TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 5 (2013) 132–172
Copyright © 2013 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12010

Human Semi-Supervised Learning

Bryan R. Gibson,[a] Timothy T. Rogers,[b] Xiaojin Zhu[a]

[a]*Department of Computer Sciences, University of Wisconsin-Madison*
[b]*Department of Psychology, University of Wisconsin-Madison*

Received 20 February 2012; received in revised form 31 July 2012; accepted 21 August 2012

Abstract

Most empirical work in human categorization has studied learning in either fully supervised or fully unsupervised scenarios. Most real-world learning scenarios, however, are semi-supervised: Learners receive a great deal of unlabeled information from the world, coupled with occasional experiences in which items are directly labeled by a knowledgeable source. A large body of work in machine learning has investigated how learning can exploit both labeled and unlabeled data provided to a learner. Using equivalences between models found in human categorization and machine learning research, we explain how these semi-supervised techniques can be applied to human learning. A series of experiments are described which show that semi-supervised learning models prove useful for explaining human behavior when exposed to both labeled and unlabeled data. We then discuss some machine learning models that do not have familiar human categorization counterparts. Finally, we discuss some challenges yet to be addressed in the use of semi-supervised models for modeling human categorization.

Keywords: Category learning; Semi-supervised learning; Machine learning

Cognitive psychology has long had an interest in understanding human categorization: how we come to conceive of objects in the world as belonging to different categories, and how we use categories to draw inferences about the unobserved properties of objects. Toward this end, one of the most commonly used experimental paradigms has been supervised category learning: On each trial, the participant views a stimulus and must guess to which of a small number of categories it belongs. Feedback is provided that indicates either whether the guess was correct or what the correct answer was—the learning is supervised in this sense. The experimenter then measures how rapidly the participant learns to generate correct inferences about category membership, and how the acquired knowledge generalizes to novel stimuli.

Correspondence should be sent to Bryan R. Gibson, Department of Computer Sciences, University of Wisconsin-Madison, 1210 W. Dayton St., Madison, WI 53706-1685. E-mail: bgibson@cs.wisc.edu

This use of supervised learning as an experimental procedure has proven exceedingly fertile—hundreds of experiments in this vein have been conducted and a variety of interesting regularities in human behavior have been convincingly documented. Some well-known examples include the artificial category learning experiments of Hintzman (1986), Kruschke (1992), Murphy and Smith (1982), Nosofsky (1984, 1986, 1987), Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), and Smith and Medin (1981). The empirical record arising from such studies has in turn provided a test-bed for the development of mathematical models of categorization, with considerable success. Three of the best-known theoretical models of human categorization—the exemplar theory (Hintzman, 1986; Kruschke, 1992; Nosofsky, 1986; Smith & Medin, 1981), the prototype theory (Hampton, 1993; Rosch et al., 1976; Smith & Minda, 2002), and the Rational theory (Anderson, 1991)—all have their counterparts in formal mathematical models, and all three modeling approaches have been brought to bear on a remarkably wide range of empirical findings from supervised learning experiments. A great deal of work has focused on adjudicating which theory or model provides the best overall account of human behavior in these tasks, with somewhat equivocal conclusions, and all three theories remain prevalent in the literature.

A smaller range of studies have also investigated fully unsupervised learning, in which participants study a series of stimuli that appear without category labels. Learning is subsequently assessed either by having participants judge whether test items were identical to or of the same kind as, the previously studied items (Billman & Knutson, 1996; Love, 2002; Love, Medin, & Gureckis, 2004; Posner & Keele, 1968). Models developed to explain learning in unsupervised experimental settings are typically some variant of cluster analysis (Fried & Holyoak, 1984; Pothos & Chater, 2002; Stutz & Cheeseman, 1995), and some studies have directly compared human and model behavior in fully supervised and fully unsupervised learning scenarios (Fried & Holyoak, 1984; Gureckis & Love, 2003; Love, 2002; Love et al., 2004).

In the real world, however, category learning is neither fully supervised nor fully unsupervised. Instead, everyday experience might best be described as *semi-supervised*: Much of the time we simply observe items in the world and draw uncorrected inferences about the category to which they belong, though of course there are also occasions where we are directly told about an item's category membership. That is, everyday experience provides a mix of *labeled data* (in which both an item and its category label are directly observed, such as seeing a dog and hearing the word "dog") and *unlabeled data* (in which the item is observed but no label is provided, as when the learner simply views a dog).

At first glance, it might seem as though semi-supervised learning (or SSL) should not differ much from supervised learning (SL). After all, when faced with the problem of categorizing a novel unlabeled item, the learner must leverage whatever information she has gleaned from previous labeled examples. The preceding unlabeled examples give no direct information about the mapping from example to label. A recent stream of research in machine learning suggests, however, that there is information to be gleaned from unlabeled examples that can extend the knowledge gained from labeled examples alone, greatly aiding in category learning (Chapelle, Zien, & Schölkopf, 2006; Zhu & Goldberg, 2009).

To gain an intuitive sense of why this might be so, consider the example provided by Kalish, Rogers, Lang, and Zhu (2011): A traveler camping in the wilderness of a foreign land observes an animal in the shadows and discerns that it is about 1.5 feet long. His companion tells him the animal is a dax. Sometime later he observes a larger animal, about 2.5 feet long, and his companion informs him it is a zog. From this labeled information, the traveler may infer that daxes are typically about 1.5 feet long; zogs are typically about 2.5 feet long; and the boundary between them is somewhere around 2 feet in length. However, now suppose the knowledgeable companion leaves to collect firewood, and the traveler is left alone to observe animals scuffling in the shadows. Over time he observes animals of different lengths, and he notices that there are two clusters: a group of smaller animals about a foot in length, and a group of larger animals about 2 feet in length. The traveler might reasonably infer that the smaller group consists of daxes and the larger group of zogs, and might adjust his beliefs about these categories accordingly, despite not receiving any further instruction about true category labels. For instance, he may end up deciding that daxes tend to be about a foot long (rather than 1.5 feet); that zogs tend to be about 2 feet long (rather than 2.5 feet); and that the boundary between these is about 1.5 feet in length (rather than 2 feet). Thus, the probability density of the unlabeled distribution might reasonably be used to adjust conclusions about both the central tendencies and boundary between categories, compared with the conclusions drawn from labeled data alone. Such an adjustment will be beneficial to the learner given certain assumptions about the relationship between the unlabeled distribution and the category structure. Although unsupervised learning models are specifically designed to make use of this unlabeled distribution information, they do not describe how to make use of any labeled information encountered. The ability to handle both labeled and unlabeled information is what characterizes and differentiates SSL models from supervised and unsupervised models.

Interestingly, very little empirical work in cognitive science has investigated how human beings jointly exploit labeled and unlabeled data. This lacuna is somewhat puzzling, perhaps, because most of the best-known computational models for human category learning can be fairly easily extended to encompass SSL. Indeed, SSL has been extensively studied in the machine learning community, where a wide variety of models have been developed, each suited to a different set of assumptions about the relationship between the category label and the distribution information provided by the unlabeled items. Some of these models have been adapted or *lifted* from well-understood supervised models that are formally equivalent to the familiar exemplar, prototype, and Rational models in cognitive science. Others adopt assumptions that have no direct counterpart in current psychological theories, and so may provide a fruitful source of new hypotheses about potential category-learning mechanisms in human cognition.

The current review thus has three goals. The first is to show explicitly how well-known variants of the exemplar, prototype, and Rational models of categorization can be lifted to make use of both labeled and unlabeled learning experiences. Toward this end, we will briefly review the standard variants of these models and, following Griffiths, Sanborn, Canini, Navarro, and Tenenbaum (2011) and others (Neal, 1998; Griffiths, Canini, Sanborn, & Navarro, 2007; Sanborn, Griffiths, & Navarro, 2006), will show how these relate to well-studied models in machine learning. We will then illustrate how the machine learning models are lifted to make use of unlabeled data, and we will specify the underlying assumptions this requires. We will take the resulting models as candidate hypotheses about human SSL.

Second, we will review empirical studies of human SSL published in the last 5 years, and we will consider to what extent the documented patterns of behavior are consistent with predictions of the different SSL models. We will see that there is now considerable evidence that human beings do combine labeled and unlabeled learning experiences in ways consistent with some of the models; and that the different model variants can sometimes make differing predictions, only some of which are consistent with observed behavior. This in turn suggests that SSL may provide a means of adjudicating different cognitive theories of category learning.

Third, we will consider some of the open questions and challenges faced by the SSL research program. We will briefly describe some machine learning models that have no current counterpart in cognitive psychology, and how these may be used to develop new hypotheses about human SSL. We will also highlight some characteristics of human cognition and learning that challenge the extension of the approach to more complex and realistic learning scenarios.

1. Psychological and machine learning models of categorization

In this section, we will review well-known instantiations of the exemplar, prototype, and Rational models of categorization, illustrate how they are formally equivalent to models from machine learning, and explain how the machine learning models are *lifted* to make use of both unlabeled and labeled information. As the psychological and machine learning models are formally identical, the lifted variants of the machine learning models provide candidate hypotheses about human SSL. Some of the important relationships between the psychological and machine learning models have been discussed in detail by other researchers (Ashby & Alfonso-Reese, 1995; Fried & Holyoak, 1984; Griffiths et al., 2007; Nosofsky, 1991; Sanborn et al., 2006), but we review these relationships here for readers who are not intimately familiar with the details since a good understanding of the semi-supervised variants will depend on clear exposition of the standard models.

Before beginning, it will be useful to define the categorization task itself, to indicate very generally how mathematical models in psychology and machine learning have been brought to bear on the task, and to introduce some notation. A standard categorization task asks a learner to label a previously unseen item x_i after viewing a set of labeled examples $(x, y)_{1:i-1}$. In this notation, x_i indicates a multidimensional feature vector that describes a single stimulus item (with i indexing the order in which items are seen over time), and y_i indicates the category label associated with each item. In both psychology and machine learning, the probabilistic way of modeling human category decisions for x_i is to calculate $P(y_i = k \mid x_i, (x, y)_{1:i-1})$, that is, the probability that a person will choose label $y_i = k$ for each of $k \in K$ categories given the current item x_i and the preceding labeled evidence $(x, y)_{1:i-1}$, that is, the "training examples" viewed prior to the query item x_i.

A common way to compute the probability $P(y_i = k \mid x_i, (x, y)_{1:i-1})$ is via the Bayes rule. Formally, the Bayes rule states

$$P(y_i = k \mid x_i, (x, y)_{1:i-1}) = \frac{P(x_i \mid y_i = k, (x, y)_{1:i-1}) P(y_i = k \mid (x, y)_{1:i-1})}{\sum_{k'} P(x_i \mid y_i = k', (x, y)_{1:i-1}) P(y_i = k' \mid (x, y)_{1:i-1})}. \quad (1)$$

On the right-hand side, the first term in the numerator, $P(x_i \mid y_i = k, (x, y)_{1:i-1})$, is the *likelihood*, which specifies the probability of observing item x_i assuming it has the label $y_i = k$. The second term in the numerator, $P(y_i = k \mid (x, y)_{1:i-1})$, is the *prior*, which specifies the probability, prior to observing x_i, that x_i will have label $y_i = k$. The left-hand side, $P(y_i = k \mid x_i, (x, y)_{1:i-1})$, is the *posterior*, which indicates the probability that k is the correct label after seeing x_i. The denominator is a normalization factor so that the posterior probability sums to 1. Once the posterior probability is computed, one can classify x_i by the most likely label:

$$\hat{y}_i = \arg\max_{k \in K} P(y_i = k \mid x_i, (x, y)_{1:i-1}) = \arg\max_{k \in K} P(x_i \mid y_i = k, (x, y)_{1:i-1}) P(y_i = k \mid (x, y)_{1:i-1}). \quad (2)$$

The above classification rule minimizes expected error. Alternatively, one can *sample* the class label in a practice known as Gibbs classification in machine learning:

$$\hat{y}_i \sim P(y_i = k \mid x_i, (x, y)_{1:i-1}) \quad (3)$$

which corresponds to probability matching in psychology (Myers, 1976; Vulkan, 2000).

In machine learning, there exist a variety of models for computing the posterior via Bayes rule. In all of these models, the prior is typically a multinomial distribution over the values y_i may take (i.e., the different category labels). Thus, the primary difference between probabilistic machine learning models is in how the likelihood term is calculated. Interestingly, three common machine learning models of this computation bear a striking resemblance to the exemplar, prototype, and Rational models of human categorization. Indeed, certain parametrization of the psychological models are formally identical to the machine learning models. This identity is, perhaps, surprising since the primary goal of the psychological work has been to fit observed human behavior in artificial category learning experiments. Many early theorists, with the notable exceptions of Anderson (1991) and Shepard (1991), did not explicitly consider whether the probabilities computed by a given model were correct in any formal sense (see e.g., Hintzman, 1986; Medin & Schaffer, 1978; Rosch et al., 1976). The fact that the psychological and probabilistic models are formally equivalent thus suggests to some researchers that human categorization decisions are optimal in some respects—that is, the decisions people make are shaped by estimates of the true posterior probability distribution and so represent the best decisions that can be made given prior beliefs and learning episodes (Anderson, 1991; Griffiths et al., 2011; Sanborn et al., 2006; Tenenbaum, Griffiths, & Kemp, 2006).

The equivalence of psychological and probabilistic models is also useful for another reason: It allows us to leverage insights from machine learning to develop explicit hypotheses about human SSL. A considerable amount of work in machine learning has focused on how best to exploit both labeled data, consisting of (x,y) pairs, and *unlabeled* data, consisting of only the x observations without y labels. The modification of a supervised model to make use of unlabeled data is sometimes called *lifting* the model. In machine learning, the primary motivation for lifting supervised models has been that labeled data are often expensive—that is, data labeling can be time-consuming and often requires an expert in the field. In contrast, unlabeled data are usually plentiful and inexpensive to acquire in large quantities. A key discovery has been that, under certain well-specified assumptions, semi-supervised models can use the potentially inexpensive unlabeled data to greatly improve classifier performance compared with supervised models alone (Balcan & Blum, 2010).

By definition, unlabeled data do not come with labels and so cannot be used directly for supervised learning. Instead, these data provide information about the marginal $P(x)$, that is, the distribution of items in the feature space. To use this information for category learning, assumptions must be made about the nature of the unlabeled item distribution and the relationship between $P(x)$ and $P(y \mid x)$. These assumptions then "steer" how category learning proceeds. SSL is the learning paradigm that adopts such assumptions to make use of both labeled and unlabeled data when learning to categorize.

There are many types of SSL assumptions (Chapelle et al., 2006; Zhu & Goldberg, 2009) that can be used to support different kinds of learning models. The assumption most germane to existing psychological models of categorization is likely the *mixture model assumption*, which states that all items are drawn independently from a probability distribution composed of a mixture of underlying components. The observed distribution of unlabeled examples can thus be used to infer the underlying mixture components, while the comparatively infrequent labeled examples can be used to label each component. We will use the mixture model assumption to create lifted variants of the prototype and Rational models of human SSL. The exemplar model is a non-parametric model that requires a slightly different assumption.

Finally, one further point should be noted in relating probabilistic machine learning models to psychological models. In classic machine learning, models are typically trained in *batch mode*: The model is fit exactly once to the full history of prior training examples, and the trained model is then used to classify new items without further training. In the case of human SSL, this strategy seems inappropriate. Unlabeled "test" items provide information about the marginal distribution $P(x)$, and so can be used for further learning. Moreover, under some theoretical approaches it seems unreasonable to suppose that people are able use the entire history of prior learning experiences when updating their current model to accommodate a new observation. Instead, with each new unlabeled item encountered, the learner presumably uses the current model acquired from previous labeled and unlabeled items to generate a guess about the correct label. The new item can then be used to update the current model, using whatever information happens to be stored under a given theoretical approach. This kind of learning, in which the model is continually updated with each new labeled or unlabeled example, is referred to as *online learning*. Although online learning clearly provides a better analog to human behavior, it sometimes complicates the derivation of the model. The SSL models derived in this section are all suited to online rather than batch learning, and we will note where this differs from analogous batch-learning models.

With these points as background, we are now ready to review the development of each of the three SSL models.

1.1. A semi-supervised exemplar model

One common model of human categorization is the *exemplar* model, which stores all previously viewed examples and uses these to estimate the most likely category label for novel query items. The Generalized Context Model (GCM) proposed by Nosofsky (1986, 2011) is probably the best known of this class in cognitive science. To facilitate comparison with machine learning models, we consider a specific parametrization of the full GCM model, in which two free parameters, memory strength and dimensional scaling weights (see Nosofsky, 2011), are fixed to one.

With this simplification, the GCM model can be described as

$$P(y_i = k \mid x_i, (x, y)_{1:i-1}) = \frac{b^{(k)}\left(\sum_{j:y_j=k} s(x_i, x_j)\right)}{\sum_{k':k'\in K} b(k')\left(\sum_{j':y_j=k'} s(x_i, x_{j'})\right)} \tag{4}$$

where $b^{(k)}$ is the bias on category k and $s(x_i, x_j)$ is a scaled similarity measure between item x_i and x_j. The bias term b serves the same role as the prior in the Bayes rule: It indicates the probability of encountering a label with value k prior to observing the query item. Intuitively, it is easy to see that the probability of the query item x_i sharing the same label as a stored item x_j grows with the similarity s between the queried and stored items. Consequently, the probability that the query item receives label k depends on its similarity to all items in category k and its similarity to all other items in the contrasting categories.

This formulation does not specify how the similarity between the queried and stored examples is to be computed. In machine learning, a common choice for s is the Gaussian function, which, in 1D is defined by

$$s(x_i, x_j) = \exp\left[-\frac{1}{2\sigma^2}(x_i - x_j)^2\right] \tag{5}$$

where σ^2 is the variance. In psychological models, it is more common to employ an exponential similarity gradient, following Shepard (1986). Shepard's (1986) arguments, however, were premised on the assumption that the item distribution x was uniform over discrete dimensions (see Anderson, 1991); in the studies we consider below, the items are sampled from a mixture of Gaussian distributions in a fully continuous space. Empirically, at least one study has found that Gaussian similarity functions can provide a better fit to human behavior for such stimuli (Nosofsky, 1985). Moreover, an interesting property of this class of model is that, in the limit, the estimate of $P(y_i = k \mid x_i, (x,y)_{1:i-1})$ is not affected by the shape of the similarity gradient (often referred to as a *kernel* in machine learning). For these reasons, we have followed the more common convention in machine learning and employed a Gaussian similarity function in what follows.

1.1.1. Kernel density estimation

A clear analog to exemplar models in machine learning is *Kernel Density Estimation* (KDE), which is a method for estimating the likelihood term $P(x \mid y = k)$ in the Bayes rule. Like exemplar models, each labeled example (x, y) is retained in KDE and is used to compare against the current query item.[1] One model that makes use of the likelihood estimate provided by KDE is the Nadaraya–Watson kernel estimator (Nadaraya, 1964; Shi, Feldman, & Griffiths, 2008; Wasserman, 2006), a regression function that returns a real value. When this estimator is adapted to categorization, the real value provides a direct estimate of the conditional probability $P(y_i = k \mid x_i, (x,y)_{1:i-1})$. Given training data $(x,y)_{1:i-1}$, the categorization function is

$$P(y_i = k \mid x_i, (x,y)_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathcal{K}(\frac{x_i - x_j}{h})\delta(y_j, k)}{\sum_{j'=1}^{i-1} \mathcal{K}(\frac{x_i - x_{j'}}{h})} \tag{6}$$

where the kernel function $\mathcal{K}$ determines the weight between the query item x_i and each of the 1,…, i-1 exemplars x_j, and where $\delta(y,k) = 1$ when $y = k$ and 0 otherwise.

From this description, the equivalence between Eqs. 6 and 4 may not be immediately obvious. Under certain parameter settings, however, the equivalence becomes clear. We repeat the exemplar formulation here for convenience:

$$P(y_i = k \mid x_i, (x,y)_{1:i-1}) = \frac{b^{(k)}\left(\sum_{j:y_j=k} s(x_i, x_j)\right)}{\sum_{k':k' \in K} b^{(k')}\left(\sum_{j':y_j=k'} s(x_i, x'_j)\right)}. \tag{7}$$

The kernel function $\mathcal{K}$ acts like the similarity function $s(x_i, x_j)$, returning a value that gives a sense of the "similarity" between the query x_i and an exemplar x_j. The hyperparameter h is known as the *bandwidth* parameter and controls how the effect of each exemplar diminishes with distance. Using a Gaussian function for s (in the exemplar model) and a Gaussian kernel for $\mathcal{K}$ (in the machine learning model), and setting the bandwidth h to one standard deviation of this Gaussian, the functions become identical:

$$s(x_i, x_j) = \exp\left[\frac{1}{2h^2}(x_i - x_j)^2\right] = \exp\left[\frac{1}{2}\left(\frac{x_i - x_j}{h}\right)^2\right] = \mathcal{K}\left(\frac{x_i - x_j}{h}\right). \tag{8}$$

Setting $b^{(k)} = 1$ for all k completes the equivalence. This parametrization of the Nadaraya–Watson KDE is therefore formally identical to the parametrization of the GCM described in (4) with the additional constraint that all categories are assumed to be equally likely a priori.

1.1.2. Lifting the exemplar model

To derive the semi-supervised exemplar model, we describe a lifted version of KDE and make use of the equivalence between the Nadaraya–Watson KDE and the GCM model. The standard model is lifted as follows: When an item x_i is queried for a label, the supervised model returns $P(y_i = k \mid x_i)$ for all $k = 1,\ldots, K$ categories. Normally, in supervised learning, the true label y_i will then be received and the labeled (x_i, y_i) pair added to the training set in preparation for the next query item x_{i+1}. In the semi-supervised setting, x_i may remain unlabeled, so that no ground truth y_i label is received. Instead of tossing out this unlabeled x_i, as would happen in the supervised case, the real value $P(y_i = k \mid x_i)$ is calculated for all $k = 1,\ldots, K$ and these values are considered *soft labels* on x_i. The labels are "soft" because they are uncertain—each category k has probability $P(y_i = k \mid x_i)$ of being the correct label, rather than a "hard" or certain probability of 0 or 1. The x_i, together with the soft labels, is then added to the training set as a pseudo labeled exemplar. Equivalently, one can think of K copies of x_i, each with a distinct label and fractional weight $P(y_i = k \mid x_i)$. For such unlabeled training items, it is necessary to retain $P(y_i = k \mid x_i)$ for all $k = 1,\ldots, K$. Thus, we now maintain $(x_i, \mathbf{y}_i)$ pairs where $\mathbf{y}_i$ is a vector with $\mathbf{y}_{ik} = P(y_i = k \mid x_i)$, $k = 1, \ldots, K$. If x_i is labeled with $y_i = k^*$, the corresponding $\mathbf{y}_{ik^*} = 1$ while $\mathbf{y}_{ik} = 0$ for all other values of k. Algorithm 1 describes the model in detail.

Algorithm 1
Semi-supervised exemplar model

```
**Given:** Kernel bandwidth $h$
**for** $i = 1,2,\ldots$ **do**
    Receive $x_i$ and predict its label using
        $\arg\max_{k \in K} P(y_i = k \mid x_i, (x,y)_{1:i-1}) = \frac{\sum_{j=1}^{i-1} \mathcal{K}(\frac{x_i - x_j}{h}) \mathbf{y}_{jk}}{\sum_{j'=1}^{i-1} \mathcal{K}(\frac{x_i - x_{j'}}{h})}.$    (9)
    **if** $x_i$ is labeled with $y_i = k^*$ **then**
        Set $\mathbf{y}_{ik} = \begin{cases} 1, & \text{if } k = k^* \\ 0, & \text{and o.w.} \end{cases}$,   for $k$=1,…, $K$
**else**
        Set $\mathbf{y}_{ik} = P(y_i = k \mid x_i, (x,y)_{1:i-1})$ for $k = 1, \ldots, K$.
    **end if**
        Add $(x_i, \mathbf{y}_i)$ as an exemplar.
**end for**
```

The effect of the lifting is that unlabeled items are now "split" among the different categories. When an item is very likely to be in class k (i.e., $P(y = k \mid x) \approx 1$), it will be treated similarly to a labeled item, contributing almost all of its "weight" to category k. When the label is more uncertain, the unlabeled item will distribute its "weight" among the different categories. We note that this model is similar to an idea proposed by

Nosofksy and Palmeri (1997) to explain test-item effects in standard supervised category-learning experiments as well as the model proposed in Zaki and Nosofsky (2007) to explain results seen in experiments designed to discriminate exemplar and prototype learning.

1.2. A semi-supervised prototype model

Unlike the exemplar model, where learning is accomplished by storing all individual training items, learning in the *prototype* model consists of summarizing each category and discarding the training items themselves. The summary is achieved by assuming that each category can be represented with a parametric distribution $P(x \mid y = k)$, so that only the distribution parameters for each category need be retained. The parameters associated with a given category constitute the category *prototype*. Prototypes do not necessarily correspond to any particular labeled item but are abstract representations of all labeled items in the category they represent. For example, if we assume that each category $P(x \mid y = k)$ has a Gaussian distribution, then the corresponding prototype can be represented by the parameters $\mu^{(k)}$ (mean or "component center") and $\sigma^{2(k)}$ (variance or "spread"). Typically, the number of categories K in the model is fixed in advance, before any labeled examples are seen, so that the number of stored prototypes does not grow with the number of examples. A new item is labeled by comparing it to each stored prototype.

A variety of different prototype models have been proposed in the literature. To illustrate the link to machine learning, we will consider the model proposed by Minda and Smith (2011), in which the prototype is simply the sample mean of labeled training examples in a given category. Query items are labeled using the same method as in the exemplar model, by comparison to a set of stored representations. The difference is that the stored representations are category prototypes, and not the labeled training items themselves. Thus, it is not surprising that the formal description of the model is very similar:

$$P(y_i = k \,|\, x_i, (x, y)_{1:i-1}) = \frac{b^{(k)} s(x_i, \bar{x}^{(k)})}{\sum_{k:k' \in K} b^{(k')} s(x_i, \bar{x}^{(k')})} \tag{10}$$

where $\bar{x}^{(k)}$ is the prototype for category k and $s(x_i, \bar{x}^{(k)})$ is a similarity function as in Eq. 4, except that now x_i is compared with a single summary representation $\bar{x}^{(k)}$ of each category k. Just as in the exemplar model, the bias term $b^{(k)}$ encodes the prior belief on label k.

Different specifications of the similarity function s lead to different formal models. For instance, if the prototype is construed as a multivariate Gaussian describing the distribution of features among labeled members of a category, then the similarity function in the equation above can be defined with a function estimating, for each category, the probability that the novel item was generated from its distribution.

1.2.1. Gaussian mixture models

A machine learning analog to prototype models is the *mixture model*, in which items are assumed to be generated from some mixture of underlying components. Each component is represented by a set of parameters that are learned from the data, with the number of components fixed before learning. A common variation is the Gaussian Mixture Model (GMM), where each category is represented by a single component corresponding to a Gaussian distribution. The GMM has the parameters $\theta = \{\alpha,\mu,\Sigma\}$, where α is the set of non-negative mixing parameters $\{\alpha^{(1)}, \ldots, \alpha^{(K)}\}$ (normalized so that $\sum_{k=1:K} \alpha^{(k)} = 1$) defining how often each category is sampled from, μ is a vector of the corresponding K means $(\mu^{(1)}, \ldots, \mu^{(K)})$, and Σ is a set of covariance matrices $(\Sigma^{(1)}, \ldots, \Sigma^{(K)})$. When x is one-dimensional, the covariance matrices are replaced by variances $\sigma^{2(1)}, \ldots, \sigma^{2(K)}$. The model is defined by the joint probability $P(x_i, y_i \mid \theta) = P(y_i \mid \theta)P(x_i \mid y_i, \theta)$ where

$$P(y_i = k \mid \theta) = \alpha^{(k)}, \tag{11}$$

$$P(x_i \mid y_i = k, \theta) = \mathcal{N}(x_i; \mu^{(k)}, \Sigma^{(k)}). \tag{12}$$

Note that the $i - 1$ training examples seen prior to the query x_i are not used directly to label new items but instead are used to estimate the parameters θ, typically via the maximum likelihood estimate. We denote the parameter estimates after seeing the $i - 1$ examples as $\hat{\theta}_i$. The probability distribution over category labels for the query item x_i is then computed as the posterior

$$P(y_i = k \mid x_i, \hat{\theta}_i) = \frac{P(x_i, y_i = k \mid \hat{\theta}_i)}{\sum_{k' \in K} P(x_i, y_i = k' \mid \hat{\theta}_i)} = \frac{P(x_i \mid y_i = k, \hat{\theta}_i)P(y_i = k \mid \hat{\theta}_i)}{\sum_{k' \in K} P(x_i \mid y_i = k', \hat{\theta}_i)P(y_i = k' \mid \hat{\theta}_i)} \tag{13}$$

with the most likely label found by taking arg $\max_k P(y_i = k \mid x_i, \hat{\theta}_i)$.

As was the case when comparing KDE and exemplar models, GMMs are identical to prototype models under a certain parametrization. Again, we repeat the equation for the prototype model for convenience:

$$P(y_i = k \mid x_i, (x, y)_{1:i-1}) = \frac{b^{(k)} s(x_i, \bar{x}^{(k)})}{\sum_{k':k' \in K} b^{(k')} s(x_i, \bar{x}^{(k')})}. \tag{14}$$

As in the exemplar model, we define the function s to be a Gaussian (5). Unlike the exemplar model, where we compare the query x_i to each labeled example, here we only compare it to the set of K prototypes $\{\bar{x}^{(k)} : k \in K\}$ corresponding to the K categories. For each category, the point $\bar{x}^{(k)}$ is equal to the sample mean $\hat{\mu}^{(k)}$ for that category in the GMM formulation, while the covariance $\hat{\sigma}^{2(k)}$ enters s implicitly via the definition of the multivariate Gaussian probability density function. The set of $\hat{\alpha}$ corresponds to the set of $b^{(k)}$. Thus, under these settings the prototype model is equivalent to the GMM used in machine learning.

1.2.2. Lifting the prototype model

Recall that, in the prototype and GMM frameworks, the number of prototypes is fixed, usually equal to the number of categories, and each prototype is encoded by parameters learned from the training set. In the supervised setting, these parameters can be computed in closed form by the maximum likelihood estimate. In the semi-supervised setting, the closed-form computation is no longer possible because it is not clear to which category each unlabeled item belongs, and consequently, it is not clear to which parameter estimates the item should contribute. To make use of unlabeled data, the maximum likelihood estimate is instead computed using an approximation method, typically the *expectation maximization* (EM) algorithm (Dempster, Laird, & Rubin, 1977). This is an iterative procedure that first fixes the parameters of each component and uses these to calculate the label probabilities for each unlabeled item (the E step), then fixes these labelings to re-estimate the parameters (the M step). These two steps are iterated until the labelings and parameter estimates stop changing significantly.

In the E step, it is useful to think of each unlabeled item as splitting its label across all categories, contributing to each category by an amount equal to the posterior probability that the label is correct, just as in the exemplar/KDE model. These "soft" labels can be stored in a vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iK})$ for each unlabeled item x_i. For labeled items, the provided labels are stored in a similar vector in which $\mathbf{y}_{ik^*} = 1$ if $y_i = k^*$ and 0 otherwise. The key difference from the exemplar/KDE approach is that these label probabilities are subsequently used to adjust the parameter estimates (prototypes) associated with each category. The contribution of each unlabeled item to the parameter estimates for category k is weighted by the posterior probability, estimated in the E step, that the item belongs to category k. Thus, while an item unlikely to belong to the category contributes little to its parameter estimates, an item very likely to belong to the category will be treated like a "hard" label. An item whose category status is uncertain will contribute slightly to the estimate of many category prototypes.

Once the category parameter estimates are adjusted, however, this then alters the posterior probability distributions over labels for the unlabeled items—so these must be re-computed. The new labeling probabilities then change the parameter estimates for each category, and so on. When both the parameter estimates and the probability distributions over labels have stabilized, the joint probability over labeled data and the marginal probability over unlabeled data are maximized, but only locally. The particular local maximum found depends on the initialization of the algorithm. A common way of initializing the process is to use labeled data to find the initial estimate of the parameters. The reader is referred to Zhu et al. (2010) for a full description of the lifted algorithm for the online SSL prototype model.

1.3. A semi-supervised Rational model

The exemplar and prototype models lie at the two extremes of a continuum. The exemplar model assumes that every new learning example is stored in memory so the representational elements grow with the number of training examples, and no assumptions are made about the number or distribution of categories. The prototype model assumes that there are a fixed number of components (category prototypes) whose distributional parameters must be learned, and these are the only representational elements stored in memory.

Several models have been proposed that exist between these extremes (e.g., the Varying Abstraction model; see Vanpaemel, Storms, & Ons, 2005). Here, we will consider Anderson's Rational model (Anderson, 1990, 1991) as it has been highly influential and closely resembles models encountered in machine learning.

Whereas exemplar models treat the individual labeled items $(x, y)_{1:i-1}$ as the base representational elements, and prototype models treat prototypes of each category ($k \in K$) as the base elements, the base elements in the Rational model can range from individual items to large clusters depending on the best fit to the data and prior beliefs. A learned model consists of a partitioning of the labeled data into clusters. In general, each category may be represented by a number of clusters, where the number is not fixed prior to learning but can grow indefinitely with more training examples.

A version of the Rational algorithm, slightly modified from the presentation in Anderson (1991), is presented in Algorithm 2.[2] Each labeled example is assigned either to an existing cluster or, with some probability, to a completely new cluster that contains only the new item. If a new cluster is created, it is then included as one of the existing clusters under consideration when the next labeled item is assigned.

Algorithm 2
Rational model of categorization

```
Given: the cluster assignments $z_{1:i-1}$ assigning $x_{1:i-1}$ to clusters in $L$:
for each cluster $l \in L$ do
    calculate $P(z_i = l \mid x_i, x_{1:i-1}, z_{1:i-1})$, the probability that $x_i$ comes from cluster $l$.
end for
Also, let $P(z_i = l' \mid x_i)$ be the probability that $x_i$ comes from a new cluster $l'$.
Assign $x_i$ to the cluster with maximum probability:
    $z_i = \arg\max_{l \in \{L, l'\}} \begin{cases} P(z_i = l \mid x_i, x_{1:i-1}, z_{1:i-1}) \\ P(z_i = l' \mid x_i) \end{cases}$    (15)
If the assigned cluster is the new $l'$, add $l'$ to $L$.
```

The term $P(z_i = l' \mid x_i)$ controls the probability that a given item will be assigned to a new cluster, with the effect that the number of representational elements in a trained model will vary with this term. This probability in turn depends on a "coupling parameter" that specifies the prior probability of any two items being drawn from the same cluster. When the coupling parameter is low, $P(z_i = l' \mid x_i)$ is high, so each labeled item will likely be placed in its own cluster, similar to the exemplar model. When the coupling parameter is high, $P(z_i = l' \mid x_i)$ is low and relatively few clusters will be learned, similar to the prototype model. In Anderson (1991), the coupling parameter is assumed to be fixed in advance of training.

1.3.1. Dirichlet process mixture models

Dirichlet Process Mixture Models (DPMMs) are to KDEs and GMMs as the Rational model is to exemplar and prototype models: DPMMs allow the number of components of the mixture model to grow dynamically with the number of data points observed. Anderson's Rational model was in fact shown to be equivalent to the DPMM (Griffiths et al., 2011; Neal, 1998; Sanborn et al., 2006). To understand how this model works in a categorization task, it is first necessary to understand DPMMs as generative models that specify both (a) how a dynamically growing number of components in a mixture model can be generated using a Dirichlet Process (DP) (Teh, 2010), and (b) how items are generated from the growing mixture model.

The DP takes two parameters: a concentration parameter α (similar to $P(z_i = l' \mid x_i)$ in the Rational model) and a base distribution H. The concentration parameter α is a positive real number that regulates how often a new component for the mixture will be generated. The base distribution H is a distribution over parameters θ of the components of the mixture model. Each component is itself a distribution $F(x;\theta)$ and each time a new component is added to the mixture its parameters are sampled from the base distribution H. For example, consider a mixture of 1D Gaussian components all sharing a fixed variance σ^2. Each component is defined by its mean μ (the cluster center) so that for the component l, $\theta^{(l)} = \mu^{(l)}$ and $F(x;\theta^{(l)}) = \mathcal{N}(x;\mu^{(l)},\sigma^2)$. The base distribution H is then some distribution over all possible $\mu \in R$. For instance, it could be another 1D Gaussian, $H = \mathcal{N}(0,1)$.

Now consider how a sequence of items $x_1, x_2, \ldots$ might be generated from a DPMM. To generate the first item x_1, we must determine from which component of the mixture it is to be sampled. Since the number of components is zero in the beginning, the only option is to create a new component for the mixture. We assign the item an integer index z_1 that indicates from which component it will be sampled. In this case, there is only one component, so $z_1 = 1$. We then define the first component by drawing its parameters from the base distribution H: $\theta^{(1)} \sim H$. In the 1D example, we draw the new mean $\mu^{(1)} \sim \mathcal{N}(0,1)$. Then, to generate the actual item, we sample an observation from this component: $x_1 \sim F(\theta^{(1)})$, for example, $x_1 \sim \mathcal{N}(\mu^{(1)},\sigma^2)$.

To generate x_2, we must again decide from which component to draw the sample and then indicate this with a component index z_2. In this case, there are two options: The item can be drawn from the existing component, in which case $z_2 = 1$, so that x_2 and x_1 "belong to" the same component and $x_2 \sim F(\theta^{(1)})$, for example, $x_2 \sim \mathcal{N}(\mu^{(1)},\sigma^2)$. Alternatively, x_2 can be drawn from a new component, in which case $z_2 = 2$ and the parameters for this new component are sampled: $\theta^{(2)} \sim H$, for example, $\mu^{(2)} \sim \mathcal{N}(0,1)$. The new item is drawn from this component: $x_2 \sim F(\theta^{(2)})$, for example, $x_2 \sim \mathcal{N}(\mu^{(2)},\sigma^2)$. Importantly, the assignment that z_2 receives—that is, the decision about which component will be used to generate the sample—is defined by the following probability:

$$P(z_i = l \mid z_{1:i-1}) = \begin{cases} \frac{n_l}{i-1+\alpha} & \text{if } n_l > 0 \\ \frac{\alpha}{i-1+\alpha} & \text{if } n_l = 0, \end{cases} \tag{16}$$

where n_l is the number of items assigned to component l and α is the concentration parameter. Thus, components that have generated many previous items are more likely to be used to generate the new item; and larger values of α make it more likely that a new component will be added to the mixture. This generative process continues for $i = 3,4,\ldots,$ producing (a) a sequence of items x_i, (b) a component assignment index z_i for each item, and (c) an increasing number of components in the mixture. For n items, the component assignments $z_{1:n}$ represent a particular *partitioning* of the items over components. That is, some subset of $x_{1:n}$ will be "assigned to" (i.e., were generated from) component 1, some to component 2, and so on. Eq. 16 then defines the conditional distribution over such partitions.

Given the above DPMM generative process, we can compute the useful conditional probability $P(x_i \mid x_{1:i-1})$. This probability describes the distribution over possible values for a novel item x_i, given past examples $x_{1:i-1}$. It is useful because, as we shall see later, a slight variant will allow us to use DPMM for categorization. The computation requires three steps. First, the posterior probability over the hidden cluster assignments $P(z_{1:i-1} \mid x_{1:i-1})$ is computed. Second, the probability of the new index $P(z_i \mid z_{1:i-1})$ is computed. Finally, the probability of the new item $P(x_i \mid z_i)$ is computed. Multiplying these terms and marginalizing out $z_{1:i}$ and θ gives the desired $P(x_i \mid x_{1:i-1})$.

In categorization, we wish to predict the category label y_i for a novel item x_i after viewing a preceding sequence of $(x,y)_{1:i-1}$ pairs (i.e., the labeled training data). We may consider the labels y_i to be simply another feature dimension, replacing x_i with the extended vector (x_i, y_i), for instance, $x_i \in \mathbb{R}$ and $y_i \in \{-1, 1\}$ for two-category classification. In this case, θ is extended accordingly so that $\theta^{(l)} = (\mu^{(l)}, p^{(l)})$, where p is the parameter for a Bernoulli distribution (i.e., the head probability of a coin toss). Similarly, $F(x, y; \theta^{(l)}) = \mathcal{N}(x; \mu^{(l)}, \sigma^2) \cdot \text{Bernoulli}(y; p)$. Finally, the base distribution H can be a product of Gaussian and Beta distributions. With this extension, categorization can be viewed as a variant of the conditional probability above:

$$P(y_i = k \mid x_i, (x,y)_{1:i-1}) \propto \sum_{z_{1:i}} P(y_i = k, x_i \mid z_i) P(z_i \mid z_{1:i-1}) P(z_{1:i-1} \mid (x,y)_{1:i-1}), \qquad (17)$$

where we marginalize over the possible $z_{1:i}$ partitions, and θ is marginalized out implicitly. The first term on the right-hand side of Eq. 17 is defined by the distribution F. The second term, $P(z_i \mid z_{1:i-1})$, is the conditional distribution over partitions defined in Eq. 16. Exact computation of Eq. 17 requires the sum to be taken over all possible partitions $z_{1:i-1}$, resulting in an intractable combinatorial explosion. Instead, sampling methods such as *particle filtering* are often used, where a subset of all possible partitions is used to approximate $P(z_{1:i-1} \mid (x,y)_{1:i-1})$. A discussion of this technique goes beyond the scope of this paper and we instead refer the reader to Zhu et al. (2010); Bishop (2007), and the references therein.

1.3.2. Lifting the Rational model

Section 1.3.1 described how DPMMs can be used to model mixtures of components where the underlying number of components is unknown and can grow dynamically with the data. That was in the supervised case, where the data consist of labeled (x,y) pairs.

Just as in the semi-supervised exemplar and prototype models, lifting the DPMM requires modifications to accommodate both labeled (*x*,*y*) pairs and unlabeled *x* items with no corresponding ground truth labels *y*. Specifically, the probability of partition assignments $z_{1:i}$ in Eq. 17 assumes that all data consist of (*x*,*y*) pairs, that is, labeled data. What is needed is a method of calculating the probability distribution over partition assignments $P(z_{1:i-1} \mid (x,y)_{1:i-1})$ when some of the *y*'s are unlabeled. With Bayes rule, this conditional probability can be shown to be proportional to the product of three probabilities:

$$P(x_{1:i-1} \mid z_{1:i-1}) P(y_{1:i-1} \mid z_{1:i-1}) P(z_{1:i-1}). \tag{18}$$

In this equation, only the middle quantity $P(y_{1:i-1} \mid z_{1:i-1})$ depends on the label history $y_{1:i-1}$. The first quantity captures the probability distribution of both labeled a unlabeled items in the feature space *x*, while the third quantity is obtained from the definition of a DP in Eq. 16. The second quantity, then, is what differentiates supervised from SSL in this framework. If we let *L* be the set of indices between 1, …, *i*-1 that correspond to labeled data, then the unknown labels marginalize out, resulting in

$$P(y_{1:i-1} \mid z_{1:i-1}) = P(y_L \mid z_L). \tag{19}$$

As in the supervised case, the particle filtering method can be used to approximate the intractable sum over all possible $z_{1:i-1}$. Readers interested in more detail are again referred to Zhu et al. (2010).

The key point is that the probability distribution over partition assignments, which is central to the Rational/DPMM approach, is influenced here by the distribution of both labeled and unlabeled examples in the feature space, as well as by the labels given to the labeled items. Unlabeled data thus influence category learning by influencing which partitions of the feature space are most probable.

2. Experiments and model assessment

Before we consider whether the semi-supervised models of Section 1 provide candidate psychological models, we must first consider the evidence that human participants are affected by both labeled and unlabeled data when learning a categorization task. It may well be that people ignore and are unaffected by unlabeled examples for the purpose of categorization; or alternatively that they perform unsupervised clustering of unlabeled data and only use labeled data to figure out which labels "go with" which clusters. In either case, there would be little reason to consider models of SSL further. Here, we review three recent studies that investigated these questions.

Experiments 1 and 2 examined human learner sensitivity to the distributions of unlabeled examples in categorization. In both, participants first encountered a small set of items, each appearing with one of two possible labels, that together suggested a particular boundary between the two categories. Subsequently, participants classified a large set of unlabeled items sampled from a bimodal distribution in which the low-density region did not align with the boundary implied by the preceding labeled items. If participants learn from unlabeled items, their beliefs about the location of the category boundary should change with exposure to the unlabeled distribution. If learning is solely based on the supervised experience, beliefs about the boundary location should not change. If learning is based solely on unsupervised learning, beliefs about the boundary location should only reflect the unlabeled distribution and should not be influenced by the initial labeled experience.

2.1. Experiment 1

The first study was designed simply to assess whether human categorization decisions are influenced by the distribution of unlabeled examples (Zhu, Rogers, Qian, & Kalish, 2007). To our knowledge, this was the first study designed to explicitly address this SSL question. Twenty-two students at the University of Wisconsin completed a binary categorization task with complex novel shapes varying in a single continuous parameter $x \in [-2,2]$ as shown by the examples in Fig. 1. The two categories were denoted by $y = 0$ or $y = 1$. Participants first received two labeled items: $(x,y) = (-1,0)$ and $(1,1)$, repeated 10 times each in random order. These items were "labeled" in that feedback indicating the correct response was provided after each trial. Participants next classified 795 unlabeled test examples in one of two experimental conditions, differing only in how the majority of the unlabeled items were generated. In the L-shift condition, 690 of the unlabeled test items were drawn from a mixture of two Gaussians with a trough shifted

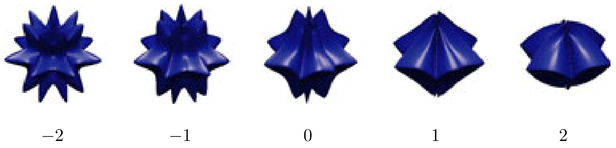


Fig. 1. Example stimuli used in Experiment 1, with corresponding x values.

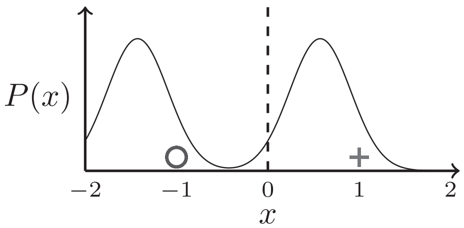


Fig. 2. Example of the data set used in the L-shift condition of Experiment 1. Labeled points are represented as negative (○) and positive (+). The black curve is the bimodal distribution $P(x)$ from which unlabeled items were drawn. The dashed vertical line represents the boundary implied by the labeled points alone. Note that the trough in the unlabeled distribution is shifted to the left with respect to the supervised learning boundary.

to the left of the boundary implied by the labeled training items (see Fig. 2). The other condition, R-shift, varied only in that the trough between the Gaussians was now shifted to the right of the implied labeled boundary. In both conditions, the remaining unlabeled test items were items drawn from a grid across the entire range of *x*, ensuring that both unlabeled distributions spanned the same range. The grid items appeared in random order at the beginning and end of the unsupervised phase, allowing us to measure the category boundary participants learned immediately following the supervised experience and following exposure to the unlabeled bimodal distribution. Methodological details are provided in Zhu et al. (2007).

Fig. 3 shows a summary of the results by pooling human behavior by condition and fitting logistic regression curves to show the conditional probability $P(y = 1 \mid x)$. Two subsets of the data are examined. The early subset shows behavior on the first 50 unlabeled test items (presented immediately after the labeled training phase), whereas the late subset shows behavior on the final 50 unlabeled test items (presented at the end of exposure to unlabeled data).

Comparing the early items, the two groups look essentially the same and the curves overlap. On the late items the curves are substantially different. The decision threshold, that is, the value of *x* producing $P(y = 1 \mid x) = 0.5$, shifted in opposite directions in the two conditions, moving to the left in the L-shift condition and to the right in the R-shift condition. In the late subset, the majority of participants classified the items $x \in [-0.07, 0.50]$ differently in the two conditions. If participants were unaffected by unlabeled data, the late test curves should be identical to the early curves and overlap. The fact that they do not indicates that participants *are* affected by the unlabeled data for this categorization task. To statistically test these observations, decision boundaries for the early and late grid-test items were computed separately for each participant using logistic regression on the participant's categorization decisions. A repeated measures analysis of variance assessing the influence of early versus late and L-shift versus R-shift on the location of the decision showed a significant interaction between the two factors

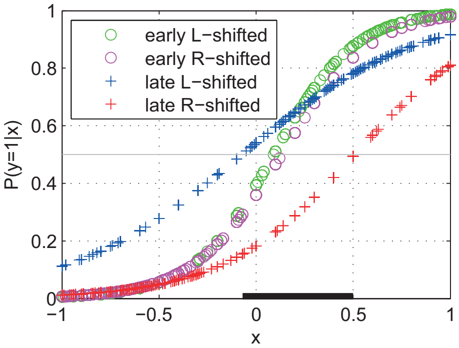


Fig. 3. Experiment 1 results from shift in unlabeled distribution. The thick black line marks items on which the majority of human categorization differs in the two conditions.

($F(1,18) = 7.82$, $p < 0.02$), indicating that after exposure to the unlabeled data, the decision boundary shifted in significantly different directions for the two groups. Thus, exposure to the unlabeled bimodal distribution appears to alter participants' beliefs about the location of the category boundary.

2.2. *Experiment 2*

The second study had a somewhat different goal—namely to investigate whether SSL might provide part of an explanation as to why people are often prone to form incorrect beliefs about social categories (Kalish et al., 2011). The experiment is useful for current purposes, however, because it revealed similar effects to those reported by Zhu et al. (2007) even though it used quite different stimuli and a different method for measuring the effect of unlabeled items. In this experiment, the unlabeled distribution was held constant, while the location of the original labeled examples varied across experimental groups.

Forty-three undergraduates viewed schematic images of women varying along the single dimension of width. The women were described as coming from one of two islands. As in Experiment 1, each participant first completed a supervised phase where a labeled example from each category (i.e., "Island") was presented five times in random order for a total of 10 labeled examples. In the L-labeled condition, participants viewed two relatively thin stimuli (pixel-widths of 80 and 115), whereas those in the R-labeled condition viewed two somewhat wider stimuli (pixel-widths of 135 and 165). All participants then classified a set of unlabeled items without feedback. In the experimental conditions, both L-labeled and R-labeled groups viewed the same set of unlabeled items, including 37 initial test items sampled from a uniform grid along the full stimulus range, 300 items sampled from a mixture of two Gaussian distributions, and a final set of 37 test items sampled from the grid. The mixture of Gaussians was constructed so that the modes of the distribution lay midway between the labeled points in the L-Labeled and R-labeled conditions (see Fig. 4). In a control condition, participants received the same L-labeled

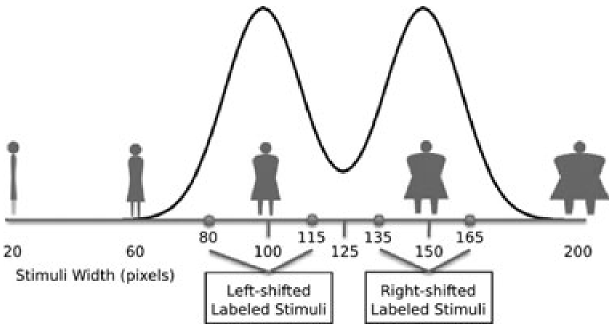


Fig. 4. Examples of the Island Women stimuli, the labeled points, and the bimodal distribution from which unlabeled items are sampled.

or R-labeled experience, but only viewed items lying on a grid between the two labeled items in the unsupervised phase.

In Zhu et al. (2007), the trough of the unlabeled distribution fell between the labeled points. In contrast, in this study the two labeled points both fell to one side of the trough in the unlabeled distribution, resulting in an even stronger conflict between the boundaries suggested by supervised and unsupervised experience. Given this mismatch, would learners still be affected by the unlabeled distributions? To answer this question, the authors considered three different measures. First, like Zhu et al. (2007), they considered how participants categorized unlabeled items along the grid prior to and following exposure to the bimodal unlabeled distribution. Second, following the unsupervised phase of the experiment, they asked participants to explicitly indicate where the boundary was located by adjusting a slider that controlled the width of a pictured stimulus. Finally, using the same slider, they asked participants to indicate the "most typical" example of each category.

All three measures showed beliefs about category structure to be strongly shaped by the distribution of the unlabeled examples. In the control condition, participant behavior strongly reflected their supervised learning experience: The estimate of the implicit category boundary and the participants' explicit reports of the boundary were closely aligned with and not significantly different from the midpoint between the labeled examples, while their judgments of the most typical example of each class aligned closely with and did not differ significantly from the labeled examples they had received. In comparison, implicit boundary estimates in the experimental groups were significantly shifted toward the trough in the unlabeled distributions—that is, toward the right in the L-labeled condition, and toward the left in the R-labeled condition. This shift was reflected even more strongly in the explicit boundary judgments. Moreover, choices about the most typical examples of each category aligned closely with the modes of the unlabeled distribution, shifting very dramatically away from the labeled items observed in the beginning of the experiment. Perhaps most interesting, the majority of participants in each condition actually altered their beliefs about one of the two labeled examples, coming to classify it with the opposite label than that viewed during the supervised phase.

Given these substantial effects of unlabeled data, one might inquire whether participants accurately remember the labeled examples and simply change their beliefs about the accuracy of the earlier supervisory feedback, or whether their memory for the labeled items itself changes. Kalish et al. (2011) addressed this question in a follow-up experiment where, following exposure to the unlabeled items, participants used the slider in an attempt to reproduce the two labeled items that had appeared at the beginning of the study. Strikingly, their reproduction were also strongly influenced by the unlabeled data, lining up closely with the two modes of the unlabeled distribution, even though, in actuality, the two labeled points lay on either side of one of the modes. Thus, memory for the original labeled examples appeared to be distorted by exposure to the unlabeled items.

One might further wonder whether the labeled experience has any impact at all in these studies beyond providing basic information about which "cluster" in the unlabeled distribution should get which label. Kalish et al. (2011) were able to show that the labeled information does, in fact, have a persisting influence even after extensive unlabeled experience: Despite being exposed to exactly the same unlabeled items, participants in the L-labeled and R-labeled conditions of these studies did not end up with exactly the same beliefs about the location of the boundary. Instead, the L-labeled group's final boundary was displaced significantly to the left of the R-labeled group's final boundary, indicating some lasting effect of the original supervisory experience.

Finally, this study rules out an alternative explanation of the effects of unlabeled data in these experiments. In the Zhu et al. (2007) study, because participants in the different experimental groups viewed different sets of unlabeled items, it was possible that the observed differences in categorization boundaries might arise from perceptual contrast effects. For instance, a given stimulus in that study might look "more pointy" or "less pointy" depending on how pointy the preceding stimulus was. It is conceivable that these local perceptual contrast effects might lead to consistent differences in the estimated category boundary depending on the location of the trough in the unlabeled distribution. In the study of Kalish et al. (2011), however, both experimental groups viewed the exact same set of unlabeled items, in the same fixed order, but nevertheless showed systematic differences in their estimate of the category boundary depending on their supervised experience. Thus, the learning in this study appears to be truly semi-supervised, reflecting contributions from both labeled and unlabeled experience.

2.3. Experiment 3

Although the preceding studies investigated the effects of the *distribution* of labeled and unlabeled examples, the final study we consider investigated effects of the *order* in which unlabeled items are encountered. Such effects have been previously reported in the literature, which generally shows that participants continue to learn from the unlabeled test items presented following a fully supervised learning session (Palmeri & Flanery, 1999, 2002; Zaki & Nosofsky, 2007). Here, we focus on work described by Zhu et al. (2010) because it employed the same "blob" stimuli used in Zhu et al. (2007). This allows for a clear comparison in the next section of how well the models developed in Section 1 account for the effects of both the distribution and ordering of unlabeled examples on human categorization.

In this study, 40 undergraduates learned to classify the stimuli used in Zhu et al. (2007) (Fig. 1). Similar to Experiments 1 and 2, participants first classified, with corrective feedback, two labeled items $(x,y) = (-2,0)$ and $(2,1)$ repeated five times each, followed by 81 unlabeled items (i.e., presented without feedback) on a regular grid between $x = -2$ and $x = 2$. The two experimental conditions were identical except for the order in which the unlabeled items were presented: In the L to R condition, test items were presented from smallest to largest values of x, while in the R to L condition this ordering was reversed. The central question was whether the participants' category boundary would differ depending upon the order in which the test items were presented.

Fig. 5 shows a plot of $P(y_i = 1 \mid x_i, (x,y)_{1:i-1})$, estimated by the fraction of subjects in each condition who classified x_i with label $y_i = 1$. The test item ordering clearly had a

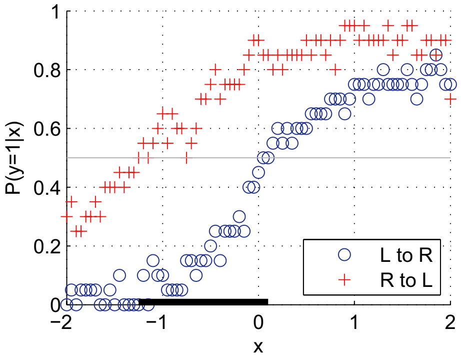


Fig. 5. Experiment 3 results from ordered unlabeled examples. The thick black line marks items on which the majority of human categorization differs in the two conditions.

strong effect on categorization. Participants in the L to R condition tended to classify more test items as $\hat{y} = 0$, while those in the R to L condition tended to classify more as $\hat{y} = 1$. A paired-sample t-test shows that the two conditions differed significantly ($p < 3 \times 10^{-7}$). Thus, the category boundary acquired appears to depend, not just upon the distribution of unlabeled items, but also upon the order in which these are encountered.

2.4. Model fitting experiments

These empirical studies indicate that, at least in the simple case of one-dimensional stimuli with two mutually exclusive categories, human category learning is sensitive to both the labeled data and the distribution and ordering of unlabeled data. In the remainder of this section, we consider how well the SSL models proposed in Section 1 fit the observed human behavior. We will focus on fitting the models to the data from Experiments 1 and 3, since these employed the same stimulus set and the same experimental procedure, allowing for comparison between model family as well as experiment. The central question we pose is how well, both qualitatively and quantitatively, the different SSL models described earlier can produce behavior that fits the observed human behavior.

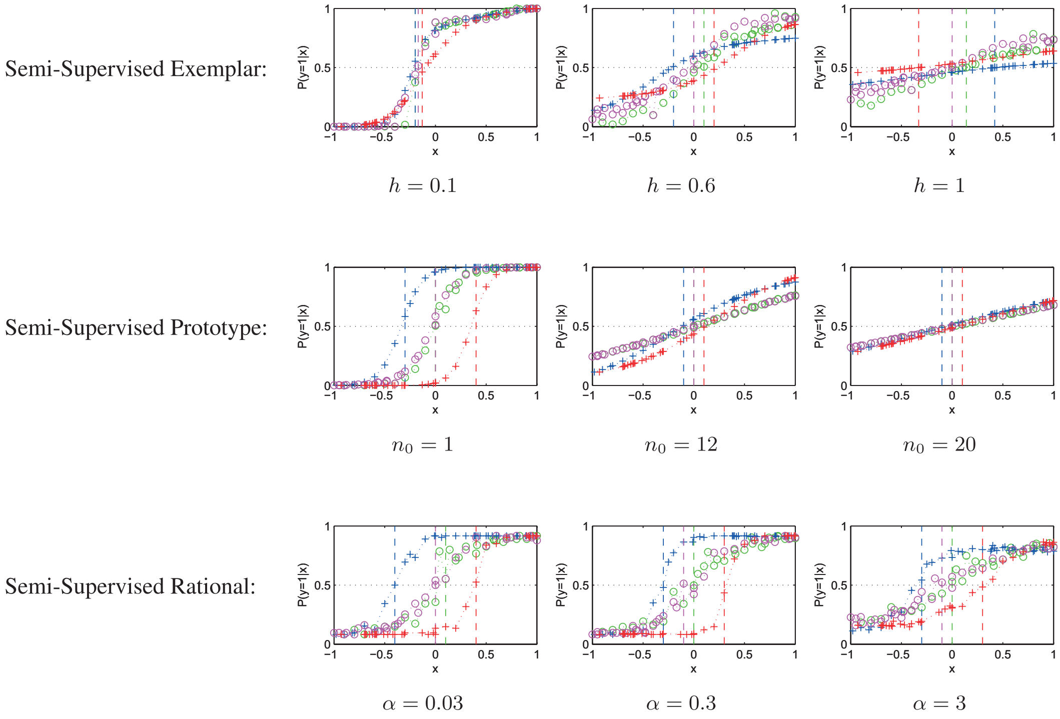
Before presenting the results, we must consider how best to parametrize the different models. In each of the three models we have considered, some parameters must be chosen by the theorist a priori. The subsequent behavior of the model will vary depending on the choices made. Thus, each of the three general models entail a *family* of associated models, with each individual model corresponding to a particular choice of parameters. The theorist interested in understanding which theoretical approach offers the best account of human SSL cannot just consider the fit of specific individual models within each type but must consider how best to adjudicate the different model families. In what follows, we illustrate that the different models can lead to quite different behaviors depending on the parametrization, with some models apparently providing a qualitatively better match than others. We then consider one method for empirically adjudicating the different model families, note some shortcomings of this approach, and sketch an alternative approach for use in future work.

Each model family potentially has several associated parameters. A full exploration of the parameter space for every model family is beyond the scope of this review. Instead, we will consider the effects of varying a single critical parameter for each model. For the semi-supervised exemplar model, the parameter we consider is h, the *bandwidth*, which determines how the influence of a stored example on a query item diminishes with distance. Small values of h mean that only very similar items are strongly weighted in the categorization decision, whereas larger values mean that distal items still receive substantial weight. For the semi-supervised prototype model, the critical parameter is n_0, a count of *pseudo-items* used to initialize the Gaussians for each category to be learned. This parameter effectively captures the strength of prior beliefs about the frequencies, modes, and variances of the two category prototypes. Higher values indicate stronger prior beliefs in the initial distributions, meaning that more evidence is required to shift the parameter estimates to reflect the clusters in the data. All Gaussian distributions are initialized to have a mean of zero and variance of one in these simulations. For the semi-supervised Rational model, the critical parameter is α, which controls the likelihood that each newly encountered item will begin a new cluster. When α is high, many clusters will be created, and when α is low, few clusters will be created.

Fig. 6 shows the behavior of each model family (rows), under different values of the critical parameter (columns), and fit to a data set similar to that used in Experiment 1 (Fig. 2). The middle column shows the parametrization that provides a good fit to the empirical data, while the other two columns show much smaller and larger parametrization to exhibit the variability of the models. Each item in the experiment is represented as a real-valued number x, and each label is coded as $y \in \{0,1\}$. The data set includes two labeled items at $x = -1$ and $x = 1$ followed by a large number of unlabeled data from the bimodal distribution. The curves show $P(y_i = 1 \mid x_i, (x, y)_{1:i-1})$ for the different values of x in the first 50 unlabeled test trials after 10 exposures to two labeled items (early), as well as the last 50 unlabeled test trials (late).

Two observations are of note. First, all three models show patterns of behavior qualitatively similar to those observed in the empirical data in Fig. 3: Following exposure to the unlabeled items, the decision curves shift toward the trough, that is, leftward in the left-shifted case and rightward in the right-shifted case. Second, the exact behavior of each model family does vary considerably depending on the choice of parameters, with some models varying more than others. All three models appear to provide a relatively good match to the empirical data under some but not other parameter choices.

The same models were also fit to the data set used in Experiment 3, which included two labeled items at $(x,y) = (-2,0)$ and $(x,y) = (2,1)$ followed by 81 unlabeled items lying on a grid in the range $(-2,2)$ and ordered either smallest to largest (L to R) or in the reverse direction (R to L). Fig. 7 shows the behavior of the three models under three different settings for the relevant critical parameter. All three models are clearly influenced by the ordering of the unlabeled items, with the direction of the effect consistent with that observed in the behavioral data and the magnitude varying substantially with



$h = 0.1$ $h = 0.6$ $h = 1$

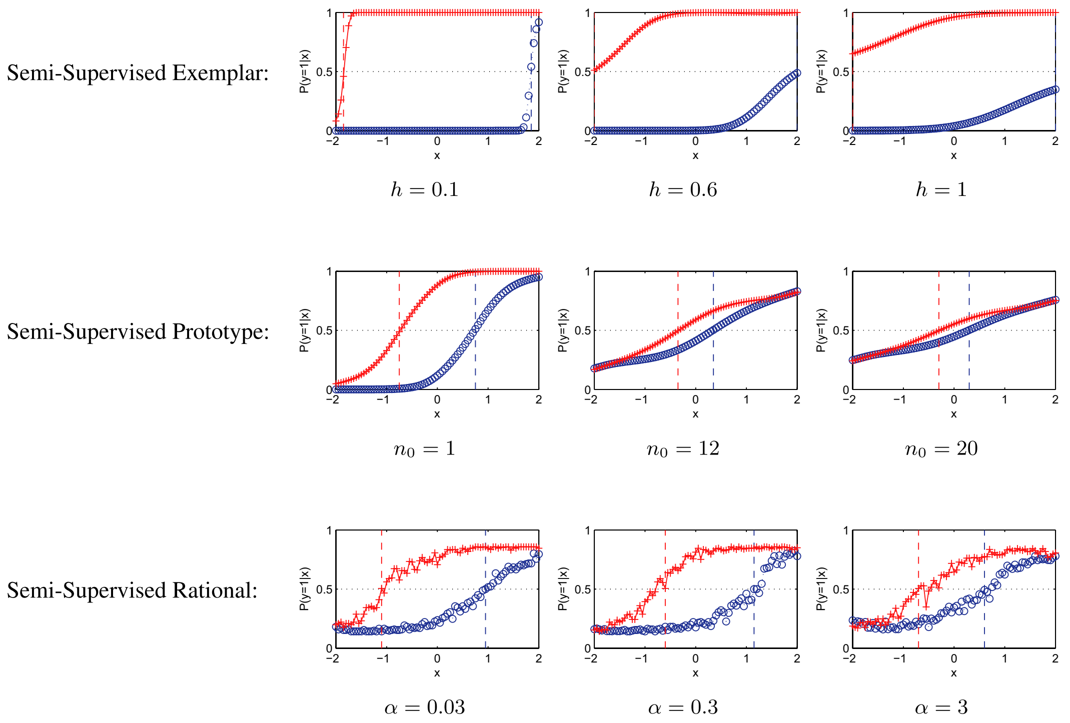
$n_0 = 1$ $n_0 = 12$ $n_0 = 20$

$\alpha = 0.03$ $\alpha = 0.3$ $\alpha = 3$

Fig. 6. Model predictions $P(y_i = 1 \mid x_i, (x, y)_{1:i-1})$ on the shifted distribution task of Experiment 1. Items in early represented by ○, in late by +. The L-shift condition is colored in blue, R-shift in red. Vertical dashed lines represent derived decision boundaries.

model choice and parametrization. In this case, however, one model family makes predictions that are grossly inconsistent with the observed data. Specifically, the semi-supervised exemplar model predicts, under each parametrization, that participants should *almost always guess the same label* in this experiment. To see this, note that in the R to L condition the conditional probability $P(y = 1 \mid x)$ is almost always larger than 0.5. In other words, when the test items are presented in decreasing order, the semi-supervised exemplar model assigns almost all test items the label $y = 1$. Similarly, in the R to L condition the model assigns almost all test items the label $y = 0$. This pattern is qualitatively different from that shown by the participants and by the semi-supervised prototype and Rational models, which all show a right-shifted boundary in the left-to-right ordering and a left-shifted boundary in the right-to-left condition. Thus, by inspection, the behavioral data appear to disconfirm the semi-supervised exemplar model but are qualitatively consistent with the semi-supervised prototype and Rational models.

We are left with the question of how best to quantitatively adjudicate the different model families. Here, we will consider the approach employed by Zhu et al. (2010), which uses a training/test set split procedure to quantitatively measure how well different models match the observed results in each task, a method commonly used in machine



$h = 0.1$ $h = 0.6$ $h = 1$

$n_0 = 1$ $n_0 = 12$ $n_0 = 20$

$\alpha = 0.03$ $\alpha = 0.3$ $\alpha = 3$

Fig. 7. Model predictions $P(y_i = 1 \mid x_i, (x,y)_{1:i-1})$ on the order task of Experiment 3. L to R condition represented by ○, R to L by +. Vertical dashed lines represent derived decision boundaries.

learning for model comparison. The data set under consideration is the conjoined results of Experiments 1 and 3—studies that employed the same stimuli and test procedures, and so which might reasonably be expected to involve the same model parametrization. The procedure involves dividing all the trials in these two experiments into a *training set* that includes 90% of trials selected at random, and a *test set*, which includes the remaining 10% of trials. Training involves examining the performance of each model family under a wide range of parameter settings and selecting the settings $\hat{\theta}$ that provide the best fit to the training dataset as measured by log likelihood $\ell_{train}(\hat{\theta})$. The resulting model for each type is used to generate predicted labelings for the test items, which are in turn compared to the aggregate experimental results for the corresponding items using log likelihood $\ell_{test}(\hat{\theta})$. This method of creating a training and test split of the data set results in a better comparison of generalization between models.

Table 1 shows the best-fitting parameter values $\hat{\theta}$ found using the training set and the associated log likelihoods $\ell_{test}(\hat{\theta})$ for each model family applied to the test set. Larger scores indicate a better fit. Under this measure, the prototype and Rational models provide a better match to the empirical data than does the exemplar model; and the Rational model shows an even better match to the data than does the prototype model.

Table 1
Log likelihood on the held aside test set

	Exemplar	Prototype	RMC
$\hat{\theta}$	$h = 0.6$	$n_0 = 12$	$\alpha = 0.3$
$\ell_{test}(\hat{\theta})$	−3,727	−2,460	−2,169

With regard to the question posed earlier, then, both the qualitative and quantitative analyses suggest three tentative conclusions. First, all three model families can provide a qualitatively good match to the behavioral data observed in some experiments, under some parametrization. Indeed, the three model families appear to provide approximately equally good fits to the data, considering just the best parametrization of each. These results thus establish the face validity of all three families in accounting for human SSL. Second, the model-fitting results for Experiment 3 establish that the different model families can make qualitatively different predictions about expected behavior in semi-supervised category learning experiments, only some of which are consistent with the observed behavior. Thus, SSL may prove a useful experimental paradigm for adjudicating between the different theories of human categorization. In the present case, for instance, evidence from Experiment 3 appears to disconfirm predictions of the semi-supervised exemplar model. Finally, the quantitative analysis above provides some evidence that, across both experiments, the semi-supervised Rational model provides a somewhat better fit than the alternative approaches.

These results must be viewed as preliminary in at least three respects, however. First, the approach we have sketched for quantitatively adjudicating model families, though in relatively common usage, has an important flaw. The reason is that the model families may not be matched for their complexity. Even though we have limited each model to a single tuning parameter, the intrinsic complexity of those models may still vary. The semi-supervised prototype (i.e., GMM) family, for instance, models each category with a single Gaussian, and consequently can only learn to form a single threshold decision boundary or an interval decision boundary in the 1D case. In the semi-supervised Rational (i.e., DPMM) model, in contrast, the number of Gaussian components can grow indefinitely, and as a consequence, this model is capable of learning multiple boundaries in the 1D case. So, though both model families have a single tuning parameter, the complexity and expressive power of the Rational model is greater. Richer model families are more likely to out-perform less complex model families on training data, but they are also more prone to overfitting. A fairer comparison of models, then, would control the complexity of the model families under consideration—a strategy we leave for future work.

Second, these analyses consider just two experiments, both with relatively simple one-dimensional, two-category learning problems. In contrast, standard variants of exemplar, prototype, and Rational models have been assessed against a great number of more complex category learning problems. It also remains for future work, then, to determine whether these preliminary results will apply across a comparably broad range of findings.

We emphasize that these results do not conclusively favor the lifted Rational over the lifted exemplar and prototype SSL models.

Third, although all three model variants do a good job of fitting some of the empirical data, there remain alternative explanations of the observed effects that do not require the assumption that human learning is semi-supervised. One hypothesis that has been repeatedly raised in discussion of this work is the possibility that participants learn and endeavor to retain important information from the supervised learning examples, beyond the location of the category boundary. For example, participants might learn from supervised experience that examples of the two categories occur about equally frequently. When subsequently classifying unlabeled items, they may take care to retain this information, ensuring that they assign category labels with equal frequency. Such a strategy would lead them to place a boundary in a location that divides the unlabeled items approximately in half, consistent with the behavior reported in Experiments 1 and 2. Or participants may notice from the supervised experience that both categories are about equally variable, and may assign labels to unlabeled examples to preserve this aspect of the supervised distributions. Future work must be conducted to formulate model that correspond to these hypotheses and to design experiments that discriminate them.

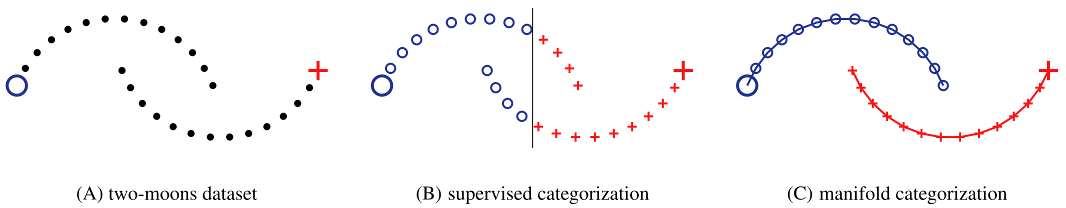
These caveats notwithstanding, there are two general conclusions we would like to draw. First, the preceding results establish the face validity of the general hypothesis that human category learning is semi-supervised—that is, influenced jointly by labeled and unlabeled experience—and also of the more specific hypotheses expressed in the formal models developed earlier. Second, because the different models can make differing predictions about human behavior in SSL scenarios, SSL may provide a new way of testing the theoretical commitments that underlie exemplar, prototype, and Rational models. A great deal of research in cognitive science has been concerned with understanding which method provides the most useful framework for understanding human categorization. This debate has been useful in elaborating computational descriptions of the different theories, but with the result that all three models seem capable of explaining much if not all of the available data, sometimes simply through the addition of parameters. In some respects this is not surprising: It is possible to show that, under appropriate parametrization, supervised variants of all three models will eventually converge on the same posterior probability distribution (i.e., the true posterior). However, as we have seen, this is not necessarily true of lifted variants of the same models. The models we have described rely on assumptions about how unlabeled items are distributed, and how labelings relate to these distributions. It can be shown that the lifted models also converge on the true posterior, but only so long as the corresponding assumptions are valid. When the assumptions are not valid, the different semi-supervised models can fail, in somewhat different ways. The test-item ordering effects reported above represent one such failure: All of the models assume that observations are independently and identically distributed, whereas this is clearly not the case when the test items are ordered. As a consequence of this violation of an assumption, the different models show different behaviors depending on the ordering of the unlabeled items; and these effects provide the basis for figuring out which model framework best accounts for human behavior. Other violations of SSL assumptions—such as, for instance, the assumption that the true category boundary will lie in the "gap" between clusters in feature space, which was violated in Experiment 2—may likewise provide opportunities for understanding which SSL models best help to explain human behavior.

3. Other SSL assumptions

The preceding sections focused on semi-supervised models that adopt the mixture-model assumption since these bear the most transparent relationship to familiar models in psychology. There are, however, several other kinds of assumptions that can be adopted to support SSL (Chapelle et al., 2006; Zhu & Goldberg, 2009). These include large separation assumptions (that items which are separated by a large gap in feature space tend to belong to different categories, such as in Semi-Supervised Support Vector Machines Chapelle, Sindhwani, & Keerthi, 2008; Joachims, 1999; Lawrence & Jordan, 2005; Vapnik, 1998), and the multi-view assumption (that the features can be split into separate "views" that tend to be conditionally independent given the category, and separate learners on each view may cooperate; one example being Co-Training; see Blum & Mitchell, 1998). In this section, we will briefly consider another such assumption, the *manifold* assumption, to illustrate how these models might lead to new hypotheses about how humans use unlabeled data in categorization tasks.

3.1. Manifolds

The manifold assumption, a graph-based method, holds that (a) data are distributed along an underlying lower-dimensional manifold in the feature space, and (b) category membership propagates along this manifold. This propagation may produce labelings that conflict with the labelings suggested by Euclidean proximity in the original feature space. To see this, consider the classic "two-moons" data set shown in Fig. 8A. The data consist of items distributed in a 2D continuous space, with two labeled points and a set of unlabeled points. A simple supervised learner that classifies items solely on the basis of their similarity to labeled points in the 2D space (such as, for instance, the nearest-neighbor


(A) two-moons dataset

(B) supervised categorization

(C) manifold categorization

Fig. 8. The two-moons data set, with ○ and + representing labeled points and black dots representing unlabeled points. (B) and (C) represent the categorizations learned by a supervised linear-boundary learner and a semi-supervised manifold learner, respectively.

classifier using Euclidean distance) will produce the classification shown in Fig. 8B. Under the manifold assumption, however, the solution shown in Fig. 8C is more appropriate: Each category is associated with a one-dimensional manifold embedded in the two-dimensional space. This assignment produces a classification solution that violates Euclidean proximity in the 2D space. For instance, the leftmost point of the bottom "moon" is more similar to the blue than to the red labeled item in the 2D space, but it is classified as belonging to the red category. Note that this is not a simple case of linear dimension-reduction, where all the data can be projected onto a single dimension without loss of information. Standard linear dimension-reduction methods like principal components analysis will not recover the two categories shown in Fig. 8B. Instead, this behavior arises from the fact that the data lie along one-dimensional manifolds embedded in the 2D space. If the data are assumed to be so distributed, the key learning task is to find the most likely manifold in the original feature space—a task for which unlabeled examples can provide important information. When the manifold has been learned, it can be used to determine how category labels should generalize. Unlabeled items can be classified based on their proximity to labeled items *along the manifold*, rather than their proximity in the original space.

This example is somewhat abstract, but there are many real-world problems in cognitive science for which the manifold assumption might prove useful. Consider, for instance, the fact that face recognition is largely viewpoint-invariant. When shown a person's face from the front and told his or her name, the learner can still identify the person from a three-quarter profile. One does not need to be given separate (picture, name) pairs from every possible angle to successfully identify the individual. Several different models have been advanced in the literature to explain viewpoint-invariant face recognition, and in many cases these correspond to variants of the exemplar and prototype models familiar from earlier sections (Tarr & Bulthoff, 1998). The manifold assumption permits another kind of hypothesis. Specifically, the visual images corresponding to different views of a given face may lie on a lower-dimensional manifold along which the name label is able to propagate. Fig. 9 demonstrates this idea. Two sets of images were taken of the heads of two individuals as they rotated one full turn, starting from looking at the camera, to looking directly away from the camera and back to directly at the camera again. Each image is composed of a set of pixels, each associated with a single

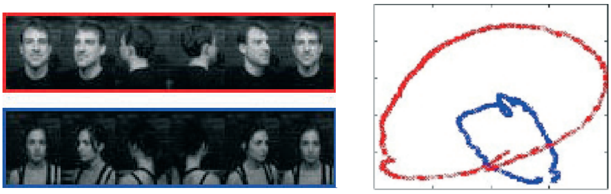


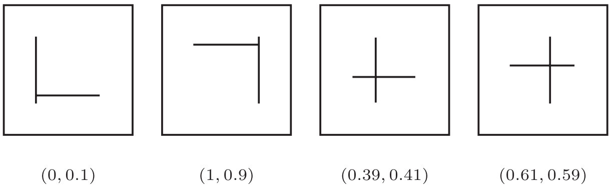
Fig. 9. On the left are examples from collections of pictures taken of the heads of two individual rotating in space. On the right are the 2D representations of each collection, denoted by color, where each point corresponds to a single image in a collection.

continuous "intensity" value. Thus, each image can be considered a point in a continuous space with each pixel corresponding to a dimension. With this conception, we can visualize the distribution of images in this high-dimensional space using Principle Components Analysis (PCA) to reduce each set to two dimensions. For each set, the points form a clear one-dimensional manifold in two-dimensional space, in this case a circle. It is easy to imagine a label associated with a front-facing image propagating along this circle so that identification can still be made without seeing additional labels for each point. It is also easy to see that a labeled image of one individual will tend to propagate to other images of the same individual, which lie along the same manifold, rather than to images depicting the other individual, even if these images are closer in the encompassing feature space.

More generally, many problems in cognitive science require the learner to figure out the "right" similarity relations existing among a set of stimuli in a given domain; and in many cases the relations that need to be learned are not transparently available directly from the sensory structure of the environment. The manifold assumption in SSL provides one way of learning relationships that do not correspond to a simple linear dimensional reduction of the encompassing high-dimensional space, and of using these relations to determine how category labels should generalize. Thus, it is of considerable interest to know whether people are capable of using manifolds to generalize category labels. Gibson, Zhu, Rogers, Kalish, and Harrison (2010) designed a set of experiments to understand under what conditions humans are capable of SSL using manifolds in a synthetic setting. The results suggest that manifold-learning is possible, but only when strong hints are provided and there is no alternative, simpler explanation of the data.

3.2. Experiment 4: Learning manifolds

In this experiment, 139 participants were asked to provide a binary label for a set of "cards" shown on a computer screen after viewing a small set of labeled examples. The feature space consisted of two dimensions $x_1, x_2 \in [0,1]$. The stimuli consisted of two crossing lines, one vertical and one horizontal, with x_1 specifying the position of the vertical line left to right horizontally and x_2 the position of the horizontal line from bottom to top vertically. Examples of the stimuli are shown in Fig. 10. Participants viewed all of the unlabeled items together on the screen at one time and "labeled" these items by dragging them to a bin on either the left or right side of the screen.



$(0, 0.1)$ $(1, 0.9)$ $(0.39, 0.41)$ $(0.61, 0.59)$

Fig. 10. Examples of stimuli used in the manifold experiment with associated (x_1, x_2) feature values.

Participants performed this task in one of two conditions differing only in how the unlabeled items were distributed. In the *manifold* condition, items were distributed in a manner similar to the two-moons dataset described above (moons$^{\mathrm{U}}$), whereas in the *uniform* condition items were distributed on a uniform grid over the stimulus space (grid$^{\mathrm{U}}$). The authors also manipulated the number of labeled items shown (two in 2^l vs. four in 4^l), and whether close neighbors in the stimulus space were visually highlighted in the display (indicated by h). In the manifold condition with two labeled points, the ends of the manifolds are labeled, just as in Fig. 8A. In the same condition with four labeled points, two additional labels were added at points inconsistent with a single linear boundary in either dimension. The same labeled points were employed in the uniform condition.

The resulting data sets are shown in Fig. 11. When given just two labeled points, participants did not use the manifold to generalize the label, even when hints to the structure were provided by highlighting close neighbors in the feature space. Instead, these participants always learned a linear axis-parallel category boundary consistent with the labeled points. When four labeled points were provided, but no hints were provided via highlighting, participants again failed to follow the manifold when labeling, although they did abandon the single linear boundary solution since it did not match the labeled data. Only when four labels were provided and close neighbors were highlighted in the display (4^lmoons$^{\mathrm{U}}$h) did subjects exhibit manifold learning behavior.

One possible interpretation of this result is that participants did not learn the manifold at all, but simply selected whichever cards were shown by the highlighting to be similar to the preceding card. Three points of evidence contradict this interpretation, however. First, despite the highlighting, participants in 2^lmoons$^{\mathrm{U}}$h condition failed to learn the manifold—only when hints were combined with sufficient labeled evidence (as in the 4^lmoons$^{\mathrm{U}}$h condition) did the manifold pattern emerge. If participants were simply following the highlighting, manifold learning should have been observed in both conditions. Second, if participants were simply following the highlighting, they should never classify items that were not highlighted by the user interface. In fact, these "un-highlighted" classifications were observed fairly frequently—an average of 17 of 78 categorizations—by participants in the 4^lmoons$^{\mathrm{U}}$h condition. Third, in a control experiment conducted with identical labeled and unlabeled points, but with the neighbor

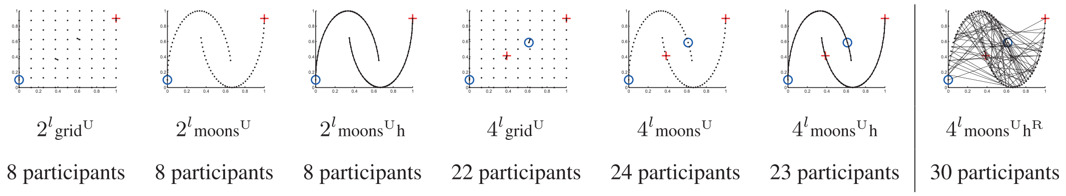


Fig. 11. The six conditions for the manifold experiment, followed by the "random neighbor" experimental condition. Large symbols indicate labeled items; dots indicate unlabeled items. Highlighting shown to participants is represented as graph edges.

relations used for highlighting randomized (see the $4^{l}\mathrm{moons}^{\mathrm{U}}\mathrm{h}^{\mathrm{R}}$ condition in Fig. 11), participants essentially ignored the highlighting—suggesting that, in the original study, participants must have a sense of the underlying manifold and did not blindly follow the highlighting.

This study provides some preliminary evidence that when labeled evidence rules out the preferred single-linear-boundary solutions, and when the task provides some perceptual cues, people are capable of using unlabeled data to learn non-linear manifolds. Semi-supervised learning models developed under the manifold assumption may provide hypotheses about learning processes in such scenarios. Of course, the particular cues to manifold structure used in this study—highlighting near neighbors in the graphical display—were highly artificial, as was the task itself (sorting cards bisected by orthogonal lines). It is not difficult, however, to think of more realistic real-world cues to manifold structure. In the case of 3D object recognition, for instance, people presumably have considerable prior knowledge about how objects are capable of moving in space. Participants may be able to better employ manifold structure in category learning if provided with learning scenarios that allow them to exploit such knowledge—for instance, if told that the different images are "snapshots" of two objects as they move around in some environment. Thus, the import of manifold-based SSL for problems in cognitive science remains to be assessed in future work. More generally, there exist a variety of other semi-supervised assumptions in machine learning that likewise may provide a source of hypotheses about ways in which human beings can capitalize on unlabeled data to support category learning.

4. Some challenges for models of SSL

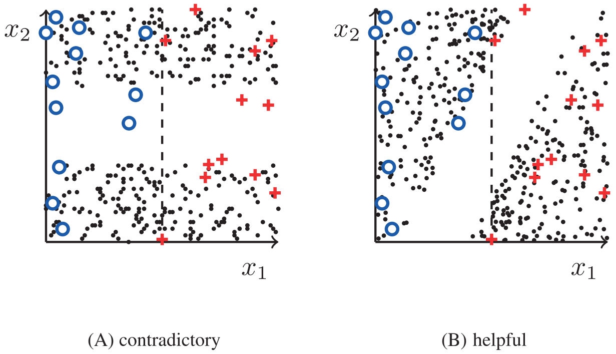
The preceding sections suggest that human behavior in categorization tasks can be influenced by both labeled and unlabeled experiences, in ways that are qualitatively consistent with some SSL models. In this final section, we consider some of the ways in which human behavior does not seem to be well accounted for by standard SSL models.

The issues we identify mainly stem from the consideration of SSL in higher dimensional stimulus spaces. With the exception of the manifold-learning work, the experiments and models described earlier assumed that the items to be categorized varied along a single dimension (or at least along a 1D subspace in a multidimensional feature space). Although generalization of the models themselves to higher dimensional spaces is straightforward, the application of these ideas to human SSL raises some challenges that are not clearly met by current machine learning models.

4.1. Feature weighting and SSL

One issue concerns the question of feature-selection or feature-weighting: When stimuli vary in many different ways, human beings have the ability to selectively attend to or weight various feature dimensions differently when making a categorization decision. Feature-weighting (and its extreme form, feature selection) is, of course, an issue that has been extensively studied in both machine learning and cognitive science. Such approaches, however, generally rely upon the fact that learning is fully supervised: Features receive a strong weighting when they are useful for discriminating among categories, and they receive lower weightings when they are not. In SSL, it is less clear how feature weighting should be handled: The unlabeled distribution of items provides clues about how the items should be partitioned, but should the learner attend only to the distribution along feature dimensions that, from the labeled items, seem important to the categorization task, or should the full distribution in the feature space matter?

To illustrate the issue, consider the learning problems in Fig. 12. In both, labeled and unlabeled items are situated in a 2D feature space, and the labeled items (colored symbols) suggest a vertical boundary in x_1. In the left panel, unlabeled items are distributed uniformly across the space, but with a substantial gap in x_2. The gap in the distribution is orthogonal to the boundary suggested by the labeled data—thus, a semi-supervised learner may be prone to "selecting" the wrong feature dimension (x_2) if she is heavily influenced by the distribution of unlabeled items. In this case, it seems that the right thing to do is to use the labeled examples to determine the feature weights—for instance, heavily weighting x_1—and to ignore the distribution on the other dimension. This approach, however, will not work well with the distribution shown in the right panel. Here, the labeled items suggest the same boundary in x_1, but the distribution of unlabeled items suggests that an oblique boundary in both x_1 and x_2 might be more appropriate. A semi-supervised learner that weights feature dimensions based solely on labeled examples will neglect important distributional information in this case. The question of how best to weight feature dimensions in SSL—and the empirical question of how human beings do so—remains to be addressed.



(A) contradictory

(B) helpful

Fig. 12. An example of a contradictory unlabeled distribution on the left, and a helpful unlabeled distribution on the right. Points labeled negative are represented by ○, points labeled positive by +, and points unlabeled by black dots. The labeled boundary is indicated by a dashed line.

4.2. Integral and separable feature dimensions

A second challenge arises from the fact that human beings do not treat all perceptual feature dimensions equally. In some cases, perceptual dimensions that are independent in principle appear to be "coupled" in human experience—it is difficult for people to respond to one such dimension without also being influenced by others (Shepard, 1991).

One common example of such *integral* dimensions is the saturation and brightness of a given color where one dimension cannot be varied without a perceived change in the other (the first row of Fig. 13). Other perceptual features are easily separable: People can selectively process information on one such dimension without being greatly influenced by the other. The frequency and orientation of a Gabor patch, for instance, constitute psychologically *separable* dimensions (the second row of Fig. 13).

It has long been known that separability of stimulus feature dimensions can strongly influence behavior in supervised category-learning tasks where all training items are labeled. For instance, when dimensions are separable, SL models better fit human data if the distance between stimuli is measured using a city-block rather than a Euclidean metric (Nosofsky, 1987).

Zhu, Gibson, and Rogers (2011) recently demonstrated, however, that the distinction may have a further influence on categorization behavior in the context of SSL. In a simple 2D semi-supervised task, participants learned to classify two labeled items in a short supervised phase, then categorized a large number of unlabeled items. The data were composed of four clusters arranged in a "diamond" configuration, as seen in Fig. 14. Two labeled items appeared at the center of the left-most and top-most clusters and thus were consistent with three single linear boundaries: a vertical boundary in x_1, a horizontal boundary in x_2, or an oblique boundary in both x_1 and x_2.

When the stimulus dimensions were separable (frequency and orientation of Gabor patches, as in the second row of Fig. 13), participants usually selected an axis-parallel boundary—only 17 of 45 participants selected an oblique boundary that took both feature dimensions into account. In contrast, when the dimensions were integral (brightness and

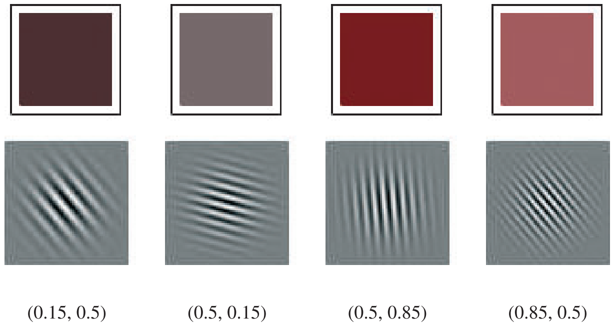


Fig. 13. Examples of color stimuli varying in brightness and saturation (first row) and Gabor patch stimuli varying in frequency and rotation (second row), with the corresponding values (x_1, x_2) in feature space.

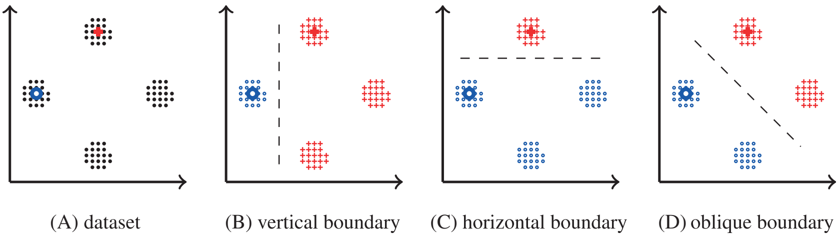


Fig. 14. The data set used in Zhu (2011), with two labeled points (○ and +) and unlabeled items arranged in four clusters. The three plots to the right show three possible labelings of the unlabeled data produced using a single linear boundary, each consistent with the labeled points.

saturation of a colored square, as in the first row of Fig. 13), all participants (34 of 34) selected an oblique boundary. This result suggests that, when dimensions are separable, participants may attend selectively to only a subset of the stimulus properties, without noticing or processing values along other dimensions. If this is so, there are obvious implications for SSL—specifically, learning may not be influenced by the distribution of unlabeled items along unattended stimulus dimensions. This in turn suggests that SSL may be less apparent in multi-dimensional tasks with separable stimulus dimensions: If the interesting distributional information from unlabeled items happens to be carried by stimulus dimensions that are unattended/unselected, then this information will not be available to influence learning. Such a result was reported by Vandist, De Schryver, and Rosseel (2009), who failed to find evidence for SSL that employed Gabor patches varying in frequency and orientation as the feature dimensions. On the other hand, participants may be more strongly influenced by the full distribution of unlabeled items in the feature space when stimulus dimensions are integral.

Rogers, Kalish, Gibson, Harrison, and Zhu (2010) recently reported results consistent with this hypothesis. Participants learned to classify bisected circles that varied in the orientation of the bisecting line and the radius of the circle (Nosofsky, 1986). In a semi-supervised condition, 32 labeled learning trials were intermixed with 400 unlabeled trials drawn from a distribution with a prominent gap aligned with the true category boundary. In two control conditions, the same labeled items were presented all together in a single block, or intermixed with unrelated filler trials. In this setup, a semi-supervised learner might be expected to learn the correct boundary fastest in the experimental condition, because the gap in the unlabeled distribution supplements the labeled experiences to provide an important cue to the location of the boundary. To the contrary, performance over the course of learning, and on a subsequent set of unlabeled test trials drawn from a grid, was no different, and relatively poor, across the three conditions, with several participants placing the boundary on the irrelevant stimulus dimension.

Size and orientation are clearly separable psychological dimensions to which participants can selectively attend. Thus, one explanation of this failure is that, when participants selectively attend to the irrelevant dimension, they fail to "notice" or be influenced by the trough in the unlabeled distribution on the unattended dimension, thus reducing the influence of SSL overall. To test this hypothesis, the authors conducted a second experiment identical to the first except that participants were pressured to respond very rapidly, within 400 ms of stimulus onset. Allocation of selective attention is commonly thought to require at least 500 ms (Wolfe & Horowitz, 2004), so the authors reasoned that this time-pressure would prevent learners from selectively "screening out" information on the unattended dimension. They therefore predicted a larger effect of SSL in the speeded condition. Consistent with this prediction, the authors observed a substantial SSL effect: Accuracy improved more rapidly, with a larger proportion of participants learning to criterion, in the semi-supervised condition compared to both control conditions. Moreover, performance was better overall when participants had limited time to process and respond to the stimulus—suggesting that the unlimited response window in the first experiment was actually hindering learning. These results suggest that, when stimulus dimensions are separable, participants may selectively attend to some dimensions and thus may be less influenced by the distribution of unlabeled items along unattended dimensions. By definition, it is difficult to selectively focus on one dimension for integral-dimension stimuli. This suggests that SSL may have generally larger effects for such items, though this hypothesis remains to be directly assessed.

More generally, machine learning has no analog to the distinction between integral and separable dimensions. An important future challenge will be to better understand how such models might be adapted to account for these kinds of effects in people.

4.3. Perceptual spaces are shaped by unlabeled distributions

A third related issue is that the perceived similarity of novel stimuli is not static in human cognition but can vary substantially with learning and with perceptual adaptation. Again, such effects have been the focus of extensive study in the context of fully supervised category learning (Goldstone, 1994) but have additional implications in the context of SSL. Specifically, recent evidence suggests that perception can be radically re-shaped by statistical structure among unlabeled stimuli over quite short periods of time. Stilp and colleagues exposed participants to complex sounds varying in two psycho-acoustic dimensions (attack/decay and spectral shape) carefully matched for perceptual discriminability (Stilp, Rogers, & Kluender, 2010). Sounds were sampled frequently from a diagonal line in the space, and only occasionally from the orthogonal line, thus inducing strong covariance between the two dimensions. After 7 min of passive exposure to these sounds, participants performed an AXB discrimination task for pairs of sounds sampled either along the original line or orthogonal to it. Immediately following the passive exposure, participants were "deaf" to differences between sounds varying orthogonally to the direction of experienced covariation—as though their auditory system collapsed the original two dimensional space to a single dimension, aligned with the first principal component of the perceptual space. With further exposure to the test sounds, sensitivity to the orthogonal dimensional gradually recovered to pre-exposure levels.

There are, of course, many examples of perceptual change following extensive learning with some novel stimulus domain. The Stilp et al. (2010) study adds to this literature in three ways that may be especially challenging for developing models of SSL. First, the learning was incidental—participants were given no instructions, and indeed conducted a distracting task (playing with an Etch-A-Sketch) while listening to the sounds. Second, the period of exposure was only 7 min, much shorter than the standard categorization study, suggesting that the perceptual structure of a given stimulus space might be changing at the same time that participants are learning to partition the space. Third, the study suggests that such perceptual effects might go beyond simply learning how to weight different stimulus dimensions. In Stilp et al. (2010), the perceptual space appeared to "collapse" from two dimensions to a single dimension on the basis of the strong covariance between the two dimensions, so it was not just the weights on the dimensions that were changing but the dimensionality of the space itself.

These effects pose challenges for applying machine learning models to human categorization, because the models typically assume that stimuli can be represented as static points within a feature space of fixed dimensionality. Indeed, many of the parameters in the General Context Model that have no counterpart in our lifted exemplar model were introduced by Nosofsky (2011) as a way of capturing both feature weighting and categorization in a single framework. Of course, machine learning also encompasses a broad variety of models for efficiently extracting or representing the information contained in multidimensional distributions. The manifold-learning assumption and related models addressed earlier provide one way of finding interesting low-dimensional structure within a high-dimensional space, for instance. The interesting challenge for future work will be to consider how these models might be combined with category learning models to simultaneously capture both their gradual learning about category membership and the perceptual changes that arise when they are exposed to structured unlabeled items.

5. Conclusion

It seems clear to us that the course of daily life provides a wealth of unlabeled experience relevant to categorization, over and above the occasional explicit labeling experiences we receive. It also seems reasonable to suppose that such experiences are important in shaping our category knowledge, but most formal models of human categorization have not taken such influences into account. SSL models developed in the context of machine learning can provide a rich source of hypotheses about how human SSL might proceed, and recent empirical results suggest that people do combine labeled and unlabeled learning experiences in ways that are highly consistent with some of these models. Moreover, different models can make differing predictions about how human beings should behave in SSL contexts, and so may provide a way of adjudicating which categorization theories are most useful for understanding human behavior. To this point, however, such work has mainly focused on learning in simple one-dimensional two-category problems, and there remain critical challenges to generalizing the approach to more realistic multi-dimensional tasks. Future work in this vein will need to focus on understanding feature-weighting and feature-separability in an SSL context.

Acknowledgments

We thank Chuck Kalish, Richard Shiffrin, Robert Nosofsky, Josh Tenenbaum, Bilge Mutlu, Martha Alibali, Joseph Harrison, Chen Lang, and Richard Qian for their helpful discussions and help with experiments. This research is supported in part by AFOSR FA9550-09-1-0313, NSF IIS-0953219, IIS-0916038. BUVI (1990).

Notes

1. This property, where the number of exemplars grows with the number of examples, makes KDE a *non-parametric* model, which is distinguished from *parametric* models such as GMMs, which are defined by a fixed number of parameters.
2. This version assumes a component distribution jointly over the feature vector *x* and label *y*, so that a single cluster can produce examples with different label *y* value. An alternative approach would be to assume that a separate model for each category, as in the A Class model of Mansinghka, Roy, Rifkin, and Tenenbaum (2007).

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

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

Balcan, M.-F., & Blum, A. (2010). A discriminative model for semi-supervised learning. *Journal of the ACM*, *57*(3), 19:1–19:46

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*(2), 458–475.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In P. Bartlett and Y. Mansour (Ed.), COLT '98: Proceedings of the eleventh annual conference on Computational learning theory (pp. 92–100). New York: ACM.

Chapelle, O., Sindhwani, V., & Keerthi, S. S. (2008). Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, *9*(Feb), 203–233.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.) (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234–257.

Gibson, B. R., Zhu, X., Rogers, T. T., Kalish, C.W., & Harrison, J. (2010). Humans learn using manifolds, reluctantly. In J. Lafferty (Ed.), *Advances in neural information processing systems 23*, vol. 24, (pp. 730–738). Red Hook, NY: Curran Associates, Inc.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *123*(2), 178–200.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In D. S. McNamara & J. G. Trafton (Ed.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 323–328). Austin, TX: Cognitive Science Society.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2011). Nonparametric bayesian models of categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 173–198). Cambridge, UK: Cambridge University Press.

Gureckis, T. M., & Love, B. C. (2003). Towards a unified account of supervised and unsupervised category learning. *Journal of experimental and theoretical artificial intelligence*, *15*(1), 1–24.

Hampton, J. A. (1993). Prototype models of concept representation. In I. Van Mechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67–95). San Diego, CA: Academic Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93* (4), 411–428.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th international conference on machine learning* (ICML 1999) (pp. 200–209). San Francisco, CA: Morgan Kaufmann.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, *120*(1), 106–118.

Kruschke, J. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Lawrence, N. D., & Jordan, M. I. (2005). Semi-supervised learning via Gaussian processes. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (pp. 753–760). Cambridge, MA: MIT Press.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review*, *9*(4), 829–835.

Love, B. C., Medin, D., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Mansinghka, V. K., Roy, D. M., Rifkin, R., & Tenenbaum, J. (2007). AClass: An online algorithm for generative classification. In M. Meila and X. Shen (Eds.), Journal of Machine Learning Research - Proceedings Track, vol. 2 (pp. 315–322). Brookline, MA: Microtome Publishing.

Medin, D., & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review; Psychological Review*, *85*(3), 207.

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 40–64). Cambridge, UK: Cambridge University Press.

Murphy, G. L., & Smith, E. E. (1982). Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior*, *21*(1), 1–20.

Myers, J. (1976). Probability learning and sequence learning. In W. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Application*, *9*, 141–142.

Neal, R. M. (1998). Markov chain sampling methods for dirichlet process mixture models (Tech. Rep. No. 9815). Department of Statistics, University of Toronto.

Nosofksy, R. M., & Palmeri, T. J. (1997). An exemplar-based random-walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Nosofsky, R. M. (1984). Choice, similarity and the context theory of classification. *Journal of experimental psychology: Learning, memory and cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception and Psychophysics*, *38*(5), 415–432.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *13*(1), 87–108.

Nosofsky, R. M. (1991). The relation between the rational model and the context model of categorization. *Psychological Science*, *2*(6), 416–421.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge, UK: Cambridge University Press.

Palmeri, T. J., & Flanery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science*, *10*, 526–530.

Palmeri, T. J., & Flanery, M. A. (2002). Memory systems and perceptual categorization. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory*, vol. 41, (pp. 141–189). San Diego: Academic Press.

Posner, M., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77* (3), 353–363.

Pothos, E., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26*(3), 303–343.

Rogers, T. T., Kalish, C. W., Gibson, B. R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2320–2325). Austin, TX: Cognitive Science Society.

Rosch, E., Mervis, C. B., Gray, D. W., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun (Ed.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 761–731). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on nosofsky. *Journal of Experimental Psychology: General*, *115*, 58–61.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Smith, E. E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 800.

Stilp, C. E., Rogers, T. T., & Kluender, K. R. (2010). Rapid efficient coding of correlated complex acoustic properties. In R. Schekman (Ed.), *Proceedings of the national academy of sciences*, vol. 107, (pp. 21914–21919). Washington, DC: National Academy of Sciences.

Stutz, J., & Cheeseman, P. (1996). Autoclass – a bayesian approach to classification. In J. Skilling & S. Sibisi (Eds.), *Maximum entropy and bayesian methods* (pp. 117–126). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Tarr, M. J., & Bulthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. *Cognition*, *67*, 1–20.

Teh, Y. W. (2010). Dirichlet processes. In C. Sammut and G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 280–287). New York: Springer.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*(7), 309–318.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception, & Psychophysics*, *71*(2), 328–341.

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 2277–2282). Mahwah, NJ: Lawrence Erlbaum Associates, Inc..

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the allocation of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*, 1–7.

Zaki, S. R., & Nosofsky, R. M. (2007). A high-distortion enhancement effect in the prototype-learning paradigm: Dramatic effects of category learning during test. *Memory & Cognition*, *35*, 2088–2096.

Zhu, X., Gibson, B. R., Jun, K., Rogers, T. T., Harrison, J., & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In J. Fürnkranz and T. Joachims (Eds.), *The 27th international conference on machine learning (ICML-10)* (pp. 1247–1254). Haifa, Israel: Omnipress.

Zhu, X., Gibson, B. R., & Rogers, T. T. (2011). Co-Training as a human collaboration policy. In W. Burgard and D. Roth, (Eds.), *The 25th conference on artificial intelligence (AAAI-11)* (pp. 852–857). Menlo Park, CA: The AAAI Press.

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, *3*(1), 1–130, Morgan & Claypool.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans perform semi-supervised classification too. In R. C. Holte and A. Howe (Eds.), *Proceedings of the 21st conference on artificial intelligence (AAAI-11)* (pp. 864–870). Menlo Park, CA: The AAAI Press.