Warning: The Exemplars in Your Category Representation May Not Be the Ones Experienced During Learning

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13902 USA

Daniel C. Silliman (dsillim1@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13902 USA

Keywords: Connectionist; Exemplar; Category learning; Computational modeling

The WARP Model of Category Learning

Research on categorization and classification learning has greatly benefitted from the use of computational modeling which requires making all theoretical assumptions explicit and provides a direct means of theory evaluation by *fitting* behavioral data. The field has advanced notably through model comparison relative to benchmark data on human category learning performance. Exemplar theory has become a leading psychological explanation largely due to the success of its formal models in fitting human data across a number of tasks (Kruschke, 1992; Nosofsky & Palmeri, 1997).

The exemplar view casts categorization as based on an explicit calculation of similarity between the to-be-categorized stimulus and instances stored in long-term memory (exemplars) associated with each category. The similarity is computed as an inverse exponential function of distance between psychological representations in a multidimensional space. This representational space can be transformed by stretching or shrinking dimensions using selective attention. The category with exemplars of greater similarity (less distance) to the stimulus is activated. This account has been extended in the ALCOVE model (Kruschke, 1992) which implements adaptive learning of attentional weights on the stimulus dimensions and association weights between each exemplar and category.

While exemplar models have shown a high degree of success in fitting behavioral data, they do not provide an account of representation learning. These models generally assume that each item in the input domain has a unique psychological representation (estimated via multidimensional scaling) that remains fixed throughout the category learning process. Further, a strict correspondence holds between the category representation and the stimulus items known to be members of that category (note: reference point models can also use centroids of clusters of exemplars).

This is in strong contrast to feedforward artificial neural networks that gradually learn representations to optimize task performance (Rumelhart, Hinton, & Williams, 1986). In standard connectionist models, each stimulus gets recoded at a "hidden" layer based on a set of optimized synapse-like weights that yield a distributed representation across the hidden nodes—which can be seen as a point in a constructed multidimensional space. A second set of weights connects these hidden nodes to an output layer of class nodes. The internal representations are incrementally repositioned in weight space via gradient descent to optimize accurate prediction at the output layer.

The Weights-as-Adaptive-Reference-Points (WARP) model is designed to bridge the reference point similarity-based approach of exemplar models with the flexibility and psychological plausibility of learned representations in neural networks. This merger of design principles is achieved by replacing the localist exemplar node representations (as in ALCOVE) with a layer that follows the foundational connectionist design principles of: 1) a forward pass that computes activation based on a function of the 'net input,' i.e., the input activations multiplied by their weights; and 2) a backward pass that modifies the weights to minimize task error and estimate the function to be approximated.

On the connectionist view, the hidden nodes are constructed dimensions that usefully transform the values of a stimulus in input space to a set of values in another representational space. On the exemplar view, each hidden node is a reference point to the location of a training item in input space and its activity indexes the proximity of that point to a stimulus. We propose a new formulation that allows the hidden nodes to function according to connectionist mechanics and yet act as reference points. The result is that the model discovers its own reference points using task-driven error minimization as opposed to making a commitment to the inputs themselves as the basis for the reference points.

The WARP model functions by taking the encoding weights to each hidden node as its "address" or reference point location in input space. As the weights change via learning by backpropagation, each node follows a trajectory in weight space from its initial random location toward a place where its task is functionality optimized. The 'net input' is the vector multiplication between the input activations and the incoming weights to a node. This is a dot product or linear algebraic measure of similarity (i.e., the angle between the vectors) as opposed to a spatial distance metric. The critical similarity computation between stimulus and reference point occurs implicitly in the forward pass. To make this work as intended, a simple, novel activation function at the hidden layer is used which takes the form of Equation 1:

$$exp[(a \cdot b) - k] \tag{1}$$

where a is the vector of input activations, b is the vector of incoming input->hidden node weights, and k is a constant value set to the number of dimensions in the category structure. The key property of this function is this: the more closely the incoming weight vector for a hidden node approximates the values of an input vector, the greater the activation of the hidden node. Over the course of training, different hidden nodes will be repositioned to parts of weight space that allow them to respond to particular regions in input space: to get better at classifying is to move the adaptive reference points to useful positions. A standard association layer connects the hidden nodes to class nodes and a softmax output layer is used to determine the class probabilities

WARP utilizes a set of connectionist-style free parameters: learning rate, number of hidden nodes (i.e., density of the implicit covering map), and range of random initialization for incoming weights; and can also incorporate a set of reference point model-style free parameters: degree of sensitivity of reference points and a response mapping constant for determining class activations.

Preliminary testing has shown promising fits to the classic behavioral benchmark of the Shepard, Hovland, and Jenkins (1961) six types of elemental category structures (dataset from Nosofsky et al., 1994). This investigation also revealed that the WARP model discovers more parsimonious reference points when available: instead of always dedicating each hidden node to a single input, WARP can develop reference points that respond strongly to particular feature correlations or unidimensional rules. In conjunction with classic exemplar-style nodes, these feature detector-style nodes allow the model to efficiently handle various and complex category structures. The use of this multi-strategy toolkit mirrors the diversity and flexibility of human category learning (Ashby, Alfonso-Reese, & Waldron, 1998).

In addition to modeling human behavior, WARP has also been initially tested for potential application as a classifier in the domain of machine learning. Different parameterizations of the model, while inappropriate for capturing the pace and nuance of human learning, show highly rapid and efficient performance on the iris flower benchmark dataset. Interestingly, the model solves the classification problem using *discriminative prototypes* that maximize distance to competing classes while minimizing distance to the target class. Continued investigations of the model are underway to better reveal the nature and diversity of the solutions WARP discovers for different types of classification problems; and to determine the power of the model in addressing the goals of psychological explanation and advancing AI.

Relevant Publications

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within-versus between-category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571.

Conaway, N., & Kurtz, K. J. (2017a). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312-1323.

Conaway, N., & Kurtz, K. J. (2017b). Solving nonlinearly separable classifications in a single-layer neural network. *Neural computation*, *29*(3), 861-866.

Honke, G. & Kurtz K.J. (in press). Similarity is as similarity does? A critical inquiry into the effect of thematic association on similarity. *Cognition*.

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & cognition*, *43*(2), 266-282.

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64*, 66-75.

References

Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, *22*(3), 352-369.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, *104*(2), 266.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.