Category Learning, Computational Perspectives

Bradley C. Love
University of Texas at Austin

Judging a person as a friend or foe, a mushroom as edible or poisonous, or a sound as an l or r are examples of categorization problems. Because people never encounter the same exact stimulus twice, they must develop categorization schemes that capture the useful regularities in their environment. One challenge for psychological research is to determine how humans acquire and represent categories. Formally, category learning can be cast as the search for the function that maps from perceptual experiences to category membership. In this light, various models of human category learning are accounts of how people approximate this function from a limited number of observations. In this entry, human category learning will be considered from this formal perspective.

A function can be seen as a machine that takes inputs and generates outputs. For example, a soda machine (after it receives payment) takes a button press selection as input and outputs the appropriate brand of soda. In algebra class, most students are taught notation for functions, like y=f(x) where y is the output, x is the input, and f is the function. For example, y=f(x)=.5556x-17.7778 is a linear function that takes as input temperatures in Fahrenheit and outputs (i.e., converts to) temperatures in Celsius. Functions can also be nonlinear, like those that compute compounding interest. Whereas the temperature conversation function has continuous outputs, category functions have a finite set of discrete outputs. For example, a vertebrate animal can be categorized as a bird, mammal, fish, reptile, or amphibian. Thus, category functions are more like the soda machine than temperature conversion example, though the inputs to the category function can be quite complex, including all that a person can sense.

A Balancing Act Between Flexibility and Bias

Any learning system faces a tradeoff that is known in statistics as the bias-variance dilemma. This tradeoff involves finding the right balance of inductive bias and flexibility when learning the category function from a limited set of examples (as people do). Inductive bias guides a model's interpretation of data. To make an analogy, people have an inductive bias to view events co-occurring in time (e.g., smoke and fire) as causally related. A strong bias constrains the form of the category function that a model considers. For example, prototype models are strongly biased to only learn linear mappings (i.e., functions) from stimuli to categories, because prototype models represent categories by a single average (i.e., abstraction) of category members. For example, a prototype model would represent the category of birds as a single point (i.e., the prototype) that is the average of the features (e.g., size, color, etc.) of all birds (e.g., eagles, robins, penguins, sparrows, etc.). In practice, prototype models are best for learning categories that have a common family resemblance structure. For example, for the category birds, most birds have characteristics in common -- they tend to be small, have wings, can fly, etc. However, other items violate this structure (e.g., penguins, bats, etc.). Thus, the prototype model will have trouble with these items as they go against its bias of categories consisting of one single clump of items. Other models, like exemplar models, are weakly biased. Rather than averaging items together in memory, the exemplar model stores each item separately, which allows it to learn any possible function. For example, an exemplar model would represent the category of birds as the collection of all birds (i.e., one point for each category member). Exemplar models can learn any category function, whether it be linear or nonlinear. However, even the exemplar model has biases as it will learn some functions more rapidly (i.e., require fewer training examples) than others.

The latter point hints at why biases can be useful. When a model's bias corresponds to the actual category structure, it will learn the correct category function more rapidly than a less biased model or a model that has an incorrect bias. When a model is too flexible, it will be overly affected by the variance (i.e., noise) of the training examples it observes. In such cases, the model will fail to learn the underlying pattern and be overly affected by the idiosyncratic properties of the examples it has observed. Thus, the category learning function learned will initially be incorrect and will not accurately classify new examples. In general, more flexible (i.e., the less biased) models require more training examples to infer the underlying form of the category function.

Bayesian methods offer a natural way to deal with the bias-variance dilemma by simultaneously considering models of varying complexity (i.e., flexibility). Bayesian methods provide a means to combine prior beliefs with the current observation to determine the probability that an item is a member of a category. Biases for an individual Bayesian model can be explicitly built into the model's prior (i.e., beliefs held prior to observing any category members), which can be loosely thought of as seeding the model with hypothetical training examples (prior to observing any actual examples). For example, one may have a prior that heads and tails are both equally likely when flipping a coin. Thus, after flipping the coin once and observing tails, one would not conclude that the coin will always come up tails. The prior serves as a bias in favor of certain hypotheses about the category function. As many examples are observed, the importance of the initial prior wanes. This prior knowledge can be quite complex. For example, prior knowledge reflecting biological theories can be incorporated into a Bayesian model designed to learn about animal categories.

When Biases Prevent Learning

The preceding discussion focuses on how biases can be helpful or harmful in promoting (i.e., speeding) the learning of the category function. Of course, very strong biases, such as in the prototype model, can actually make some category functions unlearnable. While researchers typically focus on the rate at which people learn various categories, one important question is whether a model can even learn a category structure. For example, early work in neural networks was criticized and partially abandoned because certain category functions (ones that people could readily learn) where in principle unlearnable by the models. For example, these early models could not learn nonlinear functions, such as Exclusive-Or (e.g., “a spoon is small and steel or large and wooden”).

The learning rules in these models attempt to adjust connection weights to reduce prediction error (i.e., attempt to better approximate the category function). Unfortunately, the models were overly biased toward certain solutions and incapable of learning many category functions, no matter how long the model was trained.

Learnability concerns extend to all modeling approaches. For example, Bayesian models need to be sufficiently flexible (by having a wide range of possible hypotheses about what the category function could be) to be able to eventually learn the underlying category function.

The Curse of Dimensionality

One research challenge is determining what kinds of category functions are readily learned by people. The size of the input space is a major factor. Problems tend to be easy to learn that involve only a few dimensions. In the temperature conversation example, the x in f(x) was a single number (i.e., one dimension), as opposed to a lengthy vector of many inputs. As the dimensionality of the input space increases, the number of training examples needed to support learning can increase rapidly. Many learning problems humans face appear to be high-dimensional. For example, one could view every receptor in the retina as a dimension for visual learning problems. In practice, these dimensions are not all independent and the brain can take advantage of the structure in the world. Still, one must exploit such biases to learn in large problem spaces. Category learning functions that are smooth and regular are often the most tractable for learning models. Likewise, the effective dimensionality of a category learning problem can be reduced in situations in which attention can be oriented away from certain dimensions. For example, when a mechanic tries to classify what is wrong with a vehicle, the color of the vehicle is usually irrelevant to the decision.

Further Readings

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409-429.

Briscoe, E., & Feldman, K. (in press). Conceptual complexity and the bias/variance tradeoff. *Cognition*.

Love, B. C., & Tomlinson, M. (2010). Mechanistic Models of Associative and Rule-based Category Learning. In Denis Mareschal, Paul Quinn, Stephen Lea (Eds.), The Making of Human Concepts. Oxford, UK: Oxford University Press.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, & S. M. Kosslyn (Eds.), Essays in honor of William K. Estes, vol. 1: From learning theory to connectionist theory; vol. 2: From learning processes to cognitive processes (pp. 149-167). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.