Concept Learning

Thomas J. Palmeri and David C. Noelle

Introduction

Concepts are mental representations of kinds of objects, events, or ideas. We have concepts because they allow us to see something as a kind of thing rather than just as a unique individual. Concepts allow generalization from past experiences: By treating something as a kind—as an instantiation of some concept—as a member of some category—we can use what we have learned from other examples of that kind. Concepts permit inferences: Deciding predator from prey, edible from inedible, friend from enemy involves concepts. Concepts facilitate communication: Describing something as a kind of thing may obviate the need to provide details of the thing itself. Concepts permit different levels of abstraction: We know the difference between objects and ideas, animals and plants, dogs and cats, and terriers and collies. Concepts bring cognitive economy: What we learn about animals generally can be applied to specific animals without needless replication of that knowledge throughout our conceptual hierarchy. Concepts are fundamental building blocks of human knowledge (see Margolis and Laurence, 1999).

The focus of this article is on learning mental representations of new concepts from experience. We will also address how we use mental representations of concepts to make categorization decisions and other kinds of judgments.

Overview of Concept Learning Models

One goal of model development is to test specific hypotheses regarding the kinds of representations created during learning and the kinds of processes used to act upon those representations to make decisions. In order to develop a formal model of concept learning, a modeler must specify what perceptual information is provided by the sensory system, how that information is represented, how that information is compared with what has been learned about a concept, how this previously learned knowledge is represented in memory, and how decisions are made based on comparing perceptual information with stored conceptual representations (see PATTERN RECOGNITION).

A tacit assumption of many models of concept learning is that the perceptual system extracts information from the environment, passing a perceptual representation on to a conceptual stage of processing, which in turn generates an action. In the parlance of neural networks, the perceptual system provides the inputs to the network, the association weights and activation functions encode the conceptual representations, and decision processes use the outputs to generate a response. Most concept learning models characterize the perceptual system as a dimensionality reduction device (see OBJECT RECOGNITION and OBJECT STRUCTURE, VISUAL PROCESSING). For example, in the case of vision, the input on the retina is an extremely high-dimensional representation, with every photoreceptor effectively encoding an independent dimension of sensation. The perceptual system creates a relatively low-dimensional representation by recoding the retinal input in terms of a smaller number of features or dimensions. These vectors of features or dimensions serve as the inputs to the concept learning network.

The distinction between features and dimensions is fully discussed elsewhere (e.g., Tversky, 1977). The members of a category—the extension of a concept—may be seen

as clusters of perceptual vectors in a psychological space. An important type of concept learning entails associating regions in that space with particular category labels.

A *featural representation* in a neural network essentially consists of a vector of input nodes encoding the presence or absence of primitive elements. Similar stimuli share many common features. Dissimilar stimuli correspond to uncorrelated vectors. Often in simulation modeling, feature representations have no direct relationship to the actual stimuli of an experiment, but are instead designed to capture the statistical relationships among stimuli.

A *dimensional representation* is not discrete, but represents information in terms of values along continuously varying psychological dimensions. Similar stimuli have similar values along the dimensions, occupying adjacent locations in psychological space. Often in simulation modeling, dimensional representations may be derived from physical properties of actual stimuli or may be derived from similarity ratings (or other measures of psychological proximity) made by subjects using techniques such as multidimensional scaling (although feature representations can be derived as well).

The core of this article reviews models of how concepts are learned and represented in neural networks. One important distinction between models is whether concepts have LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS (q.v.). Another is whether conceptual knowledge relies on abstractions, such as rules or prototypes, or relies on specific exemplar knowledge. Most models focus on supervised learning, where trial-by-trial feedback is supplied, but some models address unsupervised learning. Most models focus on how concepts are learned within a category learning paradigm, whereby subjects learn to produce the correct label for each stimulus, but some models address how subjects can learn to infer properties other than the category label. Most models focus on learning from induction over examples, but some address learning from instruction or from explanation as well.

Concept Learning Models

Rule Models

Early philosophers conjectured that all concepts were decomposable into necessary and sufficient conditions for membership (e.g., Plato in Margolis and Laurence [1999]). A “triangle” is a closed form with three sides, a “bachelor” is an unmarried adult male, and so forth. Conceptual rules are like carefully worded definitions provided to a student. Indeed, a strength of this hypothesis is the apparent alignment of mental representations with self-reports of conceptual knowledge. Generally, to be considered a rule, a concise definition is required—if arbitrarily complex rules are allowed, the rule hypothesis becomes vacuous, because virtually any representation can be characterized by complex rules. Therefore, rule representations typically include just a small set of dimensions— sometimes only a single dimension. Although this may seem overly restrictive, humans frequently exhibit reliance on individual dimensions during concept learning (Nosofsky, Palmeri, and McKinley, 1994).

Simple neural network units may be connected to compute logical functions (see NEURAL AUTOMATA AND ANALOG COMPUTATIONAL COMPLEXITY). A rule involving a threshold along a single dimension is the simplest example, and serves as the basis for rules in some concept learning models (e.g., Ashby et al., 1998; Erickson and Kruschke, 1998). Assuming a dimensional input representation, a single-dimension rule simply involves a learned weighted connection w_{ij} from input node a_i to output unit o_j with learned bias θ_j

$$o_j = \frac{1}{[1 + \exp(-w_{ij}a_i - \theta_j)]} \qquad (1)$$

The sigmoidal activation of this unit is proportional to the likelihood of category membership. A network of this kind essentially implements a linear decision boundary in psychological space, with the boundary orthogonal to one of the psychological dimensions, and with the position of the boundary specified by the learned bias θ_j (see Figure 1*A*).

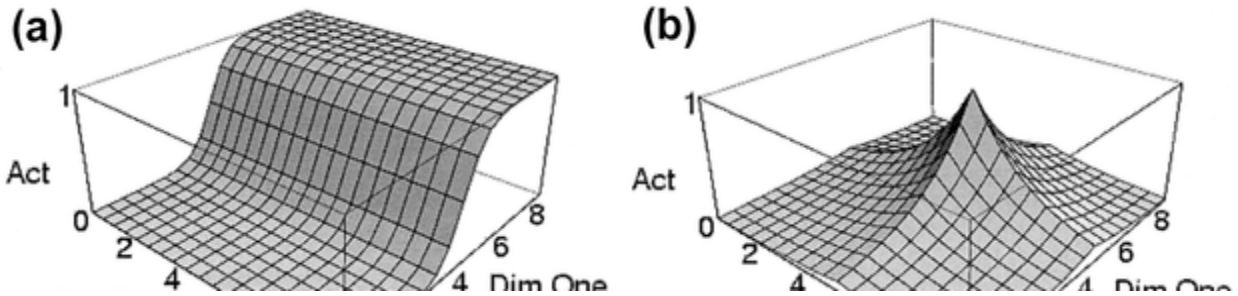
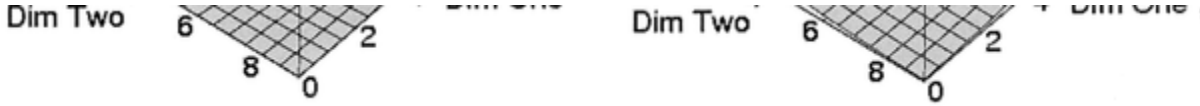


Figure 1. Examples of activation (Act) as a function of location in a two-dimensional psychological space (Dim One × Dim Two) for (a), the logistic sigmoidal function in Equations 1–3 and for (b), the radial-basis function in Equation 4.



Prototype Models

Although people often report conceptual knowledge in terms of rules, there are reasons to question rules as the universal basis for conceptual knowledge. Many common concepts are difficult to define using rules—Wittgenstein suggested the example of “games” as appearing to defy definition (see Margolis and Laurence, 1999). Instead, concepts seem to possess “family resemblances,” with instances bearing many similarities but no characteristics common to all members. Human performance often belies the existence of rules in that some items appear to be “better” members than others in terms of typicality ratings, speed of processing, and inductive power.

Such findings led to an alternative view of conceptual knowledge based on abstract prototypes (see Rosch and also Lakoff in Margolis and Laurence, 1999). A prototype need never be directly experienced, but can be formed by averaging across observed instances. In psychological space, the prototype is the centroid of a cloud of instances. New items are classified according to their relative similarity to learned prototypes, with typicality effects emerging from this process. Learning just two prototypes effectively partitions psychological space into two regions separated by a linear boundary, but this boundary is a fuzzy (probabilistic) one and is unconstrained in terms of its orientation within psychological space.

A simple two-layer prototype model assumes an input layer of features (or dimensions) with learned associations to category output nodes. Each output unit o_j corresponds to a single concept prototype, and the weights w_{ij} from inputs a_i to each o_j reflects the strength of association of particular stimulus features with that concept:

$$o_j = \frac{1}{\left[1 + \exp\left(-\sum_i w_{ij} a_i - \theta_j\right)\right]} \qquad (2)$$

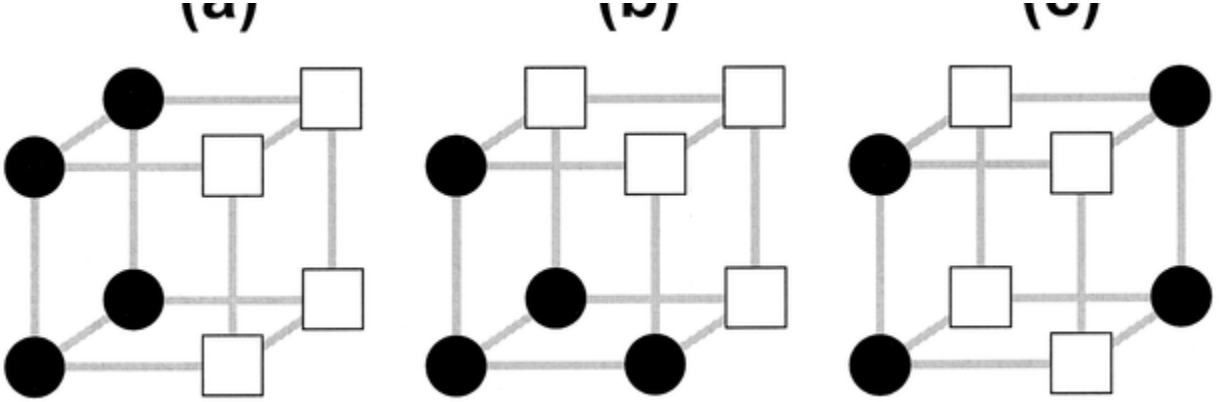
The probability $P(j)$ of categorizing a stimulus as a member of category j can simply be given by the relative activation of node j compared to all other nodes. Other more dynamic mechanisms, such as lateral inhibition, can also introduce competition between concepts in WINNER-TAKE-ALL NETWORKS (q.v.).

This simple network learns to associate input stimuli with their corresponding categories. A related approach is to train ASSOCIATIVE NETWORKS (q.v.) to reproduce the features of each training instance as well as the correct category label (e.g., McClelland and Rumelhart, 1985). This pattern completion approach permits the network to not only categorize stimuli, but also to infer other missing features as well. More powerful pattern completion arises when COMPUTING WITH ATTRACTORS (q.v.) such as in the Brain State in a Box model (see ASSOCIATIVE NETWORKS). Recurrent connections between all category label units and all feature units permit the network to encode soft constraints guiding how activation settles over time. In addition to their pattern completion properties, learned basins of attraction in such networks can instantiate nonlinear decision boundaries between categories, potentially creating a kind of prototype model that incorporates information about both the mean and the variability of a distribution of category exemplars.

Exemplar Models

Simple concept learning networks of the sort just described can learn category structures (a) and (b) depicted in Figure 2—linearly separable categories—but cannot learn category structure (c)— nonlinearly separable categories. This is the classic XOR problem that stymied early developments in neural networks (see PERCEPTRONS, ADALINES, AND BACKPROPAGATION). In contrast, multi-layered networks with an input layer, a hidden layer, and an output layer can learn these category structures. Activation of hidden and output nodes is determined by a nonlinear sigmoid function like that shown in Equation 2. Learning takes place via gradient descent on error, with knowledge fully distributed throughout the network connections (see BACKPROPAGATION: GENERAL PRINCIPLES). These “backpropagation networks” are powerful learning devices and, with sufficient hidden nodes, they can acquire concepts of nearly unlimited complexity.

Figure 2. Depictions of three category structures. Individual stimuli are composed of three binary-valued dimensions. The



dimensional values of a stimulus are specified by its location in the three-dimensional psychological space. For each structure, black circles represent stimuli in one category and white squares represent stimuli in another category.

Psychological models of concept learning attempt to model human behavior, with all its errors and apparent inefficiencies. Although backpropagation networks are powerful machine learning devices, they make relatively poor models of human concept learning. For one, backpropagation networks are insensitive to the psychological dimensional structure within the categories. From a statistical standpoint, structure (a) and structure (b) in Figure 2 have equivalent complexity, so backpropagation networks learn both types equally quickly. But people find structure (a) far easier to learn than structure (b) because structure (a) permits attending to a single dimension. Backpropagation networks generally learn linearly separable categories, such as structure (b), far more quickly than nonlinearly separable categories, such as structure (c). By contrast, people find structure (c) easier to learn. More generally, people are not constrained by linear separability, and often exhibit more rapid learning of nonlinearly separable than linearly separable categories. Finally, backpropagation networks suffer from catastrophic forgetting in which new concept learning overwrites previous concept learning. There have been many modifications to simple backpropagation networks to combat this problem. Common to many approaches is the use of semilocalized representations instead of fully distributed representations (see LOCALIZED VERSUS DISTRIBUTED REPRESENTATIONS).

Exemplar-based models assume local representations (see Smith and Medin in Margolis and Laurence, 1999). In contrast to rule and prototype models, concepts are represented extensionally, in terms of specific category instances. A number of variations of exemplar models have been proposed, and a vast set of empirical phenomena is consistent with them (Nosofsky and Kruschke, 1992). Perhaps the best-known neural network exemplar model is ALCOVE (Kruschke, 1992), which is largely derived from the generalized context model of categorization (Nosofsky, 1986), but incorporates error-driven learning.

ALCOVE is a three-layered feedforward network in which activation passes from a dimensional input layer, with each dimension scaled by a learned selective attention weight, to an exemplar-based hidden layer, to a category output layer via learned association weights. In contrast to backpropagation networks (compareFigures 1(a) and 1(b)), activation h_j of hidden exemplar node j is based on its similarity to the input stimulus

$$h_j = \exp\left[-c\left(\sum_i \alpha_i |e_{ji} - a_i|^r\right)^{1/r}\right] \qquad (3)$$

where a_i is the value of the input stimulus along dimension i, e_{ji} is the value of exemplar j along dimension i, α_i is the learned attention to dimension i, c is a similarity scaling parameter, and r determines the psychological distance metric (see also RADIAL BASIS FUNCTION NETWORKS). When optimally allocated, selective attention weights emphasize differences along diagnostic dimensions and deemphasize differences along nondiagnostic dimensions (Nosofsky, 1986).

ALCOVE also learns to associate exemplars with category outputs. Activation of category output node k is given by

$$o_k = \sum_j w_{kj} h_j \qquad (4)$$

where w_{kj} is the learned association weight between exemplar j and output node k. The probability of categorizing the input stimulus as a member of category K is given by

$$P(K) = \frac{\exp(\varphi o_K)}{\sum_k \exp(\varphi o_k)} \qquad (5)$$

where ϕ is a response mapping parameter. Attention weights (α_i) and association weights (w_{kj}) are learned by gradient descent on error.

Although exemplar models like ALCOVE have accounted for a wide variety of fundamental categorization phenomena (Nosofsky and Kruschke, 1992), until recently they have ignored the time-course of making a categorization decision. One avenue of recent theoretical development has addressed the time-course of the accumulation of perceptual evidence used to categorize a stimulus (Lamberts, 2000), modeling how perceptual processes make some information available sooner than others. Another avenue of development has examined the time-course of making categorization decisions. Nosofsky and Palmeri (1997) proposed a stochastic exemplar-retrieval model with a competitive decision process to account for both categorization response probabilities and response times. Although this theoretical development was formalized using mathematical modeling tools, such as a random walk decision process, the dynamics of these stochastic, competitive categorization models could be implemented in various neural network architectures as well.

Mixed Models

Exemplar models have been shown to account for a variety of results, including some that were originally thought to unequivocally indicate rules or prototypes as concept representations. Yet, there seems to be some emerging evidence that people do use abstractions, particularly abstract rules, as concept representations. Clearly, people can be instructed to use rules before they have experienced any examples (Palmeri, 1997; Noelle, Cottrell, and McKinley, 2002). Also, it appears that people may approach the task of learning categories by testing simple categorization rules (e.g., Nosofsky et al., 1994), although people may eventually shift to using exemplars with experience (Johansen and Palmeri, in press). One important focus of current research is developing and testing formal models with mixed representations. At one extreme are models that posit functionally independent rule-based and exemplar-based systems that race to completion (e.g., Palmeri, 1997); exemplar-based representations gain strength with repeated exemplar experience and eventually win the race. Alternatively, rule and exemplar representations may be functionally independent, but the outputs of these systems may compete based on strength of evidence rather than completion time (e.g., Ashby et al., 1998). Erickson and Kruschke (1998) proposed a neural network model (ATRIUM) with separate rule and exemplar representations that compete, with the model learning whether rule-based or exemplar-based information should be used to categorize a particular instance. Finally, other architectures have proposed combinations of rules, exemplars, and perhaps other representations within a single representational medium.

Neuroscientific Models

Neurobiological findings have begun to constrain concept learning models by ruling out mechanisms that resist reasonable neural implementation. Some recent models have included hypotheses concerning how conceptual knowledge is instantiated in the brain.

COVIS is one example of a model grounded in neurophysiology (Ashby et al., 1998). COVIS is a mixed representational model, incorporating implicit decision boundaries and explicit unidimensional rules. The implicit learning system, assumed to reside within the striatum, encodes a category decision boundary (although other representations—such as exemplar-like coarse-coded topological maps of psychological space with regions associated with category labels—also fits within this general framework). Verbal rule processing is assumed to exist in the prefrontal cortex, with the selection of rule-attended dimensions handled by a reinforcement learning process in the anterior cingulate, mediated by projections from the basal ganglia. Given a stimulus, the implicit and rule-based systems compete to provide a category response. This model has been applied to concept learning deficits seen in the very old and the very young, in patients with Parkinson's disease and Huntington's disease, in clinically depressed patients, in individuals with focal brain lesions, and in nonhuman animals.

Other neurally oriented concept learning models have focused on the role of prefrontal cortex in representing and actively maintaining rule-based information during categorization (Noelle et al., 2002; O'Reilly et al., 2002). A distinguishing feature of these models is the use of signals that encode changes in expected future reward (see REINFORCEMENT LEARNING)—emanating from the basal ganglia dopamine system (see DOPAMINE, ROLES OF)—to determine when a useful rule representation has been found and should be gated into a prefrontal working memory system. Like COVIS, these dopamine-gating models incorporate a mixed representation, integrating rules with a kind of procedural knowledge. Unlike COVIS, these models focus on procedural knowledge embedded in multilayer networks, presumably located within the cortex, rather than on a special implicit learning system within the striatum. Rule representations, stored as patterns of activity in a prefrontal attractor network (see COMPUTING WITH ATTRACTORS), do not directly compete with these procedural systems, but rather modulate them through the injection of activity. Models of this kind have provided explanations for frontal lesion data, suggested a coarse topological organization for prefrontal cortex, captured patterns of performance on dynamic classification tasks, explained interference effects in instructed category learning, and illuminated learning deficits in schizophrenia.

Discussion

In this article, we limited our discussion to the kinds of representations and processes that subserve a particular aspect of concept learning, namely learning to categorize. Recent work has investigated other topics, as well, including how people learn to infer properties other than the category label and how learning about categories may influence perceptual processing in a top-down manner (e.g., Schyns, Goldstone, and Thibaut, 1998).

An important focus of current research was outlined in this article: Do people use different kinds of concept representations, how are those representations learned, and is the dominance of particular representations modulated by experience or other task demands? In order to help answer these questions, and in order to develop neurally plausible models of concept learning, some researchers are beginning to incorporate the constraints imposed by various neuroscientific sources of evidence, including studies of patients with focal brain damage, functional imaging and evoked potential studies, and single unit recordings in animals.

Road Map: Psychology

Related Reading: Feature Analysis ♢ Object Recognition ♢ Pattern Recognition

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., and Waldron, E. M., 1998, A formal neuropsychological theory of multiple systems in category learning, *Psychol. Rev.*, 105:442–481.

Erickson, M. A., and Kruschke, J. K., 1998, Rules and exemplars in category learning, *J. Exp. Psychol.*, 127:107–140.

Johansen, M. K., and Palmeri, T. J., (in press), Are there representational shifts during category learning? *Cognitive Psychology*.

Kruschke, J. K., 1992, ALCOVE: An exemplar-based connectionist model of category learning, *Psychol. Rev.*, 99:22–44.

Lamberts, K., 2000, Information-accumulation theory of speeded categorization, *Psychol. Rev.*, 107:227–260.

Margolis, E., and Laurence, S., 1999, *Concepts: Core Readings,* Cambridge, MA: MIT Press. ◆

McClelland, J. L., and Rumelhart, D. E., 1985, Distributed memory and the representation of general and specific information, *J. Exp. Psychol.*, 114:159–188.

Noelle, D. C., Cottrell, G. W., and McKinley, C. R. M., 2002, Modeling individual differences in the specialization of an explicit rule, Manuscript under review.

Nosofsky, R. M., 1986, Attention, similarity, and the identification-categorization relationship, *J. Exp. Psychol.*, 115:39–57.

Nosofsky, R. M., and Kruschke, J. K., 1992, Investigations of an exemplar-based connectionist model of category learning, in *The Psychology of Learning and Motivation*, vol. 28 (D. L. Medin, Ed.), San Diego, CA: Academic Press, pp. 207–250. ◆

Nosofsky, R. M., and Palmeri, T. J., 1997, An exemplar-based random walk model of speeded classification, *Psychol. Rev.*, 104:266–300.

Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C., 1994, Rule-plus-exception model of classification learning, *Psychol. Rev.*, 101:53–79.

O'Reilly, R. C., Noelle, D. C., Braver, T. S., and Cohen, J. D., 2002, Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control, *Cerebral Cortex*, 12:246–257.

Palmeri, T. J., 1997, Exemplar similarity and the development of automaticity, *J. Exp. Psychol.*, 23:324–354.

Schyns, P. G., Goldstone, R. L., and Thibaut, J. P., 1998, The development of features in object concepts, *Behav. Brain Sci.*, 21:1–40. ◆

Tversky, A., 1977, Features of Similarity, *Psychology Review*, 84:327–352.

«« Previous Next »»