■ p (f i) {\displaystyle \textstyle \sum _{i=1}^{n}p(f_{i})\log p(f_{i})} , represents the cost (in bits) of optimally encoding (or transmitting) feature information when no category information is available. The value of the category utility will, in the above formulation, be non-negative.

Category utility and mutual information

Gluck & Corter (1985) and Corter & Gluck (1992) mention that the category utility is equivalent to the mutual information . Here is a simple demonstration of the nature of this equivalence. Assume a set of entities each having the same n {\displaystyle n} features, i.e., feature set F = { f i } , i = 1 … n {\displaystyle F=\{f_{i}\},\ i=1\ldots n} , with each feature variable having cardinality m {\displaystyle m} . That is, each feature has the capacity to adopt any of m {\displaystyle m} distinct values (which need not be ordered; all variables can be nominal); for the special case m = 2 {\displaystyle m=2} these features would be considered binary , but more generally, for any m {\displaystyle m} , the features are simply m-ary . For the purposes of this demonstration, without loss of generality, feature set F {\displaystyle F} can be replaced with a single aggregate variable F a {\displaystyle F_{a}} that has cardinality m n {\displaystyle m^{n}} , and adopts a unique value v i , i = 1 … m n {\displaystyle v_{i},\ i=1\ldots m^{n}} corresponding to each feature combination in the Cartesian product ⊗ F {\displaystyle \otimes F} . (Ordinality does not matter, because the mutual information is not sensitive to ordinality.) In what follows, a term such as p (F a = v i) {\displaystyle p(F_{a}=v_{i})} or simply p (v i) {\displaystyle p(v_{i})} refers to the probability with which F a {\displaystyle F_{a}} adopts the particular value v i {\displaystyle v_{i}} . (Using the aggregate feature variable F a {\displaystyle F_{a}} replaces multiple summations, and simplifies the presentation to follow.)

For this demonstration, also assume a single category variable C {\displaystyle C} , which has cardinality p {\displaystyle p} . This is equivalent to a classification system in which there are p {\displaystyle p} non-intersecting categories. In the special case of p = 2 {\displaystyle p=2} there are the two-category case discussed above. From the definition of mutual information for discrete variables, the mutual information I (F a ; C) {\displaystyle I(F_{a};C)} between the aggregate feature variable F a {\displaystyle F_{a}} and the category variable C {\displaystyle C} is given by:

where p (v i) {\displaystyle p(v_{i})} is the prior probability of feature variable F a {\displaystyle F_{a}} adopting value v i {\displaystyle v_{i}} , p (c j) {\displaystyle p(c_{j})} is the marginal probability of category variable C {\displaystyle C} adopting value c j {\displaystyle c_{j}} , and p (v i , c j) {\displaystyle p(v_{i},c_{j})} is the joint probability of variables F a {\displaystyle F_{a}} and C {\displaystyle C} simultaneously adopting those respective values. In terms of the conditional probabilities this can be re-written (or defined) as

If the original definition of the category utility from above is rewritten with C = { c , c ¯ } {\displaystyle C=\{c,{\bar {c}}\}} ,

This equation clearly has the same form as the (blue) equation expressing the mutual information between the feature set and the category variable; the difference is that the sum ∑ f i ∈ F {\displaystyle \textstyle \sum _{f_{i}\in F}} in the category utility equation runs over independent binary variables F = { f i } , i = 1 … n {\displaystyle F=\{f_{i}\},\ i=1\ldots n} , whereas the sum ∑ v i ∈ F a {\displaystyle \textstyle \sum _{v_{i}\in F_{a}}} in the mutual information runs over values of the single m n {\displaystyle m^{n}} -ary variable F a {\displaystyle F_{a}} . The two measures are actually equivalent then only when the features { f i } {\displaystyle \{f_{i}\}} , are independent (and assuming that terms in the sum corresponding to p (f i ¯) {\displaystyle p({\bar {f_{i}}})} are also added).

Insensitivity of category utility to ordinality

Like the mutual information, the category utility is not sensitive to any ordering in the feature or category variable values. That is, as far as the category utility is concerned, the category set {small,medium,large,jumbo} is not qualitatively different from the category set {desk,fish,tree,mop} since the formulation of the category utility does not account for any ordering of the class variable. Similarly, a feature variable adopting values {1,2,3,4,5} is not qualitatively different from a feature variable adopting values {fred,joe,bob,sue,elaine} . As far as the category utility or mutual information are concerned, all category and feature variables are nominal variables. For this reason, category utility does not reflect any gestalt aspects of "category goodness" that might be based on such ordering effects. One possible adjustment for this insensitivity to ordinality is given by the weighting scheme described in the article for mutual information .

Category "goodness": models and philosophy

This section provides some background on the origins of, and need for, formal measures of "category goodness" such as the category utility, and some of the history that lead to the development of this particular metric.

What makes a good category?

At least since the time of Aristotle there has been a tremendous fascination in philosophy with the nature of concepts and universals . What kind of entity is a concept such as "horse"? Such abstractions do not designate any particular individual in the world, and yet we can scarcely imagine being able to comprehend the world without their use. Does the concept "horse" therefore have an independent existence outside of the mind? If it does, then what is the locus of this independent existence? The question of locus was an important issue on which the classical schools of Plato and Aristotle famously differed. However, they remained in agreement that universals did indeed have a mind-independent existence. There was, therefore, always a fact to the matter about which concepts and universals exist in the world.

In the late Middle Ages (perhaps beginning with Occam , although Porphyry also makes a much earlier remark indicating a certain discomfort with the status quo), however, the certainty that existed on this issue began to erode, and it became acceptable among the so-called nominalists and empiricists to consider concepts and universals as strictly mental entities or conventions of language. On this view of concepts—that they are purely representational constructs—a new question then comes to the fore: "Why do we possess one set of concepts rather than another?" What makes one set of concepts "good" and another set of concepts "bad"? This is a question that modern philosophers, and subsequently machine learning theorists and cognitive scientists, have struggled with for many decades.

What purpose do concepts serve?

One approach to answering such questions is to investigate the "role" or "purpose" of concepts in cognition. Thus the answer to "What are concepts good for in the first place?" by Mill (1843 , p. 425) and many others is that classification (conception) is a precursor to induction : By imposing a particular categorization on the universe, an organism gains the ability to deal with physically non-identical objects or situations in an identical fashion, thereby gaining substantial predictive leverage (Smith & Medin 1981 ; Harnad 2005). As J.S. Mill puts it (Mill 1843 , pp. 466–468),

> The general problem of classification... [is] to provide that things shall be thought of in such groups, and those groups in such an order, as will best conduce to the remembrance and to the ascertainment of their laws... [and] one of the uses of such a classification that by drawing attention to the properties on which it is founded, and which, if the classification be good, are marks of many others, it facilitates the discovery of those others.

From this base, Mill reaches the following conclusion, which foreshadows much subsequent thinking about category goodness, including the notion of category utility:

> The ends of scientific classification are best answered when the objects are formed into groups respecting which a greater number of general propositions can be made, and those propositions

more important, than could be made respecting any other groups into which the same things could be distributed. The properties, therefore, according to which objects are classified should, if possible, be those which are causes of many other properties; or, at any rate, which are sure marks of them.

One may compare this to the "category utility hypothesis" proposed by Corter & Gluck (1992) : "A category is useful to the extent that it can be expected to improve the ability of a person to accurately predict the features of instances of that category." Mill here seems to be suggesting that the best category structure is one in which object features (properties) are maximally informative about the object's class, and, simultaneously, the object class is maximally informative about the object's features. In other words, a useful classification scheme is one in which category knowledge can be used to accurately infer object properties, and property knowledge can be used to accurately infer object classes. One may also compare this idea to Aristotle 's criterion of counter-predication for definitional predicates, as well as to the notion of concepts described in formal concept analysis .

Attempts at formalization

A variety of different measures have been suggested with an aim of formally capturing this notion of "category goodness," the best known of which is probably the " cue validity ". Cue validity of a feature f i {\displaystyle f_{i}\ } with respect to category c j {\displaystyle c_{j}\ } is defined as the conditional probability of the category given the feature (Reed 1972 ; Rosch & Mervis 1975 ; Rosch 1978), p (c j | f i) {\displaystyle p(c_{j}|f_{i})\ } , or as the deviation of the conditional probability from the category base rate (Edgell 1993 ; Kruschke & Johansen 1999), p (c j | f i) – p (c j) {\displaystyle p(c_{j}|f_{i})-p(c_{j})\ } . Clearly, these measures quantify only inference from feature to category (i.e., cue validity), but not from category to feature, i.e., the category validity p (f i | c j) {\displaystyle p(f_{i}|c_{j})\ } . Also, while the cue validity was originally intended to account for the demonstrable appearance of basic categories in human cognition—categories of a particular level of generality that are evidently preferred by human learners—a number of major flaws in the cue validity quickly emerged in this regard (Jones 1983 ; Murphy 1982 ; Corter & Gluck 1992 , and others).

One attempt to address both problems by simultaneously maximizing both feature validity and category validity was made by Jones (1983) in defining the "collocation index" as the product p (c j | f i) p (f i | c j) {\displaystyle p(c_{j}|f_{i})p(f_{i}|c_{j})\ } , but this construction was fairly ad hoc (see Corter & Gluck 1992). The category utility was introduced as a more sophisticated refinement of the cue validity, which attempts to more rigorously quantify the full inferential power of a class structure. As shown above, on a certain view the category utility is equivalent to the mutual information between the feature variable and the category variable. It has been suggested that categories having the greatest overall category utility are those that are not only those "best" in a normative sense, but also those human learners prefer to use, e.g., "basic" categories (Corter & Gluck 1992). Other related measures of category goodness are "cohesion" (Hanson & Bauer 1989 ; Gennari, Langley & Fisher 1989) and "salience" (Gennari 1989).

Applications

Category utility is used as the category evaluation measure in the popular conceptual clustering algorithm called COBWEB (Fisher 1987).

See also

Abstraction

Concept learning

Universals

Unsupervised learning

References

Corter, James E.; Gluck, Mark A. (1992), "Explaining basic categories: Feature predictability and information" (PDF) , Psychological Bulletin , 111 (2): 291– 303, doi : 10.1037/0033-2909.111.2.291

, archived from the original (PDF) on 2011-08-10

Edgell, Stephen E. (1993), "Using configural and dimensional information", in N. John Castellan (ed.), Individual and Group Decision Making: Current Issues , Hillsdale, New Jersey : Lawrence Erlbaum, pp. 43– 64

Fisher, Douglas H. (1987), "Knowledge acquisition via incremental conceptual clustering", Machine Learning , 2 (2): 139– 172, doi : 10.1007/BF00114265

Gennari, John H. (1989), "Focused concept formation", in Alberto Maria Segre (ed.), Proceedings of the Sixth International Workshop on Machine Learning , Ithaca, NY : Morgan Kaufmann, pp. 379– 382

Gennari, John H.; Langley, Pat; Fisher, Doug (1989), "Models of incremental concept formation" , Artificial Intelligence , 40 (1– 3): 11– 61, doi : 10.1016/0004-3702(89)90046-5

Gluck, Mark A.; Corter, James E. (1985), "Information, uncertainty, and the utility of categories", Program of the Seventh Annual Conference of the Cognitive Science Society , pp. 283– 287

Hanson, Stephen José; Bauer, Malcolm (1989), "Conceptual clustering, categorization, and polymorphy", Machine Learning , 3 (4): 343– 372, doi : 10.1007/BF00116838

Harnad, Stevan (2005), "To cognize is to categorize: Cognition is categorization" , in Henri Cohen & Claire Lefebvre (ed.), Handbook of Categorization in Cognitive Science , Amsterdam: Elsevier, pp. 19– 43

Jones, Gregory V. (1983), "Identifying basic categories", Psychological Bulletin , 94 (3): 423– 428, doi : 10.1037/0033-2909.94.3.423

Kruschke, John K. ; Johansen, Mark K. (1999), "A model of probabilistic category learning", Journal of Experimental Psychology: Learning, Memory, and Cognition , 25 (5): 1083– 1119, doi : 10.1037/0278-7393.25.5.1083 , PMID 10505339

Mill, John Stuart (1843), A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation , London: Longmans, Green and Co. .

Murphy, Gregory L. (1982), "Cue validity and levels of categorization", Psychological Bulletin , 91 (1): 174– 177, doi : 10.1037/0033-2909.91.1.174

Reed, Stephen K. (1972), "Pattern recognition and categorization", Cognitive Psychology , 3 (3): 382– 407, doi : 10.1016/0010-0285(72)90014-x

Rosch, Eleanor (1978), "Principles of categorization", in Eleanor Rosch & Barbara B. Lloyd (ed.), Cognition and Categorization , Hillsdale, New Jersey : Lawrence Erlbaum, pp. 27– 48

Rosch, Eleanor; Mervis, Carolyn B. (1975), "Family Resemblances: Studies in the Internal Structure of Categories", Cognitive Psychology , 7 (4): 573– 605, doi : 10.1016/0010-0285(75)90024-9 , S2CID 17258322

Smith, Edward E.; Medin, Douglas L. (1981), Categories and Concepts , Cambridge, MA : Harvard University Press

Witten, Ian H.; Frank, Eibe (2005), Data Mining: Practical Machine Learning Tools and Techniques , Amsterdam: Morgan Kaufmann, archived from the original on 2020-11-27 , retrieved 2007-02-06