Psychonomic Bulletin & Review
2008, 15 (2), 256-271
doi: 10.3758/PBR.15.2.256

Generalization and similarity in exemplar models of categorization: Insights from machine learning

Frank Jäkel
Technische Universität Berlin, Berlin, Germany
and Bernstein Center for Computational Neuroscience, Berlin, Germany

Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

and

Felix A. Wichmann
Technische Universität Berlin, Berlin, Germany
and Bernstein Center for Computational Neuroscience, Berlin, Germany

Exemplar theories of categorization depend on similarity for explaining subjects' ability to generalize to new stimuli. A major criticism of exemplar theories concerns their lack of abstraction mechanisms and thus, seemingly, of generalization ability. Here, we use insights from machine learning to demonstrate that exemplar models can actually generalize very well. Kernel methods in machine learning are akin to exemplar models and are very successful in real-world applications. Their generalization performance depends crucially on the chosen similarity measure. Although similarity plays an important role in describing generalization behavior, it is not the only factor that controls generalization performance. In machine learning, kernel methods are often combined with regularization techniques in order to ensure good generalization. These same techniques are easily incorporated in exemplar models. We show that the generalized context model (Nosofsky, 1986) and ALCOVE (Kruschke, 1992) are closely related to a statistical model called *kernel logistic regression*. We argue that generalization is central to the enterprise of understanding categorization behavior, and we suggest some ways in which insights from machine learning can offer guidance.

Intuitive definitions of categorization tend to invoke similarity, in that objects that are similar are grouped together in categories. Within a category, similarity is very high, whereas between categories, similarity is low. Similarity is at the heart of many categorization models. Prototype theories postulate that categorization depends on the similarity of stimuli to an abstracted idea (Posner & Keele, 1968; Reed, 1972), and exemplar theories calculate similarity to memory representations of previously encountered stimuli (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986). A potential problem for these models is that they put the burden of explanation onto the intuitive concept of similarity. Despite serious problems in defining similarity (Medin, Goldstone, & Gentner, 1993), models of categorization continue to rely on similarity.

The appeal of invoking similarity in categorization models stems from the need to generalize. Given a stimulus that has never been encountered before, how can it be categorized correctly on the basis of limited experience with previous stimuli? An easy answer seems to be that a new stimulus is simply categorized in the same way as similar stimuli have been before. Correct generalization to new stimuli thus depends crucially on choosing the right similarity measure. Shepard (1987) famously turned this reasoning around and used generalization to measure similarity. He also tried to deduce a similarity measure such that generalization performance likely would be good (Chater & Vitányi, 2003; Shepard, 1987; Tenenbaum & Griffiths, 2001).

In Shepard's work, the idea of a perceptual space has played a major role. The similarity measure he suggested, often called *Shepard's universal law of generalization* (or simply *Shepard's law*), operates on a mental representation assumed to be a metric space. Shepard's work on generalization and similarity (e.g., Shepard, 1957, 1987) cannot be separated from his work on categorization (e.g., Shepard & Chang, 1963; Shepard, Hovland, & Jenkins, 1961) and multidimensional scaling (MDS; e.g., Shepard, 1962). Since this work, it has become common for perceptual categorization models to assume a perceptual space and to use Shepard's law as a similarity measure in this space (Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986). Exemplar models in particular strongly rely on Shepard's work. These models are very similar to a class of popular tools in machine learning and statis-

F. Jäkel, fjaekel@cs.tu-berlin.de

tics: kernel methods. This observation was first made by Ashby and Alfonso-Reese (1995). Here, we draw parallels between recent progress in kernel methods and exemplar theories of categorization.

Kernel Methods in Machine Learning

In the past, psychological theories of learning and categorization were a major influence for engineers working to build machines capable of intelligent behavior. This influence is signified by the vast engineering literature that has been published on artificial neural networks and reinforcement learning. More recently, an increased interest in kernel methods has arisen in machine learning. Even though these methods can be implemented in simple neural networks, they are usually not psychologically or biologically motivated, but instead are seen to be grounded in statistics and functional analysis. However, as we will show here, researchers in kernel methods are often guided by the same intuitions about similarity and generalization that also guide psychologists in their theories on categorization. Hence, we will argue that theoretical progress in machine learning can also lead to new insights in psychology.

Methods based on kernel ideas are often found to have cutting-edge performance in real-world applications. For example, benchmark data sets for digit recognition are often used to compare the performance of different learning algorithms. The task for a learning algorithm in this setting is to correctly classify handwritten digits it has never been exposed to before, on the basis of experience with a limited number of examples. For a long time, a hand-tuned neural network held the world record on digit recognition benchmarks, until a much simpler kernel method, called a *support vector machine*, was shown to achieve better performance with much less effort on the part of the engineer. Today, support vector machines are found in applications ranging from bioinformatics to machine vision (Cristianini & Schölkopf, 2002).

The successful application of kernel methods to real-world classification problems has led to an explosion of theoretical work in the field of machine learning. Although a considerable amount of theory on artificial neural networks had already been published, progress had been hindered by the complexity of the neural networks used in practice. Kernel methods are built on linear methods and are therefore a lot easier to analyze than nonlinear neural networks (Schölkopf & Smola, 2002).

Preview

In this article, we will demonstrate that the generalized context model (GCM; Nosofsky, 1986) and ALCOVE (Kruschke, 1992), two well-known exemplar models, are very closely related to a machine learning method called *kernel logistic regression* (Hastie, Tibshirani, & Friedman, 2001). The link between the psychological models and the machine learning method is their use of a radial basis function (RBF) neural network (Poggio, 1990; Poggio & Girosi, 1989). In the Similarity section, we will first explain the ideas behind kernel methods and RBF networks. The Categorization section follows, where we will discuss the history of exemplar models, explain their connection to kernel methods, and point to differences in their response rules that affect how Shepard's law enters the models. The kernel view makes the differences between the models more transparent and also allows an easy comparison with methods from machine learning and statistics.

As in psychology, there is a tight relationship between similarity and generalization in machine learning. However, insights from machine learning have shown that, although it is very important to choose the right similarity measure, this is not always enough to guarantee good generalization performance. More specifically, if used naively, kernel methods are prone to overfitting. In the Generalization section, we will discuss the consequences of these insights for exemplar theories of categorization. Exemplar theories have thus far exclusively relied on similarity in order to explain generalization. In fact, a major criticism of exemplar theories has always been that they do not show any form of abstraction, and therefore they are often thought not to be capable of generalization at all. We will show that related kernel methods in machine learning assure good generalization performance with a mechanism called *regularization*, and we will argue that similar mechanisms need to be implemented in psychological models if they are to exhibit good generalization performance. To demonstrate in principle how this could be done, we will analyze ALCOVE's learning algorithm from a regularization perspective. ALCOVE's behavior can be justified on theoretical grounds, thus providing an analysis of the generalization abilities of exemplar theories.

SIMILARITY

Since many categorization models build on similarity, it seems important to understand the processes underlying similarity first. However, quantitative modeling of similarity is a real challenge. Several approaches exist that are based on very different assumptions about how similarity should be represented (Navarro, 2002). Here, we restrict ourselves to a class of similarity models based on geometric representations.

Perceptual Space

The idea that stimuli can be represented as points in a perceptual space underlies MDS and has had a major impact on categorization models. Consider, as an example, the following popular stimuli: circles with a single spoke (Shepard, 1964). Two examples are shown as insets in Figure 1. These stimuli can vary on two obvious dimensions. By varying the radius of each circle, stimuli of different perceived sizes are produced. By varying the angle of the spoke, stimuli of different perceived angles are generated. Figure 1 shows the perceptual space that is defined in this way. Each point in the plane represents the perception of one of the stimuli. The x-axis depicts the perceived angle of a stimulus, and the y-axis depicts the perceived size. The perceived size and angle are of course different from the physically specified size and angle.

To illustrate categorization in perceptual space, we have plotted two clusters of three stimuli each. The first cluster consists of large stimuli with spokes pointing to the

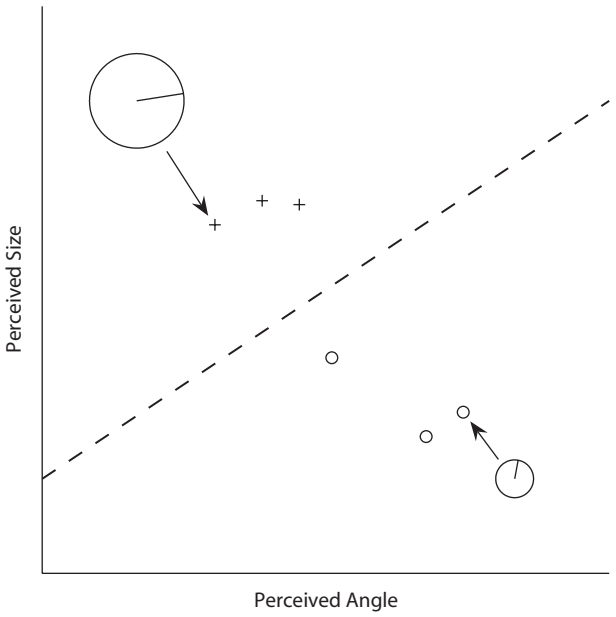

Figure 1. An illustration of a perceptual space. The stimuli are circles with a spoke and can vary on two dimensions. Two artificial categories are depicted, separated by a linear decision boundary.

right (crosses), whereas stimuli in the second cluster are smaller with spokes pointing upward (circles). It is very tempting to draw a line (not necessarily straight) separating the two clusters in order to explain a subject's categorization behavior (Ashby & Gott, 1988). The perceptron, for example, implements a linear decision rule (Rosenblatt, 1958). By comparing the similarity to the mean of the stimuli in each cluster, a prototype classifier also leads to a linear decision boundary (Posner & Keele, 1968; Reed, 1972). Exemplar theories postulate a perceptual space, too, but they can explain more complicated decision rules on the basis of the similarity to all of the stimuli shown to the subject (Kruschke, 1992; Nosofsky, 1986).

With the assumptions of a perceptual space and a metric of this space, MDS can be used to recover underlying dimensions and the configuration of stimuli. To this end, data on the similarity of different stimuli are collected. MDS then tries to embed the stimuli in the metric space so that the similarity relationship in the data is preserved as well as possible: Stimuli that are measured to be highly similar should be very close together in space, and those that are measured to be very dissimilar should have a large distance between them. In practice, MDS with either the Euclidean or the city-block metric in a low-dimensional space works surprisingly well, often leading to interpretable results (Garner, 1974; Shepard, 1964)—despite the fact that geometric approaches to similarity have been heavily criticized (Beals, Krantz, & Tversky, 1968; Jäkel, Schölkopf, & Wichmann, 2008; Tversky, 1977; Tversky & Gati, 1982).

Shepard's Universal Law of Generalization

Before an MDS analysis can be undertaken, an appropriate experimental measure of similarity needs to be found. Following the lead of Shepard (1957, 1987), the categorization literature has often relied on indirect measures of similarity. Shepard (1987) argued that generalization gradients should be used to measure similarity, and this is the stance that is taken in almost all exemplar models. In classical conditioning, generalization gradients are obtained by conditioning on a certain stimulus and measuring the response to related, but different, stimuli (Ghirlanda & Enquist, 2003; Mostofsky, 1965). For example, a dog could be conditioned to salivate in response to a 1000-Hz tone. The generalization gradient is obtained by measuring the salivation of the dog in response to neighboring frequencies. Not surprisingly, the generalization to new stimuli is higher the more similar the new stimuli are to the conditioned stimulus. Intuitively, one would like to explain generalization in terms of psychological similarity, and indeed researchers have tried to obtain measures of similarity that are independent of any generalization behavior (e.g., by integrating just-noticeable differences). In animal studies, however, this proved to be hard, which led Bush and Mosteller (1951, p. 413) to conclude, "Although there are several intuitive notions as to what is meant by 'similarity,' one usually means the properties which give rise to generalization. We see no alternative to using the amount of generalization as an operational definition of degree of 'similarity.'"

If generalization gradients are the best measure to assess similarity, Shepard (1987) reasoned, generalization gradients should be used in the construction of a perceptual space. Applying ordinal MDS to many data sets, Shepard (1987) found a pattern that is now called *Shepard's universal law of generalization*: The amount of generalization decreases (approximately) exponentially with the distance in perceptual space. Shepard (1957, 1958) had earlier used the exponential relationship to explain confusion data in humans. Shepard seems to have thought that confusions that arise in a paired-associate paradigm can, under certain circumstances, be considered as generalization gradients in humans. This is the reason why Shepard's law is not referred to as the "universal law of confusability" (Chater & Vitányi, 2003, p. 352), even though this might be what it is. Experimentally, it is often hard to tell whether generalization gradients really reflect generalization, in the literal meaning of the word, or some degree of confusion in memory or perceptual indiscriminability. Some animal learning theorists have argued that the concept of generalization is superfluous and that discrimination is the only concept that is needed (Brown, 1965), since generalization might only be a failure to discriminate. As a theoretical construct, *generalization* refers to a covert process that leads a subject to respond to a new stimulus in the same way as to a previously learned stimulus, despite the ability of the subject to tell the stimuli apart. This is the meaning that is intended by Shepard (1987), and it is also how generalization gradients are meant to be used in categorization research.

Kernels

In order to model the similarity—that is, the generalization gradient—between two stimuli x and y, we first

interpret x and y as coordinates in an n-dimensional perceptual space. The perceptual distance in this space is usually modeled as

$$d_p(x, y) = \left(\sum_{i=1}^{n} \alpha_i |x_i - y_i|^p \right)^{\frac{1}{p}}. \quad (1)$$

Most often, the distance d_p takes the specific form of either the city-block or the Euclidean distance, with p chosen to be 1 or 2, respectively. The α_is are positive weights that are needed to model the relative importance of the stimulus dimensions, which possibly change with the experimental context. Using Shepard's law, the generalization gradient between x and y is modeled as

$$k(x, y) = \exp\left(-d_p(x, y)^q\right). \quad (2)$$

We refer to the function k that models the generalization gradient as the *similarity kernel*. It is an exponential function of the distance d_p between the two stimuli in perceptual space. Deviating from Shepard's original formulation, the distance is nowadays often modified by taking it to the power of q, in order to give the model more flexibility. In models of categorization, one often finds that q is chosen to be equal to p (Nosofsky, 1990). The similarity kernel is then given as

$$k(x, y) = \exp\left(-d_p(x, y)^p\right) = \exp\left(-\sum_{i=1}^{n} \alpha_i |x_i - y_i|^p \right). \quad (3)$$

With p chosen to be 2, the similarity kernel is called a *Gaussian kernel*. The Gaussian kernel is extremely popular in machine learning. The left panel of Figure 2 shows a Gaussian kernel in two dimensions. Imagine a two-dimensional perceptual space—for example, the perceived size of a circle and the angle of a spoke within it. Stimulus y is fixed at the center of the Gaussian, and the height of the plot depicts the similarity of all other stimuli in the plane to stimulus y. The generalization gradient is rotation invariant—that is, it falls off in the same way in all directions of space. With p chosen to be 1, the similarity kernel is sometimes called a *Laplacian kernel* (in analogy to the Laplacian distribution). This case is depicted in the right panel of Figure 2. The Laplacian kernel is not rotation invariant, so the generalization gradients fall off differently in different directions of space. In particular, similarity fades away more slowly along the stimulus axes. The similarity kernel as defined in Equation 3 has several psychologically and mathematically interesting properties that we explore in detail in two other manuscripts (Jäkel, Schölkopf, & Wichmann, 2007, 2008).[1]

Neural Networks

The similarity kernel forms the basis of many categorization models (Kruschke, 1992; Love et al., 2004; Nosofsky, 1986). Exemplar theories, in particular, make heavy use of the similarity kernel. The ideas that underlie all exemplar models are that stimuli are stored in memory and that new stimuli are categorized on the basis of similarity to the stored exemplars. This idea can be formalized in a neural network. In fact, the ALCOVE model for categorization, which will be discussed in more detail in the Categorization section below, is such a neural network model.

Imagine a cell that after learning is tuned to an exemplar x_i. It will also respond to other stimuli x if they are sufficiently similar to x_i. To model the similarity, we of course use the similarity kernel given in Equation 2. In exemplar models, the similarity to several exemplars $x_1, \ldots, x_N$ is usually a weighted sum of the similarity of each exemplar:

$$f(x) = \sum_{i=1}^{N} w_i k\left(x, x_i\right) \quad (4)$$

The function that this equation computes can be represented graphically as a one-layer neural network. Figure 3 shows such a network with three exemplars. In the neural

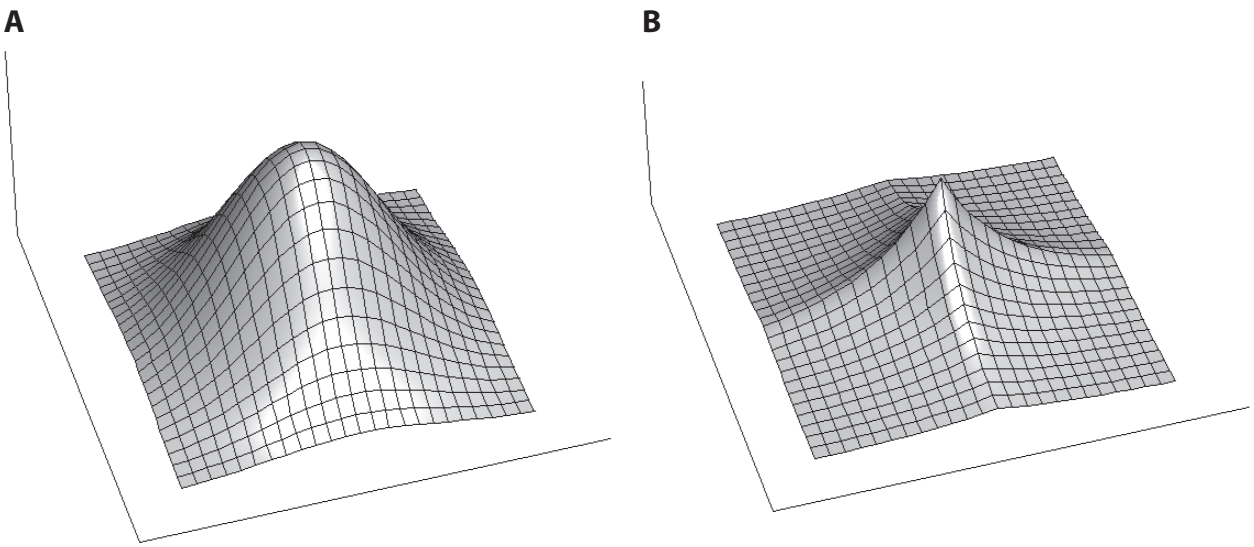

Figure 2. The similarity kernels for different values of p. For $p = 2$, a Gaussian is obtained (A), and for $p = 1$, a Laplacian is obtained (B).

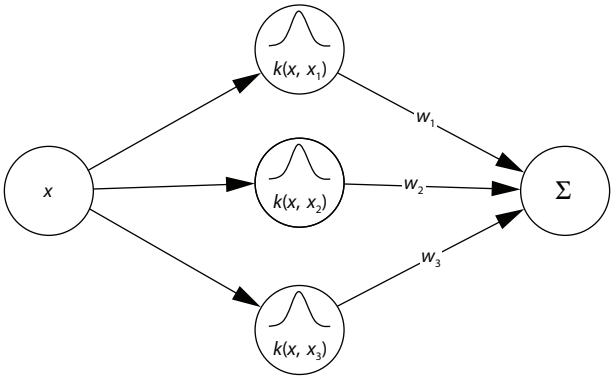

Figure 3. An RBF network calculates a weighted sum over similarity to the various exemplars. A small network with three exemplars is depicted.

network literature, nets with "tuning functions" similar to the similarity kernel are called *radial basis function* nets. These nets have repeatedly been advocated by Poggio and coworkers (Poggio, 1990; Poggio & Bizzi, 2004) as a plausible model for brain function.

For now, it is enough to imagine that all weights are set to 1. Figure 4 shows again the two categories of circles with spokes that were already depicted in Figure 1. The summed similarities to all exemplars of one of the categories (circles) are shown by gray levels. The blacker a region of perceptual space is, the more similar the stimuli in this region are to the exemplars of the category. We have used a Gaussian kernel for illustration. The generalization gradient of the Gaussian can be seen very clearly for the single stimulus close to the dashed category boundary.

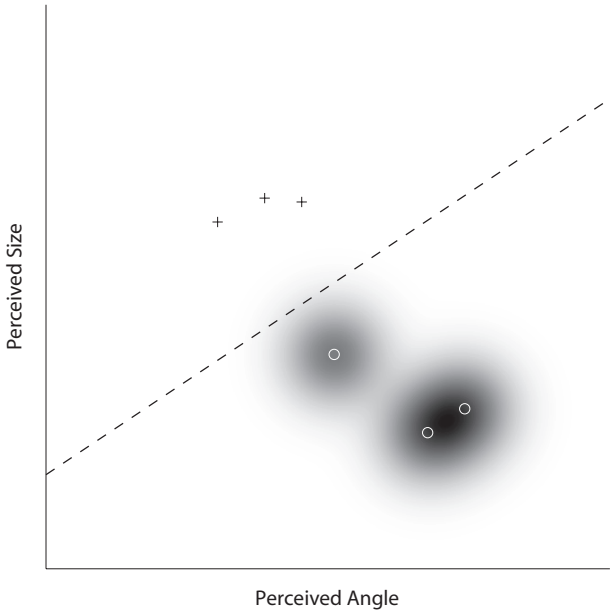

Figure 4. The summed similarity to the exemplars of one category is depicted with gray levels in the perceptual space from Figure 1. The summed similarity is akin to kernel density estimation in statistics.

For the other two stimuli in this category, the generalization gradients overlap quite a bit and form a bigger hump. Regions with a high density of exemplars therefore lead to a high output of the exemplar network (Equation 4), if all of the weights are set to 1. Hence, the output of the network can be interpreted as a measure of category membership or, if appropriately normalized, as an estimate of the probability that a stimulus from the category lies in a certain region of space. In statistics, the same idea is used in so-called *kernel density estimators* (Ashby & Alfonso-Reese, 1995).

Conclusions on Similarity

We have briefly reviewed the idea of a perceptual space and similarity measures based on Shepard's universal law of generalization. We noted that Shepard's law is akin to what is called a *kernel* in machine learning and statistics. Ashby and Alfonso-Reese (1995) have already compared exemplar theories of categorization to kernel density estimators. However, recently, methods based on kernels have attracted a lot of attention in machine learning. In what follows, we will first systematically compare two psychological exemplar theories (GCM and ALCOVE) to a method from machine learning: kernel logistic regression. We will then go on to address the issue of generalization from a machine learning point of view.

CATEGORIZATION

Historically, the first use of the similarity kernel was in an identification task (Shepard, 1957). This identification task is also the theoretical backbone of one of the most prominent exemplar models, the GCM (Nosofsky, 1986). ALCOVE (Kruschke, 1992), a connectionist variant of the GCM, also makes heavy use of the similarity kernel. In the following section, we trace the development of the GCM from the identification task and then present a detailed comparison of the GCM and ALCOVE, highlighting the differences in their uses of the similarity kernel.

Taking a kernel view of exemplar models also reveals their relationship to RBF networks and machine learning methods, especially to the kernel logistic regression method. We believe that the connections between categorization models and their heritage become clearer if they are discussed in the context of the mapping hypothesis, and this is what we will do first.

The Mapping Hypothesis

In two seminal studies, Shepard et al. (1961) and Shepard and Chang (1963) examined the relationship between identification and categorization. In identification tasks, subjects learn to call each stimulus in a set by a unique name. This may be achieved in a paired-associates paradigm in which the experimenter shows the stimuli repeatedly to the subject and asks for the corresponding name. If the subject calls the wrong name, he or she is corrected. During this process of learning, stimuli that are more similar to each other are confused more often. This is not necessarily a result of their perceptual indiscriminability. The original idea in these studies relates back to generalization gradients:

Stimuli are confused because their generalization gradients overlap, not because they cannot be discriminated. But of course, stimuli might also be confused because they have insufficient representations in memory. Over time, a subject will build a better representation of the stimuli and will be able to associate each stimulus with its unique label—at least as far as this is possible, given memory constraints and the discriminability of the stimuli.

A very simple hypothesis about categorization suggests that categorization might work in the same way as this rote-learning mechanism for identification. For each stimulus in the set, the subject has to learn a label, the only difference being that in the categorization task, labels are not uniquely identified with a stimulus. For instance, with two categories there are only two labels, but many more stimuli. In the studies cited, Shepard and colleagues hypothesized that the subjects had the same pattern of confusions as in the identification task: More similar stimuli were confused more often. Therefore, it should have been possible to predict the errors in categorization from the errors in identification. Confusions within a class do not lead to mistakes, but when stimuli from different classes are confused, that is when an error occurs. This was later called the *mapping hypothesis* (Nosofsky, 1986).

It turned out that the mapping hypothesis is not very good at predicting categorization performance, at least not for separable dimensions (Shepard et al., 1961). It does provide a better account for integral dimensions, though (Shepard & Chang, 1963). One explanation could be that categorization is more than just rote learning; some sort of abstraction, such as formation of a prototype (Posner & Keele, 1968), is happening. Another explanation was suggested by Shepard et al.: Even if the underlying representation is the same in both tasks, a subject's attention might be directed to different dimensions of the stimuli in the two tasks. This idea is formalized in Nosofsky's GCM (Nosofsky, 1986). In Nosofsky's experiments, this model proved to provide a better account of categorization performance than did the simple mapping hypothesis.

The MDS Choice Model

Since an identification model is the starting point for the GCM, it is natural to describe the model for the identification task first. In each trial, a subject has to choose a response from a set of possible responses. A very simple and widely used model for choice behavior in general was investigated by Luce (1959, 1963, 1977). The model has close connections to the method of paired comparisons and to logistic regression (Bradley, 1976; David, 1988). For an identification task, a model in the same spirit was first discussed by Shepard (1957). The probability of answering with response r_i when the stimulus was x_j is given by Luce's (1959) well-known choice rule

$$P(r_i \mid x_j) = \frac{\pi_{ij}}{\sum_{k=1}^{N} \pi_{kj}}, \tag{5}$$

where the number of stimuli and responses is N. In Shepard's identification model, π_{ij} is interpreted as the similarity between the stimuli x_i and x_j (with π_{ij} being positive). This basic model is usually supplemented with response bias terms that we will ignore for simplicity.

If no additional structure is assumed for the terms π_{ij}, nothing is gained from this formulation. Shepard (1957) assumed that π_{ij} is a monotonically decreasing function of the distance between the stimuli x_i and x_j in a psychological space. To make the model feasible, he also assumed that the psychological space is Euclidean and that the relationship between similarity and distance is exponential. Shepard's suggestion was essentially to use Equation 2, $\pi_{ij} = k(x_i, x_j)$.[2] This model has come to be known as the *MDS choice model* (Nosofsky, 1986).

For example, imagine the psychological space to be two-dimensional. Instead of having to estimate the N^2 probabilities of confusion, only the $2N$ coordinates of the stimuli have to be estimated. Since Shepard (1957) assumed the distances in the similarity kernel in Equation 2 to be Euclidean, he could use classical MDS to recover the coordinates. Later, he used his ordinal scaling method to get independent support for the shape of the similarity kernel (Shepard, 1965, 1987). Today, the similarity kernel is often assumed to be known, and the coordinates in the multidimensional space are routinely estimated by using maximum likelihood (Nosofsky, 1986).

By a simple reparameterization, $\pi'_{ij} = \log \pi_{ij}$, it is easy to see that Luce's choice rule (Equation 5) is identical to the multinomial logit model (Train, 2003):

$$P(r_i \mid x_j) = \frac{\exp(\pi'_{ij})}{\sum_{k=1}^{N} \exp(\pi'_{kj})}. \tag{6}$$

We can therefore interpret the MDS choice model in two ways. It can be interpreted as plugging the similarity of the stimuli i and j—as given by π_{ij}—into Luce's choice rule. We can also interpret it as calculating the logarithm of the similarity π'_{ij} and plugging this expression into a logit model. If we take the similarity kernel from above (Equation 3) as our measure for stimulus similarity, $\pi_{ij} = k(x_i, x_j)$, the full MDS choice model reads

$$P(r_i \mid x_j) = \frac{k(x_i, x_j)}{\sum_{k=1}^{N} k(x_k, x_j)} = \frac{\exp\left(-d_p(x_i, x_j)^p\right)}{\sum_{k=1}^{N} \exp\left(-d_p(x_k, x_j)^p\right)}.$$

Since the similarity kernel in Equation 3 is an exponential of the pth power of d_p, the distance in psychological space, the two interpretations of the MDS choice model are (1) Luce's choice rule as applied to the similarity of x_i and x_j in perceptual space, or (2) a logit model on the pth power of the distance d_p between x_i and x_j in perceptual space. Hence, when interpreted as a logit model, the MDS choice model does not make use of the similarity kernel, but rather is a logit model on the pth power of d_p.[3]

The Generalized Context Model

Using the mapping hypothesis, it is straightforward to work out the probabilities for the category responses once

the identification model is specified. A number of categories $C_1, \ldots, C_M$ with associated responses $R_1, \ldots, R_M$ are defined so that each possible stimulus $x_1, \ldots, x_N$ belongs to exactly one of the categories. The probability of observing response R_m given the stimulus x_j is then

$$\begin{aligned} P\left(R_m \mid x_j\right) &= \sum_{x_i \in C_m} P\left(r_i \mid x_j\right) \\ &= \frac{\sum_{x_i \in C_m} \pi_{ij}}{\sum_{k=1}^{N} \pi_{kj}} \\ &= \frac{\sum_{x_i \in C_m} k\left(x_j, x_i\right)}{\sum_{m=1}^{M} \sum_{x_i \in C_m} k\left(x_j, x_i\right)}. \end{aligned} \tag{7}$$

For simplicity, we have again ignored response biases. Following the MDS choice model, Nosofsky identified the similarity measure in Equation 2 with $\pi_{ij} = k(x_i, x_j)$ (Nosofsky, 1986, 1987). He called this model the *generalized context model* because, with a certain choice of similarity kernel, it can be seen as the continuous generalization of an earlier exemplar model with binary features that was called the *context model* (Medin & Schaffer, 1978). Note that the identification model is recovered if every stimulus has a unique label—that is, there is a different category for each stimulus. In order to link identification and categorization data, Nosofsky had to allow different attention weights for the dimensions in the different tasks (see Equation 1).

From a statistical viewpoint, the GCM can be seen as a multinomial logit model (see Equation 6), just as the MDS choice model is. Let us introduce the shorthand

$$f_m(x) = \sum_{x_i \in C_m} k\left(x, x_i\right)$$

for the sum of the similarities. This is a special RBF network (see Equation 4) with the weights for the exemplars in a category set to 1 and the other weights set to 0—but note that a later formulation of the GCM explicitly includes weights for exemplars (Nosofsky, 1992). Now, to see that the GCM can be interpreted as a logit model, consider the case with only two categories. The GCM (Equation 7) then simplifies to

$$\begin{aligned} P\left(R_1 \mid x_j\right) &= \frac{f_1\left(x_j\right)}{f_1\left(x_j\right) + f_2\left(x_j\right)} \\ &= \frac{\exp\left\{\log\left[f_1\left(x_j\right)\right]\right\}}{\exp\left\{\log\left[f_1\left(x_j\right)\right]\right\} + \exp\left\{\log\left[f_2\left(x_j\right)\right]\right\}} \\ &= \frac{1}{1 + \exp\left(-\left\{\log\left[f_1\left(x_j\right)\right] - \log\left[f_2\left(x_j\right)\right]\right\}\right)} \\ &= \text{logistic}\left[\log f_1\left(x_j\right) - \log f_2\left(x_j\right)\right], \end{aligned} \tag{8}$$

where we have used the definition of the logistic function: $\text{logistic}(x) = 1/[1 + \exp(-x)]$. Hence, the GCM for two categories can be seen as a logit model on the logarithm of the summed similarities.

ALCOVE

Inspired by the success of the GCM, and probably also by the general excitement about neural network models at the time, Kruschke (1992) developed a connectionist variant of the GCM. As crucial ingredients for his model, he identified both a similarity measure that can be given a tuning-curve interpretation and the attention weights that Nosofsky used. He formulated the model as a network and added a back-propagation learning algorithm to account for the adjustment of the attention weights—hence, the name *ALCOVE* ("attention learning covering map").

There are important differences between ALCOVE and the GCM. In the GCM as given in Equation 7, the similarities to all exemplars are simply added up. In ALCOVE, the output neurons collect a weighted sum of all hidden neurons. Assume once again M categories $C_1, \ldots, C_M$, with one output neuron for each category. The activation f_m of the neuron that is responsible for category C_m is defined as a weighted sum of the activation of the hidden-layer neurons:

$$f_m(x) = \sum_{i=1}^{N} w_{mi} k\left(x, x_i\right).$$

Each output neuron m has its own weights that are collected in a vector w_m. Each output neuron is an RBF network with a kernel given by the similarity measure (see Equation 4 and Figure 3). Whereas the GCM uses Luce's choice rule, ALCOVE uses the logit response rule (Equation 6) directly on the weighted similarities to the exemplars:

$$P\left(R_m \mid x_j\right) = \frac{\exp\left[f_m\left(x_j\right)\right]}{\sum_{m=1}^{M} \exp\left[f_m\left(x_j\right)\right]}.$$

An identification task can be set up by having as many categories as stimuli, but contrary to the original GCM, the identification and the categorization task cannot be linked by the mapping hypothesis. This conceptual difference between the GCM and ALCOVE should not be overlooked, because the mapping hypothesis provided the main motivation for the GCM—even though it is also important to note that the GCM with changing attention weights and response-scaling mechanisms also cannot be linked with the mapping hypothesis (see Nosofsky & Zaki, 2002).

For obvious reasons, ALCOVE is called *kernel logistic regression* in machine learning and statistics (Hastie et al., 2001). It is an RBF network combined with a logit model. In the important case of just two categories, ALCOVE reduces to

$$\begin{aligned} P\left(R_1 \mid x_j\right) &= \frac{\exp\left[f_1\left(x_j\right)\right]}{\exp\left[f_1\left(x_j\right)\right] + \exp\left[f_2\left(x_j\right)\right]} \\ &= \text{logistic}\left[f_1\left(x_j\right) - f_2\left(x_j\right)\right]. \end{aligned} \tag{9}$$

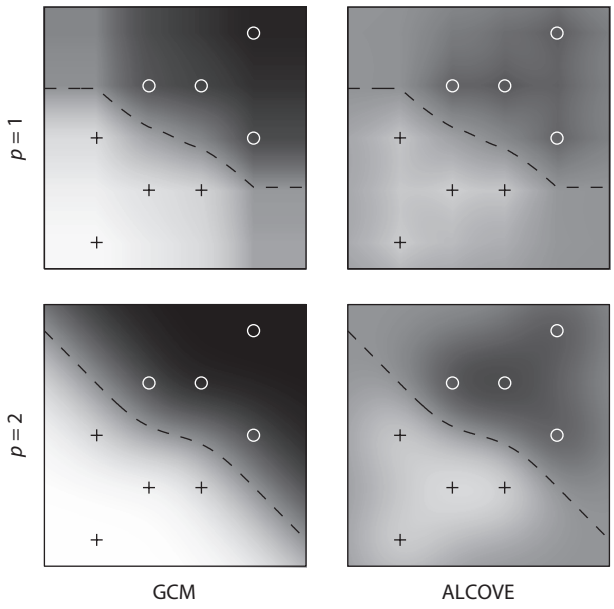

Figure 5. A comparison of GCM and ALCOVE for city-block ($p = 1$) and Euclidean ($p = 2$) metrics. The circles and crosses depict exemplars from two classes. For both categorization models, the attention and exemplar weights are set to 1. The grayscale shows the response probabilities that would be obtained for a new stimulus at each position. White areas are classified as "cross" with probability 1, whereas black areas belong to the other class. The dashed line depicts the equal-probability contour, which is the same for the GCM and ALCOVE. Outside the generalization gradients of the exemplars, the models make very different predictions.

The first term in the logistic function is a nonparametric measure for the degree that the stimulus belongs to the first category. The second term does the same for the second category. The logistic function is simply applied to the difference of the two category scales.[4]

Comparison of GCM and ALCOVE

Figure 5 shows a comparison between ALCOVE and the GCM for a simple two-category classification task. For both models, the attention and exemplar weights are set to 1. Both models are depicted with the city-block ($p = 1$) and Euclidean ($p = 2$) metrics. On the equivocality contour, the summed similarity to all exemplars of one class equals the summed similarity to all exemplars of the other (Ashby & Maddox, 1993). Since we assume that subjects are unbiased, the probability for a subject to respond with a given class is one half. The equivocality contours are shown as dashed lines. First, note that these lines are the same for the GCM and ALCOVE. ALCOVE performs logistic regression on the difference of the summed similarities in Equation 9, and therefore the choice between city-block and Euclidean metrics only makes a difference close to the exemplars. Far away from them—that is, beyond the generalization gradients of all exemplars—categorization performance drops to chance level, because the summed similarities go to 0. ALCOVE and the GCM make very different predictions on stimuli that are outside the generalization gradients for the exemplars. In Equation 8, the GCM operates on the log of the summed similarities. Therefore, points that are clearly on one side of the decision boundary are categorized more easily in the GCM, because the ratio of f_1 to f_2 can still be big, even if the absolute difference between them is small. (Note that the exact behavior of the model also depends on the exponent p.)

Because it shows no generalization beyond its generalization gradients, one could say that ALCOVE behaves like Spence's classic model for discrimination learning (Spence, 1937) and thus shows no "true" categorization behavior. In contrast to ALCOVE, the GCM is capable of categorization beyond its generalization gradients, as already noted by Nosofsky (1991b). Intuitively, if a subject has really learned to categorize the two stimuli—as opposed to only discriminate them—one would expect stimuli clearly on one side of the decision boundary to be categorized easily. This is the criterion used in animal studies to define categorization (e.g., Ohl, Scheich, & Freeman, 2001). Although this may sound like subjects would have to implement an explicit rule in order to behave accordingly, the GCM can show this behavior without representing a decision boundary explicitly.

Conclusions on Categorization

We have traced the history of exemplar models and the similarity kernel back to the work of Shepard (1957, 1958) on generalization gradients and identification tasks. A bit later, the idea to link identification and categorization via the mapping hypothesis was first tested experimentally (Shepard & Chang, 1963; Shepard et al., 1961). In parallel, Shepard (1962) developed his ideas on ordinal MDS. Using the concept of attention weights, Nosofsky (1986, 1987) was able to assemble all of these parts into a working model of categorization and to link it to the existing literature on exemplar-based categorization (Medin & Schaffer, 1978). A bit later still, Kruschke (1992) suggested a connectionist variant of the GCM that is closely related to RBF networks (Poggio, 1990) and kernel logistic regression (Hastie et al., 2001). We have seen that both the GCM and ALCOVE are based on the logit rule and the use of the similarity kernel, but with important differences. In contrast to the original GCM, ALCOVE does not use the mapping hypothesis. Also, ALCOVE does not show much categorization beyond its generalization gradients. This demonstrates that the way the generalization gradients enter the response rule in a categorization model has an influence on how new stimuli will be classified. Whether this classification is likely to be correct, however, is also influenced by other factors that we will discuss in the next section.

GENERALIZATION

Building on Shepard's (1987) work, exemplar theorists have basically completely identified similarity with generalization. However, we will argue that exclusive reliance on similarity does not necessarily lead to good generalization performance. Additional statistical considerations need to be taken into account. This will not come as a sur-

prise to the critics of exemplar theories, who have always doubted that merely remembering exemplars can lead to proper categorization. This does not, however, mean that exemplar models cannot generalize. Quite to the contrary: In machine learning, kernel methods are among the most successful tools, precisely because they are known to generalize well—if they are used wisely. We will show how exemplar theories can be made to reliably extract the structure underlying a category. To this end, we will discuss how kernel methods in machine learning and statistics deal with the problem of generalization.

Kernel Density Estimation

The category learning problem is often phrased as a density estimation problem (Aizerman, Braverman, & Rozonoer, 1964; Ashby & Alfonso-Reese, 1995; Fried & Holyoak, 1984; Nosofsky, 1990). Imagine two classes in which exemplars from each category are drawn from a probability density function that completely determines the distribution of features within each category. If a learner knew the distribution of features within a category, he or she could examine the features of a new stimulus and assign it to the category with the highest likelihood of having generated this pattern of features. Hence, learning to categorize could mean learning the distribution of features.[5]

The upper left panel in Figure 6 gives an example. It shows a two-dimensional stimulus space. The features of each stimulus are represented by coordinates in the space, which could be either a physically specified or a perceptual space. The difference of the densities of the two overlapping normal distributions is indicated by the gray levels. The darker regions correspond to high-density regions of one of the classes, whereas the lighter regions correspond to high-density regions of the other class. Probabilistic category structures similar to this one are frequently used in experiments (Ashby & Gott, 1988; Ashby & Maddox, 1992; Fried & Holyoak, 1984; McKinley & Nosofsky, 1995, 1996). We have drawn 10 exemplars from each of the distributions for illustration (circles and crosses).

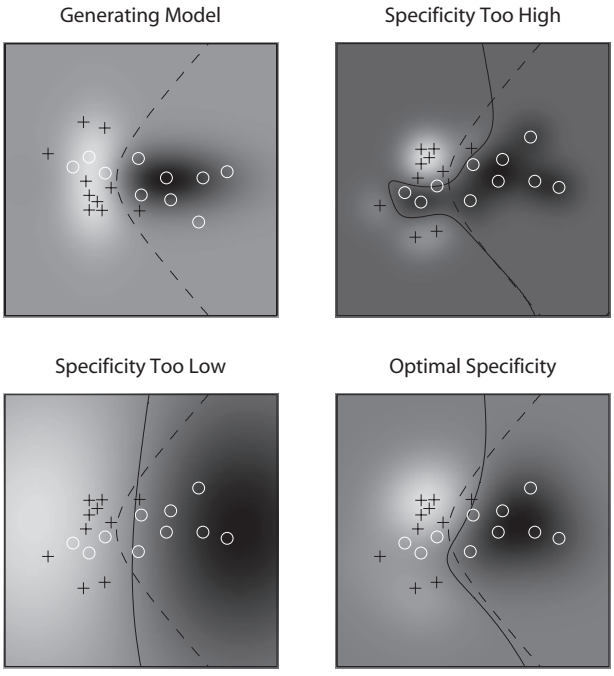

Figure 6. A two-class problem. Samples from two classes (crosses and circles) are generated from two overlapping normal distributions. In the upper left panel, the differences between the distribution densities are shown as gray levels, and the optimal decision boundary is shown as a dashed line. The other panels show kernel density estimates of the two classes with varying specificity of a Gaussian kernel, as well as the corresponding decision boundaries (solid lines).

Since we know the distributions of the two classes, we can calculate the optimal decision boundary between them, which for two normal distributions is generally quadratic (Ashby & Maddox, 1993). The optimal decision boundary is shown by the dashed lines in Figure 6. A subject trying to maximize performance—that is, correct responses—should place the decision criterion along the optimal decision boundary. On one side of the boundary, the subject should always choose one category label, and on the other side he or she should always choose the other. This sharp boundary without probabilistic responding will give the best generalization performance. However, subjects may not respond deterministically, and different models make different assumptions about how probabilistic decisions are (Ashby & Maddox, 1993). In the following discussion, we will ignore this additional complication and only talk about generalization performance under the assumption of (almost) deterministic responding. However, evidence exists that subjects do respond deterministically under certain circumstances (Ashby & Gott, 1988), and exemplar models can be and have been adapted to account for this (McKinley & Nosofsky, 1995; Nosofsky, 1991a). In fact, a considerable amount of debate has recently revolved around so-called *response-scaling mechanisms* that allow the GCM to respond more deterministically (Navarro, 2007; Nosofsky & Zaki, 2002; Smith & Minda, 1998).

Even though we can calculate the optimal decision boundary for the example in Figure 6, the subject cannot know the true distributions, because the subject has only observed a limited number of exemplars from these two categories. Therefore, the subject cannot respond optimally. One possible strategy in this case is to try to estimate the two category distributions from the observed exemplars and choose a decision boundary that would be optimal for the estimated category distributions. This can be done by assuming a particular parametric family for the category distributions and trying to estimate the parameters. For example, a category learner may assume that the distributions are normal, in which case he or she must estimate means and covariances (Fried & Holyoak, 1984). This strategy will work well if the underlying category structure to be learned is indeed approximately normal. A more flexible category learner, however, would try to avoid making very specific assumptions about the unknown distributions.

Exemplar models have been compared with the more flexible (nonparametric) kernel density estimators (Ashby & Alfonso-Reese, 1995). In the simplest exemplar model,

each data point is replaced by a kernel function—for example, a Gaussian kernel. As explained above, the summed similarity to all exemplars from one class can be seen as a density estimate. The upper right panel of Figure 6 shows just such a kernel density estimate. Black areas have a high similarity to the exemplars of one of the classes (circles), and white areas have a high similarity to exemplars of the other class (crosses). For the density estimator, a high similarity to exemplars from one class translates into a high likelihood that a new stimulus that falls in this region belongs to the corresponding category. The black solid line indicates the equivocality contour, where the similarity to the exemplars from one class equals the similarity to those from the other. This equivocality contour could be used as a decision boundary.

Finding the Right Kernel

In the example in the upper right panel of Figure 6, the specificity is chosen to be too high—that is, the width of the similarity kernel is chosen to be too narrow. New stimuli are essentially categorized in the same way as the most similar past exemplar. If this exemplar happens to lie on the wrong side of the optimal decision boundary, it is very likely that a wrong decision will be made. The similarity kernels of different exemplars hardly overlap, and therefore generalization to new stimuli is poor. The decision boundary that the category learner chooses is able to categorize all of the past exemplars perfectly, but only because the idiosyncrasies of this particular set of exemplars have been learned. This is called *overfitting*: The learner has not learned anything about the structure of the categories, but instead has only learned the labels and the exemplars by heart. The bottom left panel shows the opposite case, in which the specificity of the kernel is chosen to be too low. A wide similarity kernel means that exemplars far away from a new stimulus can influence the guess of which category it belongs to. In this case, the resulting decision boundary will also be very different from the optimal one. Hence, in order to assure good generalization performance, it is important to choose the width of the similarity kernel to be appropriate for the problem and the sample size at hand. The lower right panel of Figure 6 shows the decision boundary that results from a well-chosen kernel width.

Sometimes it may be possible for a subject to choose a reasonable kernel width before seeing the first exemplars, but in general the specificity and the relative contributions of the attention weights have to be adapted by learning as well (but see Lamberts, 1994, for effects of background knowledge on specificity). In machine learning, choosing a kernel and setting its parameters are considered to be model selection problems. In psychological models, the form of the kernel is given, but its parameters may be adapted during learning. ALCOVE adapts its attention weights during learning, but Kruschke (1992) did not directly address generalization performance. In machine learning, a common way to choose the best parameters for the kernel is by using cross-validation procedures (see Pitt, Myung, & Zhang, 2002, for an overview of model selection). Instead of trying to minimize the error on all of the known exemplars—which can always be driven to zero by choosing a narrow enough kernel, as seen in Figure 6—one tries to obtain an estimate of the generalization error by repeatedly splitting the data into a training and a test set. For a certain setting of the specificity, one asks how well the model uses the exemplars in the training set to predict the category membership of the those in the test set. The parameter value that gives the lowest estimated generalization error is the one that will be used. This procedure was applied to obtain the specificity value for the lower right panel of Figure 6. We are not suggesting that human subjects use a procedure akin to cross-validation, but we do want to point out that from a normative point of view, the choice of similarity kernel is crucial. If the similarity kernel is adaptable, subjects should then pay close attention to their generalization performance while changing it.

Overfitting With Exemplar Weights

The problem of overfitting becomes even more pressing with the introduction of exemplar weights into categorization models. Both ALCOVE and a later version of the GCM have such weights (Kruschke, 1992; Nosofsky, 1992). It is desirable to introduce these weights for several reasons. For instance, it is unlikely that subjects will be able to remember all exemplars and to attach the same weight to each of them. Probably there will be frequency and recency effects, as well as forgetting. Some of the exemplars will be more representative of a category than others and may get a greater weight. Furthermore, from a statistical point of view, the exemplar weights introduce a greater flexibility that makes it possible to learn more complicated decision boundaries. However, if these exemplar weights can be modified by learning, it follows that each exemplar will have its own free parameter—an almost sure recipe for overfitting (Pitt et al., 2002).

Recall that ALCOVE is built on an RBF network. The RBF network implements a function by expressing it as a weighted sum of kernel functions centered on the exemplars:

$$f(x) = \sum_{i=1}^{N} w_i k\left(x, x_i\right). \tag{10}$$

As noted before, Poggio has suggested RBF networks as a biologically plausible model for brain function (Poggio, 1990; Poggio & Bizzi, 2004). A common view is to see the brain as a supervised learning machine. The network gets some input, calculates a function, and receives feedback on the errors it has made. This basic setup is used in most artificial neural network approaches and underlies the back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986). Hence, learning means to adapt the weights in Equation 10 such that the error is minimized. The function that ALCOVE's back-propagation learning algorithm is trying to learn outputs a $+1$ for one of the categories and a -1 for the other.

It is possible to give the optimal weights for this function without running a back-propagation algorithm. Let f be a vector of the function values $f_i = f(x_i)$ that we want the function to take on the exemplars. Let K be a matrix with entries $k(x_i, x_j)$ in the ith row and jth column. This matrix is

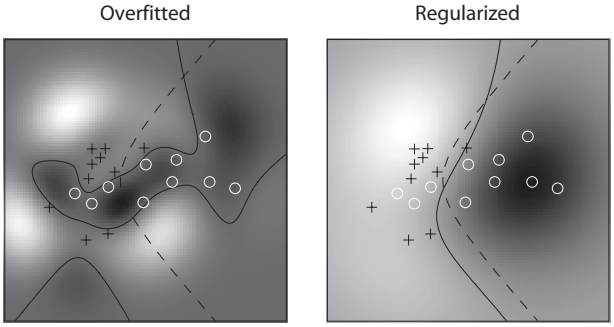

Figure 7. Unless regularized, an exemplar model with exemplar weights will overfit. The left and right panels use the same similarity kernel with the same specificity. The only difference is whether regularization was used or not. The amount of regularization was chosen to give the best possible generalization performance for the chosen specificity.

called the *kernel matrix*. Let w be the vector of weights that we seek to implement the function. With this notation, we can rewrite the neural network from Equation 10 in matrix notation as $f = K \cdot w$. In the Similarity section above, we mentioned that the similarity kernel has nice mathematical properties that are explained in a recent tutorial (Jäkel et al., 2007). The most important property is that the kernel is usually positive definite. For a positive definite kernel k, the matrix K is positive semidefinite. Most kernels used in categorization models are even strictly positive definite, and therefore invertible. Hence, we can find unique weights such that the function f makes no error at all on the exemplars: $w = K^{-1} \cdot f$. The resulting function f for the exemplars from Figure 6 is shown in the left panel of Figure 7. This function outputs a $+1$ for all exemplars from one of the categories and a -1 for all exemplars of the other category. Such a perfect (but probably useless) solution can always be found, independent of the specific exemplars and independent of the specificity of the kernel.

The fitted function does not capture the underlying regularity well; the reason is that, by freely allowing the weights to be adapted, we can override the similarity-based categorization. The exemplar weights defeat the purpose of introducing a similarity measure for the stimuli. The similarity measure is introduced because similar stimuli should be treated similarly, since very similar stimuli are very likely to belong to the same category. However, the exemplar weights can be adjusted in a way to allow even very similar stimuli to belong to different categories without interfering with each other. Imagine the case in which we only have two very similar stimuli x_1 and x_2 that have very different function values, $f(x_1) = 1$ and $f(x_2) = -1$. Say that their similarity is .99 and the self-similarity is 1. In order to make the network (Equation 10) output the right values, the small difference of .01 between their similarity and their self-similarity needs to be compensated by large weights of 100 and -100, respectively.

Regularization

One way to deal with overfitting in neural networks is regularization (Bishop, 1995; Orr & Müller, 1998). This is the approach advocated by Poggio and coworkers (Poggio, 1990; Poggio & Bizzi 2004; Poggio & Girosi, 1989; Poggio & Smale, 2003), and it is also used for kernel logistic regression (Hastie et al., 2001). The basic idea in regularization is that weights are not allowed to become too big. Since large weights can override the similarity-based categorization, the weights should be as small as possible. This is achieved by trading off the error that the classifier makes with the size of the weights. Recall that learning in the neural network setup means finding weights w such that a loss function $L(w)$ is minimized. Let us call the error that the classifier makes on the training exemplars $c(w)$. The penalty for large weights is called a *regularizer*, and we denote it $\Omega(w)$ here. With this notation, the loss function that a regularized RBF network minimizes becomes

$$L(w) = c(w) + \Omega(w). \tag{11}$$

The regularizer reflects a "complexity" constraint on the function that the network implements. It is good if the available data are fitted well, but this should not be done at all costs. The fitted function should not be too complicated, because complicated functions are more likely to overfit. Most model selection criteria trade off goodness of fit versus model complexity (Pitt et al., 2002).

The right panel of Figure 7 shows the same categorization problem as before, but this time regularization techniques were used. The gray levels code a function f of the form in Equation 10 that minimizes $L(w)$ in Equation 11, with c chosen to be squared error and Ω chosen to be linear in the squared length of the vector w. For this loss function and several other interesting ones, the optimal weights w are unique and can be found easily (Schölkopf & Smola, 2002). Because of the regularization, the category learner did not try to fit the available exemplars perfectly, but instead traded off goodness of fit with the penalty term. Clearly, the model is closer to the optimal decision boundary than one without regularization. Intuitively speaking, the regularizer penalizes the large exemplar weights necessary to make two similar stimuli have different category labels (Jäkel et al., 2007).

It should be emphasized again that the exemplar network by itself does not guarantee good generalization performance. After all, the exemplar network can always implement a function that can fit all exemplars perfectly—no matter what they look like. It is the joint choice of the kernel and the regularizer that determines the generalization performance of the network. The kernel captures some assumptions about the category structure, and the regularizer penalizes greedy optimization of goodness of fit. Different problems will require different kernels and different regularizers. In machine learning, the kernel and the regularization parameters are usually chosen by cross-validation.

Learning a Category With ALCOVE

The learning algorithm of ALCOVE greedily tries to minimize the classification error on the exemplars. We have shown above that for such models, there is a danger of overfitting. If ALCOVE is shown the same exemplars over and over again, its back-propagation algorithm can find a

solution that categorizes these exemplars perfectly—no matter what the category structure is. Since ALCOVE has been quite successful in describing subjects' learning curves in various categorization tasks, this raises the question of whether human subjects also overfit. Considering that humans seem to categorize new stimuli reliably in everyday life, this seems unlikely. Perhaps humans do overfit in experiments performed in the laboratory, however, and laboratory experiments are what exemplar theories try to model.

In most of the earlier experiments in favor of exemplar theories, subjects were shown a small number of exemplars over and over again. Remember that in the classic work of Shepard et al. (1961) and Shepard and Chang (1963), the original motivation was to see whether categorization could be described as mere rote learning of labels—this was called the *mapping hypothesis*. The GCM, too, was set up in order to link categorization with a rote-learning identification task, and the accompanying experiments used only a small number of stimuli (Nosofsky, 1986). Also, the experiments by Medin and Schaffer (1978), which are widely seen to provide good evidence for exemplar theories, have been criticized on the grounds that they used only few stimuli and poorly differentiated categories (Smith & Minda, 1998, 2000). Hence, subjects are perhaps encouraged to adopt an exemplar memorization strategy in experiments, even though they may not do so in everyday categorization. Some of the categories used in psychological experiments have so little structure that rote learning of exemplars is in fact the *only* strategy that allows subjects to solve the task (Feldman, 2000; Shepard et al., 1961).[6] If transfer items are presented in such experiments, they are only used to assess the predictions of the model (see, e.g., Medin & Schaffer, 1978; Nosofsky, 1986). There is usually no right or wrong answer for the subjects; therefore, no rational strategy exists to which a subject's behavior could be compared in order to assess his or her generalization performance.

Other experiments have explicitly compared human performance with that of an ideal observer (Ashby & Gott, 1988; Ashby & Maddox, 1992; Fried & Holyoak, 1984; McKinley & Nosofsky, 1995, 1996). Those studies used overlapping probabilistic categories like the one shown in Figure 6. This scenario is perhaps more akin to natural category learning. Contrary to many categories in psychological experiments, natural categories have a structure, and presumably it is this structure that humans learn when they learn a category. Rosch and colleagues (Rosch & Mervis, 1975; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) have argued that, on the basic level, the stimuli within a natural category share perceptual properties, and that the distribution of the properties of a category is not completely random, but also is not deterministically defined by necessary and sufficient conditions. Since very little is known about the actual structure of natural categories, we may choose to use categories like the one shown in Figure 6 as a proxy. This has the advantage that the number of possible exemplars is infinite, so that subjects never encounter the same exemplar twice. Furthermore, there is an objective way to assess a subject's generalization performance. Clearly, under these conditions, a strategy that simply remembers all encountered exemplars seems unreasonable. Nevertheless, some of the studies above have successfully fitted exemplar models to human responses (McKinley & Nosofsky, 1995, 1996).

Interestingly, in this scenario with overlapping probabilistic categories, ALCOVE will not overfit as easily, and its behavior results in exemplar networks that are regularized. A subject encounters a new stimulus but does not know its category label. He or she predicts the category of the stimulus on the basis of previous exemplars and then receives feedback about the true category label. It is reasonable to set the exemplar weights to 0 before an exemplar has been encountered. After ALCOVE is given the true category label of a new exemplar, it may be necessary to assign a large weight to the exemplar in order to output the correct label. How much the weights are allowed to change is determined by the learning rate parameter in ALCOVE. If the learning rate does not allow big changes in the weights, this is akin to regularization that penalizes large weights in order to avoid overfitting. Limiting the influence of individual points has a regularizing effect by increasing the *stability* of the solution. Indeed, solutions that are stable, in the sense of not depending too strongly on any individual training point, can be shown to generalize well with high probability (Bousquet & Elisseeff, 2002; Poggio, Rifkin, Mukherjee, & Niyogi, 2004). Note also that the feedback that the subject receives is a direct measure of the generalization error—similar to cross-validation. The prediction error is a direct measure of the subject's generalization performance, because each stimulus is a new one that has never been encountered before. This contrasts with experimental procedures in which the same stimuli are shown over and over again. Therefore, in the case in which each stimulus is new, ALCOVE does not try to minimize the error on past exemplars, but instead the prediction error on new exemplars. Early stopping in artificial neural networks is used for the same reason (Orr & Müller, 1998).

As mentioned before, regularization is used in machine learning to improve the generalization performance of kernel methods, and exemplar models in psychology should also make sure that they can generalize to new exemplars. The analysis of ALCOVE's learning algorithm shows that it is not hard to come up with psychologically plausible mechanisms akin to regularization. It would be premature to claim that humans implement regularization by choosing a small learning rate parameter. However, for ALCOVE, the learning rate parameter is indeed crucial for the model's generalization performance.

Prototype Theories, Exemplar Theories, and Generalization

The prototype versus exemplar debate is usually framed in terms of mental representations. Subjects may store a summary representation of a category, or they may store exemplars of the category. The debate can also be seen as being about which assumptions a category learner makes about the category he or she is learning (Ashby & Alfonso-Reese, 1995; Ashby & Maddox, 1993). These assumptions will determine the learner's generalization performance. Prototype theories make very strong assumptions

about the category structure: The whole category structure is summarized by the prototype. This leads to good generalization performance, even with only a few trials of learning, if the category structure to be learned is really so simple. Exemplar theories with exemplar weights, like ALCOVE, are at the other extreme. They are very flexible category learners and can learn more complicated category structures. There is a trade-off between how restrictive the assumptions of a categorization model are and the complexity of the categories it can learn. Briscoe and Feldman (2006) have recently explored the consequences of this insight for human learning. However, even though exemplar models can learn more complicated category structures, they nonetheless do make some assumptions about the category structure. These assumptions are only given implicitly by the choice of kernel and the way that the learning algorithm sets the weights. Therefore, it is a lot harder to say what these models learn from the exemplars. Nevertheless, even if they do not abstract anything from the data, they are able to learn something about the structure of the category that then enables them to generalize to new stimuli. In any case, the assumptions of the category learner need to be matched to the category structures that he or she wants to learn; otherwise, generalization performance can be extremely poor.

Unless all of the evidence in favor of exemplar theories is completely misleading because of small and ill-defined categories in the experiments (Smith & Minda, 1998, 2000), one would hope that exemplar theories can scale up to real-world categorization behavior. Kernel methods in machine learning have already proved to be successful in real-world applications. And as we have shown, these methods build on intuitions similar to those of exemplar theories. In fact, kernel methods often outperform other methods with more restrictive assumptions, such as prototype classifiers, on real-world data sets (Schölkopf & Smola, 2002). This could suggest that the restrictive assumptions of prototype theories are not met for natural categories and that more flexible mechanisms, as implemented in exemplar models, are needed to deal with real-world categories. However, it is difficult to draw any strong conclusions from this speculation. The real-world problems that machine learning methods try to solve might be very different from the categorization problems that humans usually encounter—and those problems, ill-understood as they are, provide the gold standard against which the performance of a categorization model should be compared.

The problem also remains that seemingly all of the exemplars encountered need to be stored. However, the exemplar idea might scale up to a realistic number of stimuli if not all exemplars are remembered, but only certain crucial ones. This problem has also been addressed in machine learning, in which it is also desirable to store in memory as few exemplars as necessary. Solutions that only require few exemplars to be remembered are called *sparse* in machine learning. Variants of several kernel classifiers, including kernel logistic regression, try to achieve the same categorization performance while remembering fewer exemplars (Hastie et al., 2001; Schölkopf & Smola, 2002). The idea that a few representatives may be enough has been suggested in the object recognition literature (Poggio & Edelman, 1990) and is emphasized in several categorization models (Love et al., 2004; Rosseel, 2002; Verguts, Ameel, & Storms, 2004). The interesting psychological question is, of course, which exemplars are remembered and which are not. This could merely be a question of primacy, recency, and frequency, but representational considerations could also be important. On the one hand, some exemplars are simply better representatives of a category than others. On the other, some exemplars are more important for determining the decision boundary between categories. Kernel methods in machine learning could inspire new psychological models that do not have to remember all exemplars but could still achieve good generalization performance.

Conclusions on Generalization

Generalization plays a central role in theoretical approaches to the statistical learning problem (Vapnik, 2000). In psychological categorization research, the problem of generalization is often hidden behind the prototype versus exemplars debate. Prototype theorists assume very restricted category structures and can therefore generalize well, even with very few exemplars (Smith & Minda, 1998)—especially if their assumptions are true. Exemplar theories can deal with very complicated category structures but are prone to overfitting if not regularized properly. Our contribution here is to directly address concerns about the generalization performance of exemplar theories by demonstrating how good generalization can be achieved. Our discussion was mainly guided by regularization techniques, as used in machine learning. We demonstrated that ALCOVE incidentally has mechanisms akin to regularization already built in. The questions of whether humans regularize in a similar way, and if so, what their regularization looks like, open new directions for empirical research. On a more general level, the long-standing theoretical question is, what are the regularizing assumptions that allow humans to generalize? There is evidence that humans cannot learn arbitrary category structures and that some categories are harder to learn than others (Alfonso-Reese, Ashby, & Brainard, 2002; Ashby, Waldron, Lee, & Berkman, 2001; Briscoe & Feldman, 2006; Feldman, 2000; McKinley & Nosofsky, 1995). Such results potentially give hints about the assumptions on which humans base their generalization behavior. For example, some of these assumptions could be that categories do not overlap, that decision bounds are linear, that generalization gradients are wide, that small exemplar weights are to be preferred, or that only a prototype is remembered. Furthermore, machine learning methods also suggest a middle ground between prototype and exemplar theorists by showing that flexible categorization models are possible that do not need to remember all exemplars in order to generalize well.

DISCUSSION AND SUMMARY

In one recent review, parallels were noted between the object recognition and categorization literatures (Palmeri

& Gauthier, 2004). Both literatures discuss models in which the memorization of exemplars underlies human performance in the respective tasks. In object recognition, in which the same object has to be recognized irrespective of view and lighting conditions, the exemplars take the form of memorized views. Some models in object recognition explicitly make use of RBF networks (Bülthoff & Edelman, 1992; Poggio & Edelman, 1990; Riesenhuber & Poggio, 1999). A view of an object that is similar to the view an artificial neuron is tuned to will excite it, and therefore will allow for smooth interpolation between the different views. The RBF network with its tuning curves has inspired neurophysiological work in monkeys that has found some evidence for view-tuned neurons (Logothetis, Pauls, Bülthoff, & Poggio, 1994; Logothetis, Pauls, & Poggio, 1995). On the categorization side, ALCOVE has been used to model the behavior of neurons in inferotemporal cortex during a shape categorization task (Op de Beeck, Wagemans, & Vogels, 2001, 2004). In addition, the GCM, which does not lend itself easily to a neural interpretation, has inspired single-cell recordings in monkeys (Sigala, Gabbiani, & Logothetis, 2002; Sigala & Logothetis, 2002).

Although stronger links with the object recognition literature are desirable for work in categorization, this article is more concerned with the connections of categorization to machine learning. As in psychology, it is common in machine learning to consider the problem of categorization in connection with similarity and generalization. In psychology, dissimilarity has traditionally been modeled as a distance in a multidimensional space, and the same is true for machine learning. This insight about similarity is of interest, irrespective of whether one takes a prototype or an exemplar view of categorization, since both views rely on some sort of similarity measure or distance in a multidimensional space. We showed that, because exemplar theories often use Shepard's law, they can be seen as kernel methods. In particular, the model underlying ALCOVE is the same as kernel logistic regression. Contrary to the GCM, however, ALCOVE does not show any categorization behavior beyond its generalization gradients. Furthermore, the exemplar weights of ALCOVE give the model great flexibility, which can lead to overfitting. We have suggested that regularization techniques, as used in machine learning, could be employed to avoid this problem.

In the early days of machine learning, psychology and neuroscience were major inspirations that drove research in that field. Today, mainstream machine learning is far removed from modeling learning, but instead tries to build systems that work for real-world problems. As this article has demonstrated, there are nevertheless still parallels between machine learning and psychology. Machine learning has made great progress in recent years, and results from machine learning should feed back into psychology. Apart from the insights that we have presented in this article, machine learning methods can provide standards to which the performance of both humans and models can be compared and which can suggest new experiments (Graf & Wichmann, 2004; Graf, Wichmann, Bülthoff, & Schölkopf, 2006; Wichmann, Graf, Simoncelli, Bülthoff, & Schölkopf, 2005). More importantly, theoretical work in machine learning may offer a better understanding of the core problems of learning and categorization. For example, what is the role of the complexity of the category to be learned (Alfonso-Reese et al., 2002; Briscoe & Feldman, 2006; Fass & Feldman, 2003; Feldman, 2000)? Under what circumstances does a category learner generalize well (Vapnik, 2000)? In the end, both human categorizers and machine classifiers have to solve the same problems.

AUTHOR NOTE

Correspondence related to this article may be sent to F. Jäkel, Technische Universität Berlin, Fakultät IV, Sekr. 6-4, Franklinstrasse 28/29, 10587 Berlin, Germany (e-mail: fjaekel@cs.tu-berlin.de).

REFERENCES

Aizerman, M. A., Braverman, E. M., & Rozonoer, L. I. (1964). The probability problem of pattern recognition learning and the method of potential functions. *Automation & Remote Control*, **25**, 1175-1190.

Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, **64**, 570-583.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, **39**, 216-233.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 50-71.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

Ashby, F. G., Waldron, E. M., Lee, W. W., & Berkman, A. (2001). Suboptimality in human categorization and identification. *Journal of Experimental Psychology: General*, **130**, 77-96.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, **75**, 127-142.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press, Clarendon Press.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, **2**, 499-526.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics*, **32**, 213-239.

Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias–variance tradeoff. In R. Sun, N. Miyake, & C. Schunn (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1038-1043). Mahwah, NJ: Erlbaum.

Brown, J. S. (1965). Generalization and discrimination. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 7-23). Stanford: Stanford University Press.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, **89**, 60-64.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, **58**, 413-423.

Chater, N., & Vitányi, P. M. B. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, **47**, 346-369.

Cristianini, N., & Schölkopf, B. (2002). Support vector machines and kernel methods: The new generation of learning machines. *AI Magazine*, **23**(3), 31-42.

David, H. A. (1988). *The method of paired comparisons* (2nd ed.). London: Griffin.

Fass, D., & Feldman, J. (2003). Categorization under complexity: A unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 35-42). Cambridge, MA: MIT Press.

FELDMAN, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, **407**, 630-633.

FRIED, L. S., & HOLYOAK, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

GARNER, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

GHIRLANDA, S., & ENQUIST, M. (2003). A century of generalization. *Animal Behaviour*, **66**, 15-36.

GRAF, A. B. A., & WICHMANN, F. A. (2004). Insights from machine learning applied to human visual classification. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 905-912). Cambridge, MA: MIT Press.

GRAF, A. B. A., WICHMANN, F. A., BÜLTHOFF, H. H., & SCHÖLKOPF, B. (2006). Classification of faces in man and machine. *Neural Computation*, **18**, 143-165.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

JÄKEL, F., SCHÖLKOPF, B., & WICHMANN, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, **51**, 343-358.

JÄKEL, F., SCHÖLKOPF, B., & WICHMANN, F. A. (2008). *Similarity, kernels, and the triangle inequality*. Manuscript submitted for publication.

KRUSCHKE, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

LAMBERTS, K. (1994). Flexible tuning of similarity in exemplar-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1003-1021.

LOGOTHETIS, N. K., PAULS, J., BÜLTHOFF, H. H., & POGGIO, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, **4**, 401-414.

LOGOTHETIS, N. K., PAULS, J., & POGGIO, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**, 552-563.

LOVE, B. C., MEDIN, D. L., & GURECKIS, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, **111**, 309-332.

LUCE, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

LUCE, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.

LUCE, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, **15**, 215-233.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 128-148.

MCKINLEY, S. C., & NOSOFSKY, R. M. (1996). Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 294-317.

MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

MOSTOFSKY, D. I. (ED.) (1965). *Stimulus generalization*. Stanford: Stanford University Press.

NAVARRO, D. J. (2002). *Representing stimulus similarity*. Unpublished doctoral dissertation, University of Adelaide, Adelaide, Australia.

NAVARRO, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, **51**, 85-98.

NOSOFSKY, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

NOSOFSKY, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 87-108.

NOSOFSKY, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, **34**, 393-418.

NOSOFSKY, R. M. (1991a). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 3-27.

NOSOFSKY, R. M. (1991b). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition*, **19**, 131-150.

NOSOFSKY, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 363-393). Hillsdale, NJ: Erlbaum.

NOSOFSKY, R. M., & ZAKI, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 924-940.

OHL, F. W., SCHEICH, H., & FREEMAN, W. J. (2001). Change in the pattern of ongoing cortical activity with auditory category learning. *Nature*, **412**, 733-736.

OP DE BEECK, H., WAGEMANS, J., & VOGELS, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, **4**, 1244-1252.

OP DE BEECK, H., WAGEMANS, J., & VOGELS, R. (2004). A diverse stimulus representation underlies shape categorization by primates (Abstract). *Journal of Vision*, **4**(8), 518a.

ORR, G. B., & MÜLLER, K.-R. (EDS.) (1998). *Neural networks: Tricks of the trade*. Berlin: Springer.

PALMERI, T. J., & GAUTHIER, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, **5**, 291-303.

PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

POGGIO, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, **55**, 899-910.

POGGIO, T., & BIZZI, E. (2004). Generalization in vision and motor control. *Nature*, **431**, 768-774.

POGGIO, T., & EDELMAN, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

POGGIO, T., & GIROSI, F. (1989). *A theory of networks for approximation and learning* (Tech. Rep. No. A. I. Memo No. 1140). Cambridge, MA: MIT AI LAB & Center for Biological Information Processing Whitaker College.

POGGIO, T., RIFKIN, R., MUKHERJEE, S., & NIYOGI, P. (2004). General conditions for predictivity in learning theory. *Nature*, **428**, 419-422.

POGGIO, T., & SMALE, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, **50**, 537-544.

POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

RIESENHUBER, M., & POGGIO, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019-1025.

ROSCH, E., & MERVIS, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

ROSCH, E., MERVIS, C. B., GRAY, W. D., JOHNSON, D. M., & BOYES-BRAEM, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

ROSENBLATT, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386-408.

ROSSEEL, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, **46**, 178-210.

RUMELHART, D. E., HINTON, G. E., & WILLIAMS, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press, Bradford Books.

SCHOENBERG, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, **44**, 522-536.

SCHÖLKOPF, B., & SMOLA, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

SHEPARD, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325-345.

SHEPARD, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, **65**, 242-256.
SHEPARD, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I. *Psychometrika*, **27**, 125-140.
SHEPARD, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**, 54-87.
SHEPARD, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D. I. Mostofsky (Ed.), *Stimulus generalization* (pp. 94-110). Stanford: Stanford University Press.
SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.
SHEPARD, R. N., & CHANG, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, **65**, 94-102.
SHEPARD, R. N., HOVLAND, C. I., & JENKINS, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, **75**(13, Whole No. 517), 1-42.
SIGALA, N., GABBIANI, F., & LOGOTHETIS, N. K. (2002). Visual categorization and object representation in monkeys and humans. *Journal of Cognitive Neuroscience*, **14**, 187-198.
SIGALA, N., & LOGOTHETIS, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, **415**, 318-320.
SMITH, J. D., & MINDA, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1436.
SMITH, J. D., & MINDA, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 3-27.
SPENCE, K. W. (1937). The differential response in animals to stimuli varying within a single dimension. *Psychological Review*, **44**, 430-444.
TENENBAUM, J. B., & GRIFFITHS, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral & Brain Sciences*, **24**, 629-640.
TRAIN, K. E. (2003). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.
TVERSKY, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.
TVERSKY, A., & GATI, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, **89**, 123-154.
VAPNIK, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.
VERGUTS, T., AMEEL, E., & STORMS, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition*, **32**, 379-389.
WICHMANN, F. A., GRAF, A. B. A., SIMONCELLI, E. P., BÜLTHOFF, H. H., & SCHÖLKOPF, B. (2005). Machine learning applied to perception: Decision images for gender classification. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 1489-1496). Cambridge, MA: MIT Press.

NOTES

1. Briefly, for values of p that lie between 0 and 2, the function k is a so-called *positive definite kernel* (Schoenberg, 1938). This insight opens up a large box of mathematical tools from functional analysis that can be used to gain a better understanding of psychological models of similarity. We show, for example, that these tools can be used to overcome some of the serious criticism of the metric axioms—the triangle inequality in particular—put forward by Tversky and coworkers (Beals et al., 1968; Tversky, 1977; Tversky & Gati, 1982). In fact, the same tools have greatly deepened the understanding of machine learning methods that also use positive definite kernels.

2. He used $p = 2$ together with $q = 1$.

3. Interestingly, the pth power of d_p is a metric for $0 < p < 1$, even though d_p is not a metric for that range. Tversky and Gati (1982) reported data that suggest a p smaller than 1, and hence seemingly violate the metric axioms, but those researchers also noted that an unusual metric like the pth power of d_p is a viable metric alternative to Tversky's (1977) famous contrast model (Jäkel et al., 2008).

4. In the two-category case, ALCOVE is heavily overparameterized. There is a full RBF network f_1 with as many weights as exemplars for Category 1 and a full network f_2 for Category 2. One RBF network, $f = f_1 - f_2$, with the weights set to the difference, $w_{1i} - w_{2i}$, would be enough.

5. It is potentially easier, however, to directly learn the decision function rather than trying to solve the difficult problem of density estimation first (Vapnik, 2000).

6. Unless the subject redefines the perceptual dimensions, as discussed by Shepard et al. (1961).

(Manuscript received May 11, 2007;
revision accepted for publication July 10, 2007.)