Modeling Categorization as a Dirichlet Process Mixture

Kevin Canini

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-69
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-69.html

May 18, 2007

Abstract

I describe an approach to modeling the dynamics of human category learning using a tool from nonparametric Bayesian statistics called the Dirichlet process mixture model (DPMM). The DPMM has a number of advantages over traditional models of categorization: it is interpretable as the optimal solution to the category learning problem, given certain assumptions about learners' biases; it automatically adjusts the complexity of its category representations depending on the available data; and computationally efficient algorithms exist for sampling from the DPMM, despite its apparent intractability. When applied to the data produced by previous experiments in human category learning, the DPMM usually does a better job of explaining subjects' performance than traditional models of categorization due to its increased flexibility, despite having the same number of free parameters.

1 Introduction

Despite years of progress in machine learning, the general problem of categorization remains unsolved. Fortunately, many tasks in the field of cognitive science can be phrased in terms of categorization, so there is a wealth of data available about the dynamics of categorizers who perform quite well. Hopefully, these areas of study can complement each other, with data collected from human subjects informing more intelligent machine learning algorithms, which in turn inspire new theories about the workings of the human mind.

The problem of category learning is typically posed as follows: given a sequence of $N-1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, \ldots, x_{N-1})$ and category labels $\mathbf{c}_{N-1} = (c_1, \ldots, c_{N-1})$ and an unlabeled stimulus N with features x_N, we would like an algorithm for assigning stimulus N to a category that produces results as similar as possible to that of a human categorizer. Note that this is a separate problem from learning the best-performing categorizing algorithm in an objective sense. Because human performance on this task depends on several factors, including differences between individual subjects and the particular experimental methodology, it seems that adequately explaining human behavior in general is beyond our reach. However, exploring the advantages and disadvantages of particular models in isolated contexts will hopefully shed some light on the underlying processes of the human mind.

Many algorithms have been proposed to solve the categorization problem, such as learning a decision boundary [5] and searching for deterministic rule-based category descriptions [12]. Most approaches have featured some combination of two very prominent ideas: (i) new stimuli are compared to the previously-seen stimuli (the *exemplars*) from each category, and (ii) new stimuli are compared to a central stimulus (the *prototype*) of each category, which need not be explicitly encountered during training. These two general approaches were introduced by Medin and Schaffer [8], and Posner and Keele [13], respectively. For example, the ALCOVE algorithm [7]

combines the exemplar approach with a neural network to tune the parameter weights automatically. The Varying Abstraction Model (VAM, [21]) attempts to bridge these two approaches, taking the form of an exemplar model, a prototype model, or something in-between, depending on the value of a free parameter.

The marriage of these psychological models with Bayesian statistics has given rise to a new generation of *rational* models of categorization, which attempt to cast human cognitive behavior as the optimal solutions to appropriate computational problems posed by the environment. In this framework, categorization can be solved by performing Bayesian inference with reasonable prior distributions on category structures. This idea was first introduced by Anderson in creating the Rational Model of Categorization (RMC, [2, 3]). Following Anderson's methodology, we introduce the Dirichlet process mixture model of categorization, which inherits the flexibility of the RMC and improves upon its weaknesses.

The remainder of the paper is organized as follows: in Section 2, I detail three previous psychological models: the exemplar, prototype, and VAM. In Section 3, I describe how traditional models of categorization can be interpreted as density estimation schemes, I introduce three rational models of categorization – including the Dirichlet process mixture model (DPMM) – and I mention an efficient scheme for sampling from the DPMM. In Section 4, I present the results of applying the DPMM to data from various prior experiments, and I conclude in Section 5.

2 Psychological models of categorization

Psychological models based on exemplars and prototypes can be described as a special case of the following framework: given $N-1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, \ldots, x_{N-1})$ and their associated category labels $\mathbf{c}_{N-1} = (c_1, \ldots, c_{N-1})$, the probability that a new stimulus N with features x_N belongs to some category j is given by

$$P(c_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{\eta_{N,j}\beta_j}{\sum_j \eta_{N,j}\beta_j} \tag{1}$$

where $\eta_{N,j}$ is the similarity of the stimulus N to the category j and β_j is the response bias for category j. The key difference between the models is the way they calculate the $\eta_{N,j}$ quantities.

2.1 Exemplar models

In an exemplar model, a category is represented by all of its stored instances (*exemplars*). The similarity of stimulus N to category j is calculated by summing the similarity of the stimulus to all stored instances of the category. That is,

$$\eta_{N,j} = \sum_{i|c_i=j} \sigma_{N,i}$$

where $\sigma_{N,i}$ is a symmetric measure of the similarity between the two stimuli with features x_N and x_i. It can take any form that is convenient for a particular task, but it is usually defined as a decaying exponential function of the distance between the two stimuli as per [17], that is,

$$\sigma_{N,i} = \exp(-\delta^{\alpha}_{N,i})$$

When $\alpha = 1$, the similarity decays exponentially with the distance. When $\alpha = 2$, the similarity decays according to a Gaussian bell curve with the distance. Finally, the distance $\delta_{N,i}$ between two stimuli is typically a weighted sum of the difference on each dimension of the psychological space:

$$\delta_{N,i} = c \left(\sum_d w_d \, | \, x_{N,d} - x_{i,d} |^r \right)^{1/r}$$

where c is a scaling parameter, and r specifies which distance measure to use ($r = 1$ corresponds to city-block distance, $r = 2$ corresponds to Euclidean distance, etc.). Note that as $c \to 0$, the exemplar model tends to assign a new stimulus to the largest category, and as $c \to \infty$, a new stimulus is assigned to the category of its single closest neighbor.

As an example, consider the situation depicted in Figure 1, where the unknown stimulus (denoted by a gray circle) is compared to every instance of a category (denoted by 'X's) to determine its similarity to the category. The computational complexity and memory demands of this model can become a problem as categories grow larger. Modifications to this standard approach must be made in situations where previous data is extremely abundant and decisions need to be made very quickly. However, it has been shown to explain human performance very well in many experiments, especially when memory demands are minimal and ample time is allowed for decisions to be made.

2.2 Prototype models

In a prototype model, a category j is represented by a single prototypical instance. In this formulation, the similarity of a stimulus N to category j is defined to be

$$\eta_{N,J} = \sigma_{N,p_j}$$

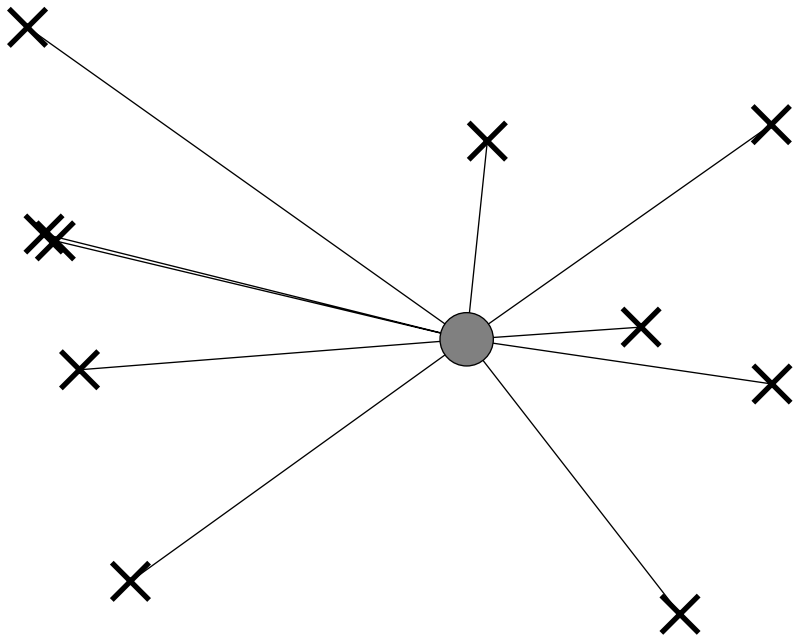

Figure 1: Determining category similarity with an exemplar model involves comparing the new stimulus to every stored instance of the category.

where η_{N,p_j} is a measure of the similarity between stimulus N and the prototype p_j of category j, defined as in the exemplar model. The category prototype is typically defined to be the center of all the instances of the category:

$$p_j = \frac{1}{N_j} \sum_{i|c_i=j} x_j$$

with N_j being the number of stimuli assigned to category j.

As an example, consider the situation depicted in Figure 2, where the unknown stimulus (denoted by a gray circle) is compared only to the category prototype (denoted by a white square) to determine its similarity to the category. The prototype is the centroid of all instances of the category.

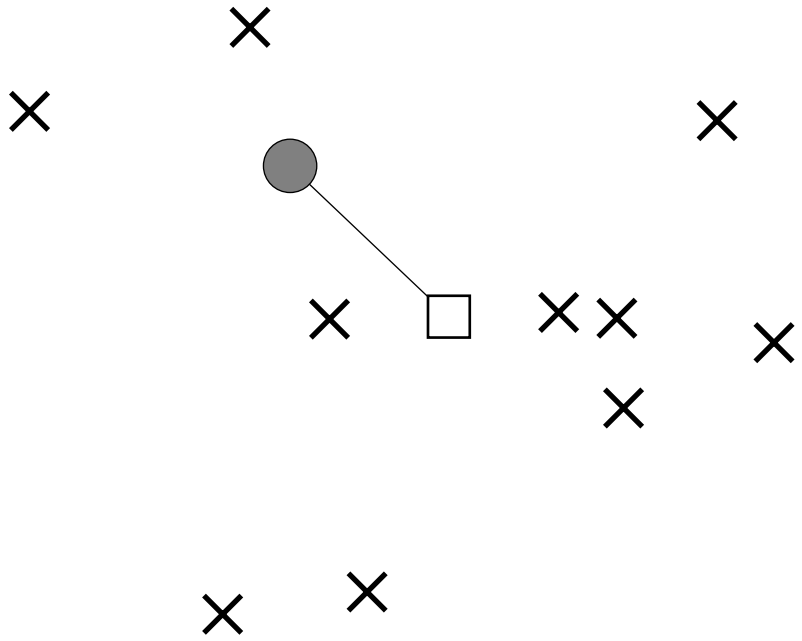

Figure 2: Determining category similarity with the prototype model involves comparing the new stimulus to the category prototype.

2.3 Comparison of exemplar and prototype models

Exemplar and prototype models both have strengths and weaknesses. Prototype models are more cognitively plausible, since it is usually difficult for a person to remember the exact composition of every stimulus ever encountered, but it is reasonable to assume that a prototypical instance near the category average can be inferred and stored.

Furthermore, exemplar models can potentially overfit the training data. If either of the parameters c or α is too large, the local surroundings of a new stimulus will be given too much importance in comparison to the global trends of the data. Furthermore, exemplar models are more sensitive to mislabeled data points that happen to be nearby the test stimulus.

However, exemplar models have the advantage of allowing for more expressive category boundaries. Prototype models are typically restricted to convex, unimodal distributions, while exemplar models can naturally create arbitrarily complicated dis-

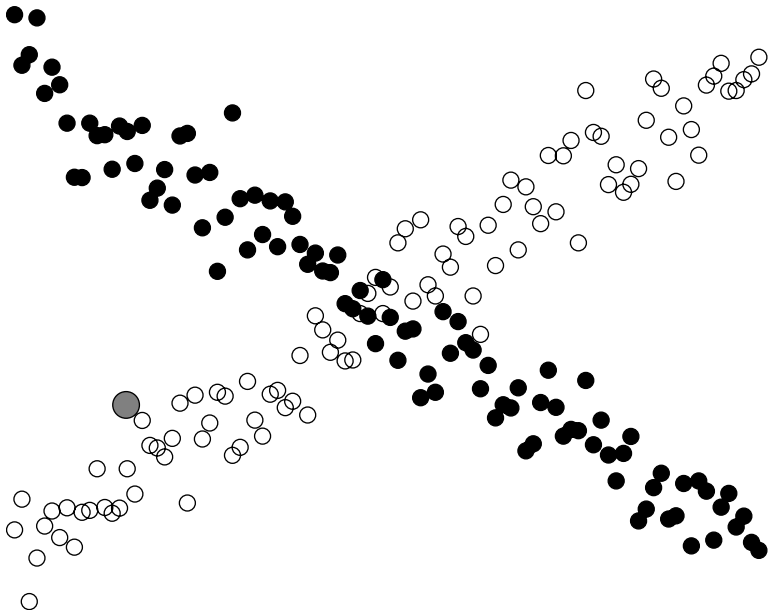

Figure 3: The two categories have the same prototype, so a basic prototype model would not be able to distinguish between them.

tributions as the data warrants.

As an example, consider the situation depicted in Figure 3, where a prototype model wouldn't be able to differentiate between the two categories, whose prototypes would be nearly identical, making differentiation very difficult. An exemplar model, however, would correctly classify the test stimulus.

On the other hand, consider the situation depicted in Figure 4. Assuming the true category boundary is linear, a prototype model would correctly classify the unknown stimulus, while an exemplar model might incorrectly classify it because of the nearby instances of the white category.

2.4 The Varying Abstraction Model

Realizing that these two models are at opposite ends of a spectrum, Vanpaemel et al. [21] showed that we can formalize a set of interpolating models by allowing the instances of each category to be partitioned into clusters, where the number of clusters K_j in category j ranges from 1 to N_j, the number of instances of the category. Then

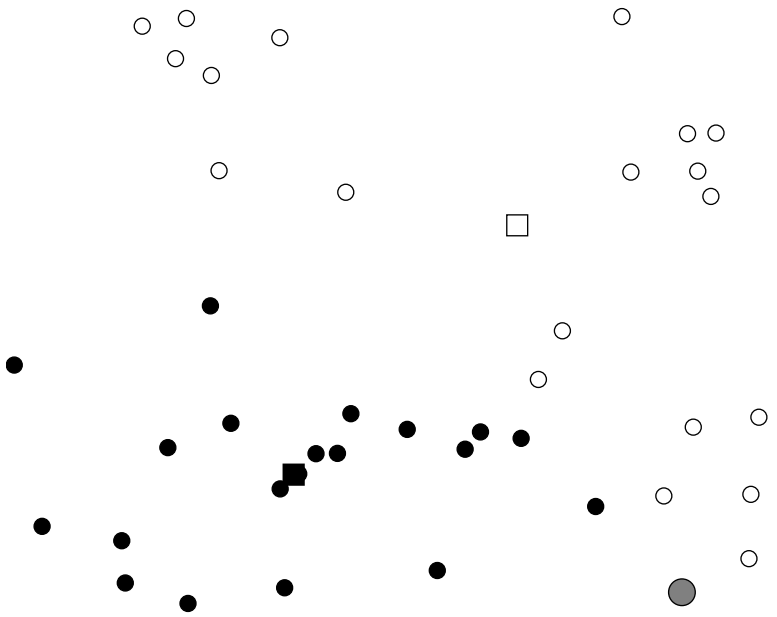

Figure 4: An exemplar model might place too much importance on the nearby instances of the white category, overshadowing the global trend of the data.

each cluster is represented by a prototype, which is defined to be the centroid of all the instances of the cluster, and the similarity of stimulus N to category j is defined to be

$$\eta_{N,j} = \sum_{k=1}^{K_j} \eta_{N,p_{j,k}} \tag{2}$$

where $p_{j,k}$ is the prototype of cluster k in category j. When $K_j = 1$ for all categories j, this is equivalent to the prototype model, and when $K_j = N_j$ for all categories j, this is equivalent to the exemplar model. Thus, this generalized model, the Varying Abstraction Model (VAM), is more flexible than both the prototype and exemplar models, so it will be able to outperform each one in both objective performance and matching human performance. The drawback to the VAM is that the parameter space is exponentially large, since we must choose a partition for each category. Any cognitively plausible model of categorization must have an acceptable computational complexity.

While the VAM provides a model with which we can interpolate between the pro-

totype and exemplar models, it provides no cognitively plausible method for choosing a partition of the category instances into clusters. Unfortunately, simply searching over all possible partitions carries an exponential computational cost and is intractable for even modestly-sized data sets. Moreover, this strategy ignores any possible biases that human learners may have towards particular types of partitions.

3 Rational models of categorization

The psychological models discussed in Section 2 attempt to explain human categorization in terms of the cognitive processes being used. They make use of similarity functions defined on pairs of stimuli that are justified in terms of psychological plausibility. In contrast to this method, we now consider *rational* models of categorization, following the example of Anderson [2]. Rational models describe the task of categorization as the optimal solution to a computational problem posed by the environment, rather than attempting to describe the underlying cognitive process being used. The models are described using ideas from Bayesian statistics, which allows us to use insights from statistical machine learning to create efficient algorithms to solve them.

As in Section 2, assume we are given a set of $N-1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, \ldots, x_{N-1})$ and their associated category labels $\mathbf{c}_{N-1} = (c_1, \ldots c_{N-1})$. Then we can find the probability that a new stimulus N with features x_N belongs to some category j by applying Bayes' rule as follows:

$$P(c_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{P(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) P(c_N = j | \mathbf{x}_{N-1}, \mathbf{c}_{N-1})}{\sum_j P(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) P(c_N = j | \mathbf{x}_{N-1}, \mathbf{c}_{N-1})} \tag{3}$$

This yields an equation of the same form as Equation (1). We can compare the similarity measure $\eta_{N,j}$ to the likelihood of stimulus N being generated by category j, $P(x_N | c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$, and the category bias β_j can be thought of as the prior probability of category j, $P(c_N = j | \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$. In fact, if we constrain the variables without any loss of generality so that $\int_{x_N} \eta_{N,j} = 1$ and $\sum_j \beta_j = 1$, then they become probability functions and the correspondence works out exactly.

Ashby and Alfonso-Reese [4] showed that the three models described in Section 2 correspond to particular solutions of the density estimation problem in statistics. In particular, the definition of η_{N,p_j} in the prototype model corresponds to estimating the

category likelihood distribution $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$ by assuming it comes from a known family of distributions and determining the most likely parameter values, which are characterized by the prototype p_j. This is an instance of a well-known technique in the statistics community called parametric density estimation.

Likewise, the definition of $\eta_{N,j}$ in the exemplar model corresponds to approximating the distribution $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$ as the average of N_j distributions, one centered on each exemplar. Thus, we have

$$P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{1}{N_j} \sum_{i|c_i=j} P_i(x_N)$$

where $P_i(x_N)$ is a distribution centered on x_i and decreasing as the distance from x_i increases. The distributions are assumed to be symmetric, so $P_i(x_N)$ can be written simply as $f(x_i, x_N)$. In the statistics literature, $f(\cdot, \cdot)$ is known as a kernel function, and this process is known as nonparametric kernel density estimation.

3.1 The Mixture Model of Categorization

Just as the VAM interpolates between traditional exemplar and prototype models, the Mixture Model of Categorization (MMC) introduced by Rosseel [14] interpolates between parametric and nonparametric density estimation. In this model, the probability $P(x_N|c_N = j, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})$ is represented as a mixture of K_j distributions. Just as in the VAM, the category instances are partitioned into clusters z, with each cluster distribution being drawn from a known family of distributions parameterized by the cluster prototype. So we have

$$P(x_N|c_N = j, \mathbf{x}_{n-1}, \mathbf{c}_{N-1}) = \sum_{k=1}^{K_j} P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) P(z_N = k|\mathbf{z}_{N-1}, c_N = j, \mathbf{c}_{N-1})$$

When each category is represented by a single cluster, this model reduces to parametric density estimation (corresponding to the prototype model), and when each stimulus has its own cluster, it reduces to nonparametric kernel density estimation (corresponding to the exemplar model). Ashby and Alfonso-Reese [4] showed that the MMC corresponds to the VAM, with the appropriate definition of $\eta_{N,p_{j,k}}$ in Equation (2).

Like the VAM, the MMC lacks a method for choosing a clustering of the category stimuli. This issue is addressed by the RMC and DPMM, which provide efficient methods for choosing category partitions that are rationally justifiable.

3.2 The Rational Model of Categorization

Anderson's Raional Model of Categorization (RMC, [2, 3]) specifies that the stimuli are partitioned into clusters, as in the VAM. Instead of choosing among all possible clusterings, though, each stimulus is assigned to a cluster in turn, according to a greedy maximum a-posteriori rule. The probability that stimulus N is assigned to cluster k is, according to Bayes' Rule,

$$P(z_N = k|x_N, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \frac{P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})P(z_N = k|\mathbf{x}_{N-1}, \mathbf{z}_{N-1})}{\sum_k P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})P(z_N = k|\mathbf{x}_{N-1}, \mathbf{z}_{N-1})}$$

where $P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$ is the likelihood of generating the stimulus from cluster k, and $P(z_N = k|\mathbf{x}_{N-1}, \mathbf{z}_{N-1})$ is the prior probability that the stimulus originated from cluster k. Since stimuli are assigned to clusters in turn, k ranges over the values of the already-existing clusters, and a new, empty cluster. The cluster likelihood $P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$ can take any useful form, and is defined in [3] such that the values on the individual dimensions of x_N are independent. The prior

distribution over z_N is defined as follows:

$$P(z_N = k|\mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \begin{cases} \frac{cM_k}{(1-c)+c(N-1)} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{(1-c)}{(1-c)+c(N-1)} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases}$$

where M_k is the number of stimuli previously assigned to cluster k and the parameter c is called the coupling probability.

While the RMC is computationally efficient, its clustering algorithm is locally greedy, and there is no guarantee of the quality of the resulting partition, since it produces different cluster assignments depending on the order in which stimuli are encountered.

3.3 Dirichlet process mixture model

The Dirichlet process mixture model (DPMM) as presented in [10, 15] is akin to the RMC in that it provides a rational account for partitioning stimuli into clusters. However, its roots in Bayesian statistics provide an efficient way for sampling from it that is asymptotically exact, as opposed to the locally greedy maximum a-posteriori algorithm given for the RMC.

We assume that the stimuli $\mathbf{x}_{N-1} = (x_1, \ldots x_{N-1})$ belong to a particular category j and are generated from a cluster such that the features of the stimulus are dependent only on cluster membership. That is, the distribution over stimuli features, given a particular cluster with parameters θ_i, is

$$x_i \,|\, \theta_i \;\sim\; F(\theta_i)$$

where F is some parameterized distribution. The distribution over the θ_i parameters, i.e., the distribution over clusters, is equal to some distribution G drawn from the Dirichlet process $DP(G_0, \alpha)$, where G_0 is the base distribution of the Dirichlet process

and α is the concentration parameter for the base distribution.

$$\begin{aligned} \theta_i \,|\, G &\sim G \\ G &\sim DP(G_0, \alpha) \end{aligned}$$

With probability 1, the realization of a Dirichlet process is a discrete distribution, so the DPMM can be thought of as a countably infinite mixture model of category j, where the θ_i parameters select among the component distributions [10]. Accordingly, we can think of the values of θ_i as determining a partition of the stimuli. In practice, the values of θ_i themselves can be integrated out, with the parameters to the cluster distributions being determined solely by their constituent members.

Once G_0 and α are specified, we have a complete model for the probability of generating stimulus N from category j. Summing over all possible partitions of the category instances into clusters, we find that

$$P(x_N|\mathbf{x}_{N-1}) = \sum_{\mathbf{z}_N} P(x_N|\mathbf{x}_{N-1}, \mathbf{z}_N) P(\mathbf{z}_N|\mathbf{x}_{N-1})$$

where z_i is the cluster to which stimulus i is assigned. We can use Bayes' rule to write

$$P(\mathbf{z}_N|\mathbf{x}_{N-1}) = \frac{P(\mathbf{x}_{N-1}|\mathbf{z}_N)P(\mathbf{z}_N)}{\sum_{\mathbf{z}_N} P(\mathbf{x}_{N-1}|\mathbf{z}_N)P(\mathbf{z}_N)} \tag{4}$$

In the DPMM, the prior probability $P(\mathbf{z}_N)$ on a partition is defined as

$$P(\mathbf{z}_N) = \frac{\alpha^K}{\prod_{i=0}^{N-1}(\alpha + i)} \prod_{k=1}^{K} (M_k - 1)!$$

where α is the concentration parameter for the DPMM, K is the number of clusters, and M_k is the number of stimuli assigned to cluster k. This prior distribution arises when cluster assignments are chosen incrementally using the Chinese Restau-

rant Process (CRP, [1]). Consider a Chinese restaurant with an infinite number of infinitely-large tables. The first customer sits at the first table, and each incoming customer afterwards sits at a given table with probability proportional to the number of people already at that table, and starts a new table with probability proportional to α. The tables represent clusters within the partition, and the customers represent the individual stimuli. So

$$P(z_N = k|\mathbf{z}_{N-1}) = \begin{cases} \frac{M_k}{N-1+\alpha}, & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{N-1+\alpha}, & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases}$$

Neal [10] pointed out that Anderson's rational model of categorization is equivalent to a DPMM where α, the strength of the base distribution, is equal to $(1-c)/c$, and the latent variables assign stimuli to discrete clusters. This insight allows us to solve a rational model of cognition using statistical inference algorithms designed for use with DPMMs, rather than using Anderson's locally greedy algorithm. For example, Gibbs sampling and particle filtering allow us to perform approximate Bayesian inference on this model without incurring the exponential computational cost.

3.4 Gibbs sampling

Markov chain Monte Carlo is a scheme whereby a Markov chain is constructed so that the asymptotic distribution over its states $\mathbf{x}$ is equal to some desired distribution $f(x)$. After iterating through the Markov chain many times, samples of the desired distribution can be approximated by sampling the states of the Markov chain.

Using Gibbs sampling, a form of Markov chain Monte Carlo, we can approximately sample from the posterior distribution $P(\mathbf{z}_N|\mathbf{x}_{N-1})$ without having to consider all exponentially many possible partitions. The Markov chain is constructed to have the set of all possible partitions as its state space, and from any initial state, we

repeatedly cycle through each stimulus and reassign it to a cluster sampled from the distribution given below. The chain should eventually converge to the desired posterior distribution. For each stimulus x_i, we reassign it to a cluster by sampling from

$$P(z_i|\mathbf{z}_N\backslash z_i, \mathbf{x}_N) \propto P(x_i|\mathbf{z}_N, \mathbf{x}_N\backslash x_i)P(z_i|\mathbf{z}_N\backslash z_i)$$

where $\mathbf{z}_N\backslash z_i$ is the cluster assignments for stimuli $1, \ldots, i-1, i+1, \ldots, N$ and $\mathbf{x}_N\backslash x_i$ is the features of stimuli $1, \ldots, i-1, i+1, \ldots, N$. Due to the exchangeability property of the DPMM, the probability $P(z_i|\mathbf{z}_N\backslash z_i)$ can be computed by assuming that stimulus i is the last one to be chosen, and all the other cluster assignments have been made. So by the Chinese Restaurant Process, we can assign z_i by sampling from

$$P(z_i = k|\mathbf{z}_N\backslash z_i) = \begin{cases} \frac{M_k}{N-1+\alpha}, & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{N-1+\alpha}, & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases}$$

4 Experimental results

In order to test the DPMM's ability to model category learning, it was fit to human data from several experiments conducted by Smith and Minda in 1998 [18] and Nosofsky et al. in 1994 [11].

4.1 Smith and Minda 1998

Smith and Minda ran several experiments with human subjects in order to capture the dynamics of humans learning to differentiate between two small categories. Their key result was that the prototype model had a better fit to human performance than the exemplar model during the early stages of learning, but a worse fit during the later stages. Thus, their experiments provide a good opportunity to test the DPMM's ability to automatically find the best interpolating point between the exemplar and prototype models.

4.1.1 Experiment 1

The first experiment was set up as follows: subjects were presented with a series of stimuli, each in the form of a six-letter nonsense word. The words can be represented as bit-strings, where each bit determines which of two possible letters occurs at that position within the word. There were 14 distinct stimuli, 7 of which were designated as Category A, with the other 7 designated as Category B. Two different category structures were used in separate parts of the experiment, denoted LS (linearly separable) and NLS (not linearly separable). The stimuli used for Experiments 1:LS are listed in Table 1. The categories in Experiment 1:LS are linearly separable, meaning that each stimulus can be correctly categorized by taking a linear combination of the values on its feature dimensions. Members of a category differ from its prototype (000000 and 111111, for Category A and Category B, respectively) on at most 2

Category A		Category B	
000000	banuly	111111	kepiro
010000	benuly	111101	kepilo
100000	kanuly	110111	keniro
000101	banilo	101110	kapiry
100001	kanulo	011110	bepiry
001010	bapury	101011	kapuro
011000	bepuly	010111	beniro

Table 1: Categories A and B from Smith & Minda 1998, Experiments 1:LS and 2:LS

Category A		Category B	
000000	gafuzi	111111	wysero
100000	wafuzi	011111	gysero
010000	gyfuzi	101111	wasero
001000	gasuzi	110111	wyfero
000010	gafuri	111011	wysuro
000001	gafuzo	111110	wyseri
111101	wysezo	000100	gafezi

Table 2: Categories A and B from Smith & Minda 1998, Experiments 1:NLS and 2:NLS

dimensions.

The stimuli used for Experiments 1:NLS are listed in Table 2. Each category contains one prototypical stimulus (000000 or 111111), five stimuli each having five features in common with the prototype, and one stimulus with only one feature in common with the prototype. Note that there is no linear function of the individual features that can correctly classify every stimulus.

In each experiment, the subjects were presented with a random permutation of the 14 stimuli and asked to identify each as belonging to either Category A or Category B, receiving feedback after each stimulus. This block of 14 stimuli was repeated 28 times for each subject, and the response data was aggregated into 7 segments of 4 blocks each. The averaged responses are presented in Figures 5 (a) and 7 (a) for Experiments 1:LS and 1:NLS, respectively.

Modeling procedure

In order to compare the DPMM to the prototype and exemplar models, all three were implemented in Matlab, exposed to the same training stimuli as the human subjects, and used to categorize each stimulus after each segment of 4 blocks. All three models were implemented with a cluster probability function that treats the dimensions (individual letters) as independent features of the stimuli, so

$$P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \prod_d P(x_{N,d}|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) \tag{5}$$

where $x_{N,d}$ is the value of the dth dimension of x_N. The individual dimensions are assumed to have Bernoulli probability distributions, where the parameter is integrated out with a Beta(β_0, β_1) prior to obtain

$$P(x_{N,d} = v|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \frac{M_{k,v} + \beta_v}{M_k + \beta_0 + \beta_1} \tag{6}$$

where v is either 0 or 1, and $M_{k,v}$ is the number of stimuli with value v on the dth dimension and belonging to cluster k according to $\mathbf{z}_N$.

The prototype and exemplar models are simple enough to allow direct implementation, but since the DPMM allows the stimuli of each category to be arbitrarily clustered, it becomes computationally infeasible to calculate its response probabilities with even modest numbers of stimuli. To alleviate this problem, we used the Markov chain Monte Carlo (MCMC) algorithm described in [20] and implemented by Y. Teh [19] to approximate the DPMM's true distribution over stimuli clusterings. For each DPMM data point, we ran the MCMC algorithm with a burn-in of 1000 steps, followed by 100 samples separated by 10 steps each. The α parameter of the Dirichlet process was sampled at each step of the MCMC algorithm, using a Gamma(1,1) prior distribution.

Once the probability of stimulus N belonging to category j is determined for each model, the response rule governing a subject's behavior is given by

$$P_{\text{resp}}(j|x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1}) = \frac{\Gamma}{|\{c_1, \ldots, c_{N-1}\}|} + (1-\Gamma)\frac{P(c_N = j|x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})^\gamma}{\sum_{j'} P(c_N = j'|x_N, \mathbf{x}_{N-1}, \mathbf{c}_{N-1})^\gamma} \tag{7}$$

where $|\{c_1, \ldots, c_{N-1}\}|$ is the number of categories under consideration, $0 \leq \Gamma \leq 1$ is a guessing-rate parameter, and $\gamma \geq 1$ specifies the degree to which the subject responds deterministically or probabilistically. Larger or smaller values of Γ make the response distribution more or less uniform, respectively. When $\gamma = 1$, the subject matches the probability of his responses to the probability of category membership. When $\gamma = \infty$, the subject always responds with the most probable category. This response-scaling parameter seems to be necessary to match human performance in different contexts. In particular, it seems that there are individual differences between γ values between different subjects in the same experiments [11]. Despite its apparent importance, it is missing from a number of prominent models, such as Anderson's RMC [2, 3]. The guessing-rate parameter Γ also seems to be helpful in fitting the non-optimality of human data for some experiments. Large values of Γ could possibly be explained by fatigue, misunderstanding, memory constraints, or just a failure to cooperate.

As in Smith and Minda's original modeling of this data, the guessing parameter Γ was incorporated in each model. The guessing parameter was allowed to vary between 0 and 1 across individual subjects, but was fixed per subject across every instance of every stimulus. Furthermore, the values of β_0 and β_1 in Equation (5) were fit to each subject, with the restriction that $\beta_0 = \beta_1$. Intuitively, this captures the variation in the subjects' tendencies to represent categories by either a few large clusters or many small clusters. The γ parameter in Equation (7) was left out, so the free parameters for each model are the guessing parameter Γ from Equation (7) and the value of $\beta_0 = \beta_1$, which were all fit individually per subject as to maximize the

total log likelihood of all the subjects' responses over all training segments.

Results

The response rates of the prototype, exemplar, and DPMM models are shown in Figures 5 (b), (c), and (d), respectively, for Experiment 1:LS. Figure 6 shows the log-likelihood of the human data (interpreted as independent Bernoulli trials) under each model across time. I was able to reproduce the early advantage for the prototype model in fitting the human data, but unlike Smith and Minda, I did not see the exemplar model beginning to take a lead in the later stages of learning. Instead, the prototype model explained the human data better throughout the experiment. It is not surprising that the complexity of the exemplar model is unnecessary to explain human performance, since the categories are perfectly described by a simple prototype representation. The DPMM performed almost identically to the prototype model in all segments.

The response rates for Experiment 1:NLS are shown in Figure 7, and the log-likelihood scores are presented in Figure 8. There is a very noticeable cross-over effect in this experiment, where the distractor stimuli start off in the wrong categories but are eventually learned to be more correctly classified. The prototype model clearly fails to display this effect, while the exemplar model immediately classifies the distractors correctly. Only the DPMM comes close to capturing this behavior. The explanation given by Smith and Minda is that subjects tend to use a more prototype-based model during the early stages of learning, switching to an exemplar-based model later on. In fact, this is exactly what the DPMM does: it assigns all the stimuli in a category to a single cluster at first, but with repeated exposure, the distractor stimuli split off into a separate cluster. Thus, the DPMM resembles the prototype model at first, and moves more towards the exemplar model as time progresses.

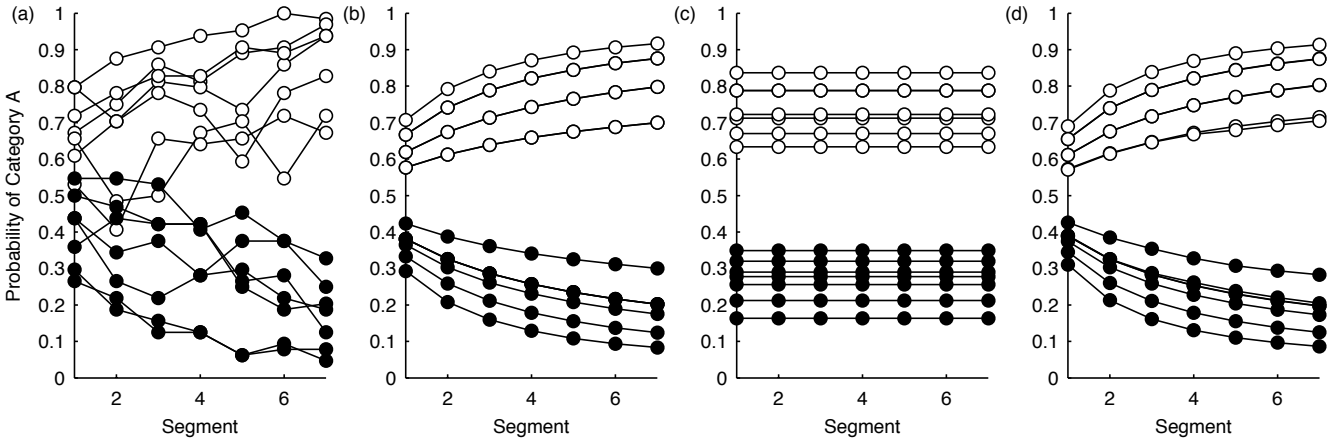

Figure 5: Human data and model predictions for Smith & Minda 1998, Experiment 1:LS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B.

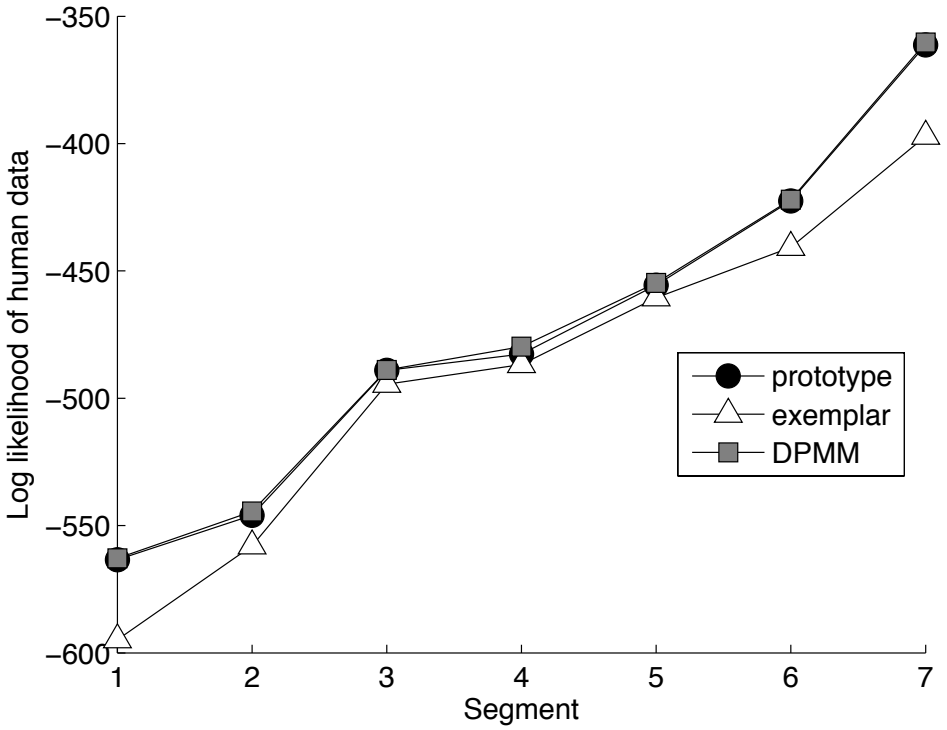

Figure 6: Log likelihood of human data for Smith & Minda 1998, Experiment 1:LS, with respect to each of the three models.

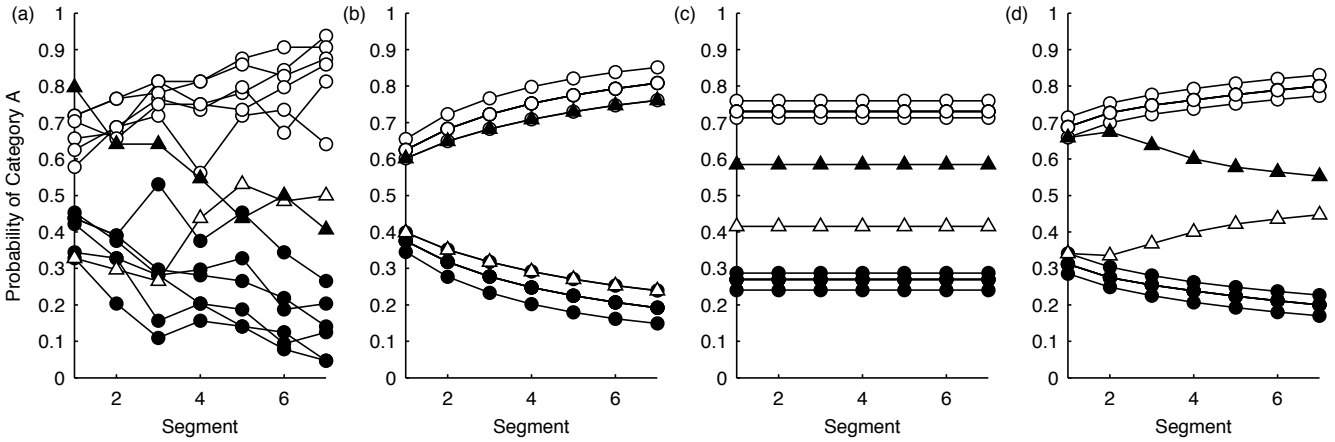

Figure 7: Human data and model predictions for Smith & Minda 1998, Experiment 1:NLS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure (111101 and 000100, respectively).

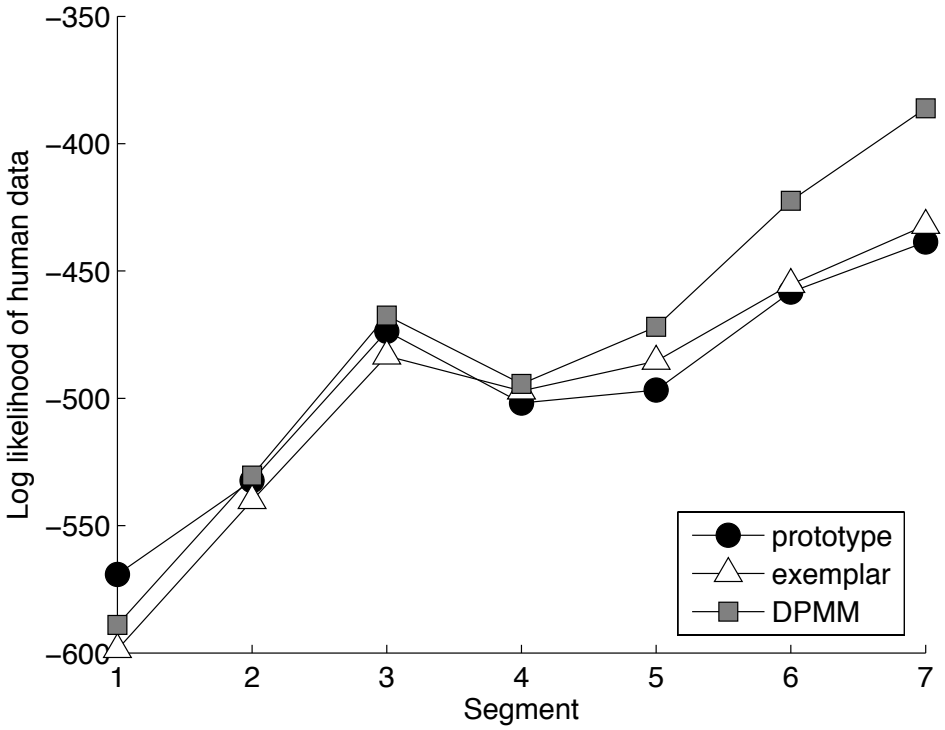

Figure 8: Log likelihood of human data for Smith & Minda 1998, Experiment 1:NLS, with respect to each of the three models.

4.1.2 Experiment 2

Smith and Minda decided to recreate Experiment 1, allowing subjects to continue learning for more trials. Their original analysis showed that the prototype model better explained human performance than the exemplar model until the very end of the experiment, and they were curious whether the exemplar model would significantly overtake the prototype model in later learning.

The data and procedure for Experiment 2 are identical to that of Experiment 1, with the exception that subjects were shown 40 blocks of the 14 stimuli rather than 28 blocks. These trials were aggregated into 10 segments of 4 blocks each. The averaged responses are shown in Figures 9 (a), and 11 (a) for Experiments 2:LS and 2:NLS, respectively.

Modeling procedure

The same modeling procedure was followed for Experiment 2 as for Experiment 1.

Results

The response rates of the prototype, exemplar, and DPMM models are shown in Figures 9 (b), (c), and (d), respectively, for Experiment 2:LS. Figure 10 shows the log-likelihood of the human data under each model across time. There are no surprises here beyond Experiment 1:LS. I was unable to reproduce the advantage in later stages of training for the exemplar model found by Smith and Minda; the prototype model maintains a steady lead throughout the experiment. Again, the DPMM explains the human data equally well as the prototype model.

The response rates for Experiment 2:NLS are shown in Figure 11, and the log-likelihood scores are presented in Figure 12. As in Experiment 1:NLS, there is a noticeable crossing-over behavior for the two distractor stimuli. Once again, the DPMM is the only model able to capture this effect, so it better fits the human data.

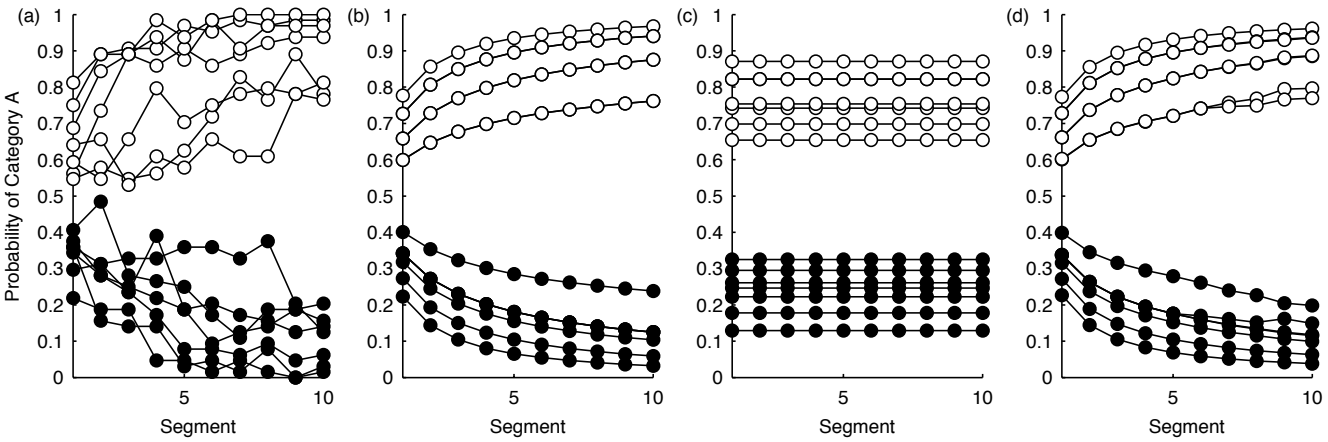

Figure 9: Human data and model predictions for Smith & Minda 1998, Experiment 2:LS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B.

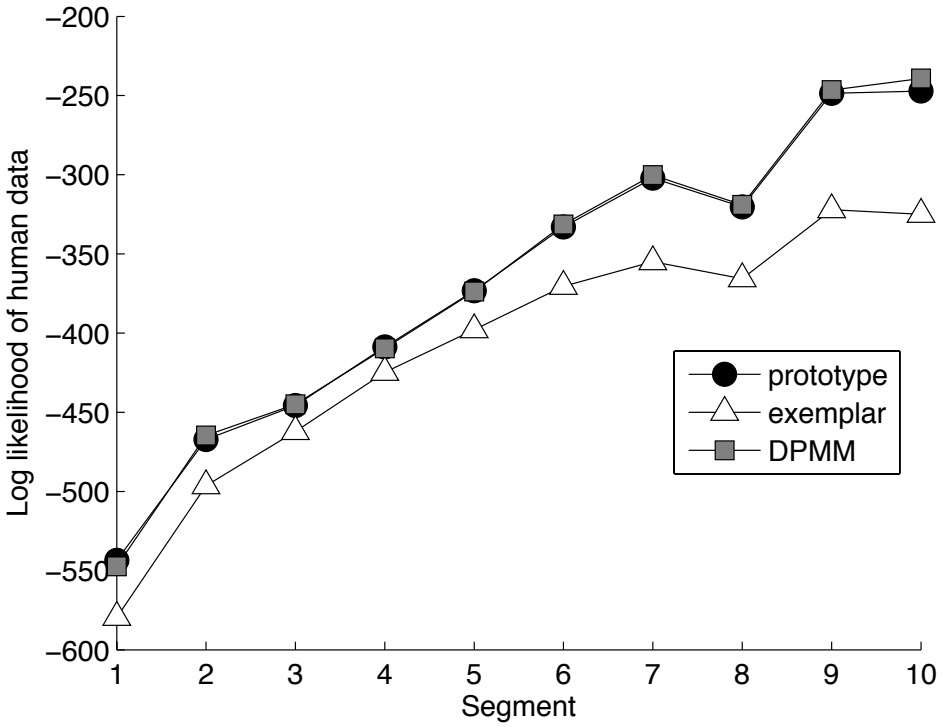

Figure 10: Log likelihood of human data for Smith & Minda 1998, Experiment 2:LS, with respect to each of the three models.

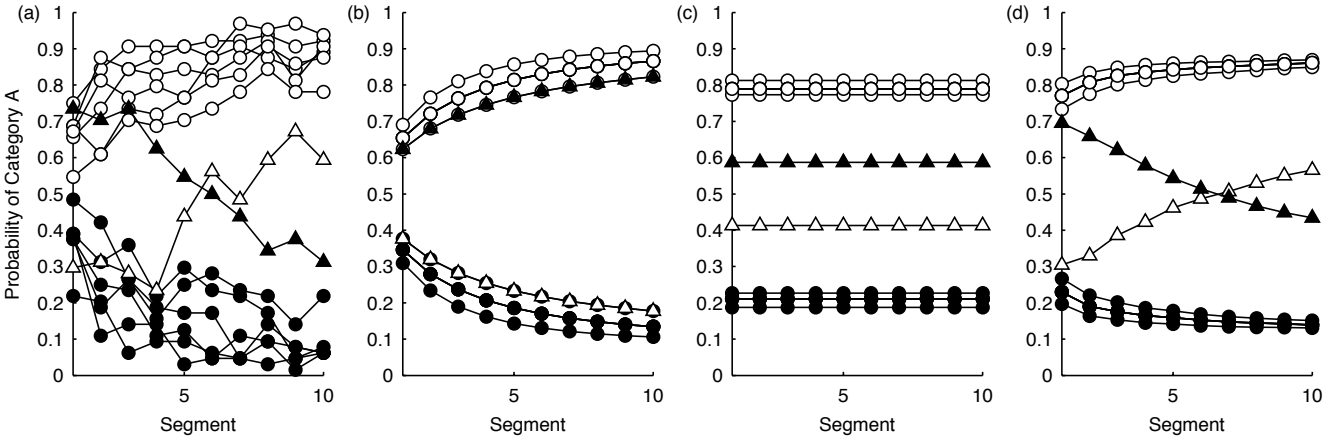

Figure 11: Human data and model predictions for Smith & Minda 1998, Experiment 2:NLS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure (111101 and 000100, respectively).

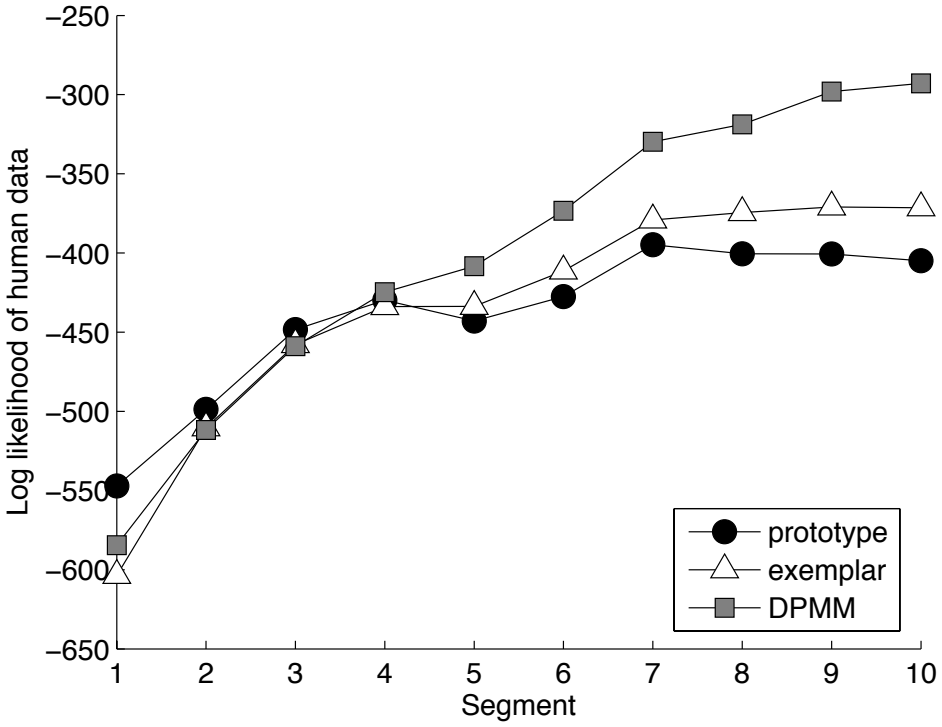

Figure 12: Log likelihood of human data for Smith & Minda 1998, Experiment 2:NLS, with respect to each of the three models.

Category A		Category B	
1010	kupo	1110	kypo
0110	bypo	1011	kupa
0001	buna	1101	kyna
1100	kyno	0111	bypa

Table 3: Categories A and B from Smith & Minda 1998, Experiment 3:LS

Category A		Category B	
0001	buna	1000	kuno
0100	byno	1010	kupo
1011	kupa	1111	kypa
0000	buno	0111	bypa

Table 4: Categories A and B from Smith & Minda 1998, Experiment 3:NLS

4.1.3 Experiment 3

The purpose of Smith and Minda's Experiment 3 was to determine if human performance in learning smaller, less-differentiated categories would be better explained by an exemplar model. They hypothesized that in this situation, exemplar-based strategies would emerge sooner and be more pronounced than in the previous experiments.

As before, subjects were presented with stimuli in the form of nonsense words. However, the words were only four letters long, and categories consisted of only four members each. Again, two different categories structures were used (identical to those used by Medin and Schwanenflugel [9] in their Experiment 2), one being linearly separable and the other being not linearly separable. The stimuli used for Experiment 3:LS are listed in Table 3. Here, category membership can be determined by counting the number of 1s in the stimulus (a linear function of the dimensional values).

The stimuli used in Experiment 3:NLS are listed in Table 4. Here, category membership cannot be determined by any linear combination of the individual dimensional values.

The procedure used in Experiment 3 is identical to that of Experiments 1 and 2, with subjects being exposed to 70 blocks of the 8 stimuli. The trials were aggregated into 10 segments of 7 blocks each, and the average responses are shown in Figures 13 (a) and 15 (a) for Experiments 3:LS and 3:NLS, respectively.

Modeling procedure

The same modeling procedure was followed for Experiment 3 as for Experiments 1 and 2.

Results

The response rates of the prototype, exemplar, and DPMM models are shown in Figures 13 (b), (c), and (d), respectively, for Experiment 3:LS. Figure 14 shows the log-likelihood of the human data under each model across time. As opposed to the findings of Smith and Minda, the prototype model dominates the exemplar model in explaining human responses throughout the experiment. Also, since the categories are less distinguished than in Experiments 1 and 2, the increased flexibility of the DPMM allows it to better capture the dynamics of human learning, so it has the strongest fit, especially in the later stages of learning.

The response rates for Experiment 3:NLS are shown in Figure 15, and the log-likelihood scores are presented in Figure 16. Here, the exemplar model does outperform the prototype model in explaining the human data from the first segment onward, as found by Smith and Minda. However, this advantage is shadowed by the even better fit provided by the DPMM. As in the previous NLS category structure, there seems to be a crossover effect (depicted by the triangular markers in Figure 15), which is captured very well by the DPMM.

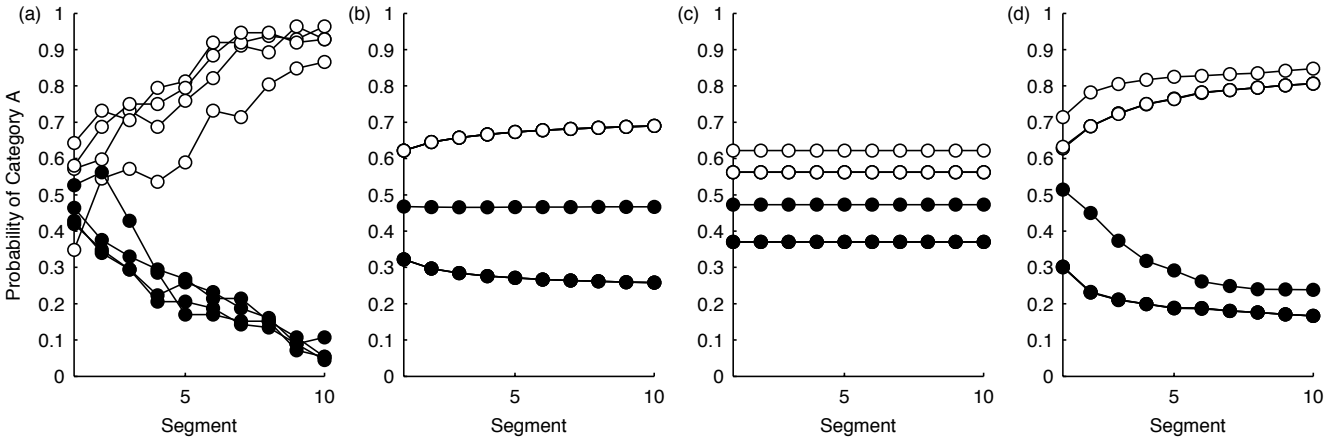

Figure 13: Human data and model predictions for Smith & Minda 1998, Experiment 3:LS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B.

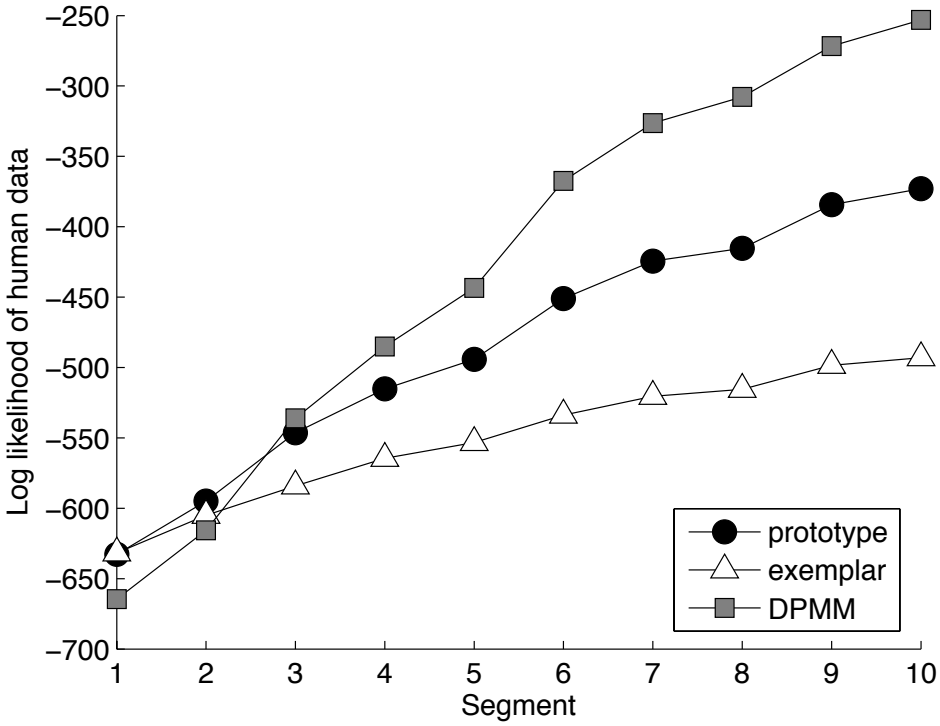

Figure 14: Log likelihood of human data for Smith & Minda 1998, Experiment 3:LS, with respect to each of the three models.

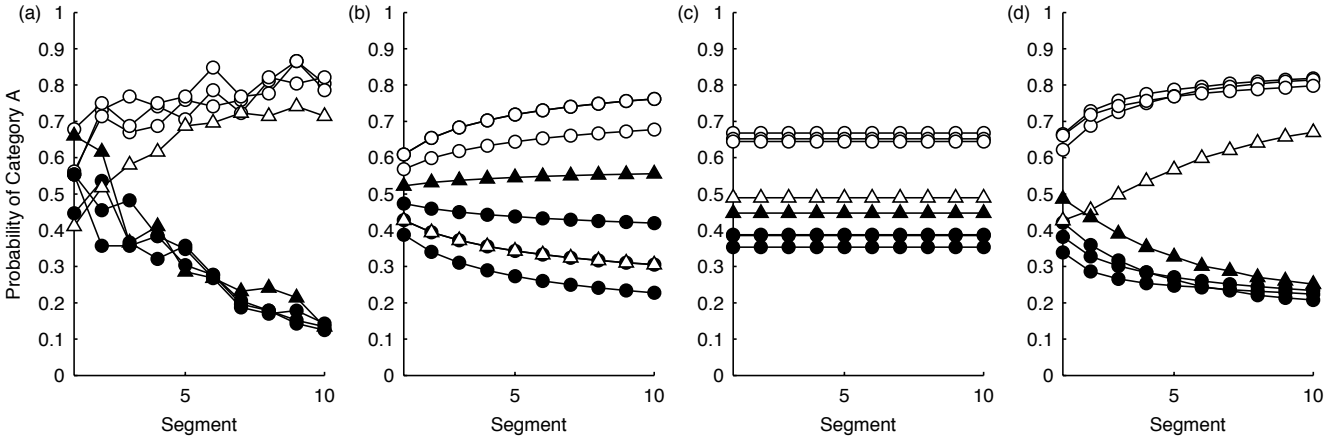

Figure 15: Human data and model predictions for Smith & Minda 1998, Experiment 3:NLS. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the distractor stimuli (1011 and 1000, respectively).

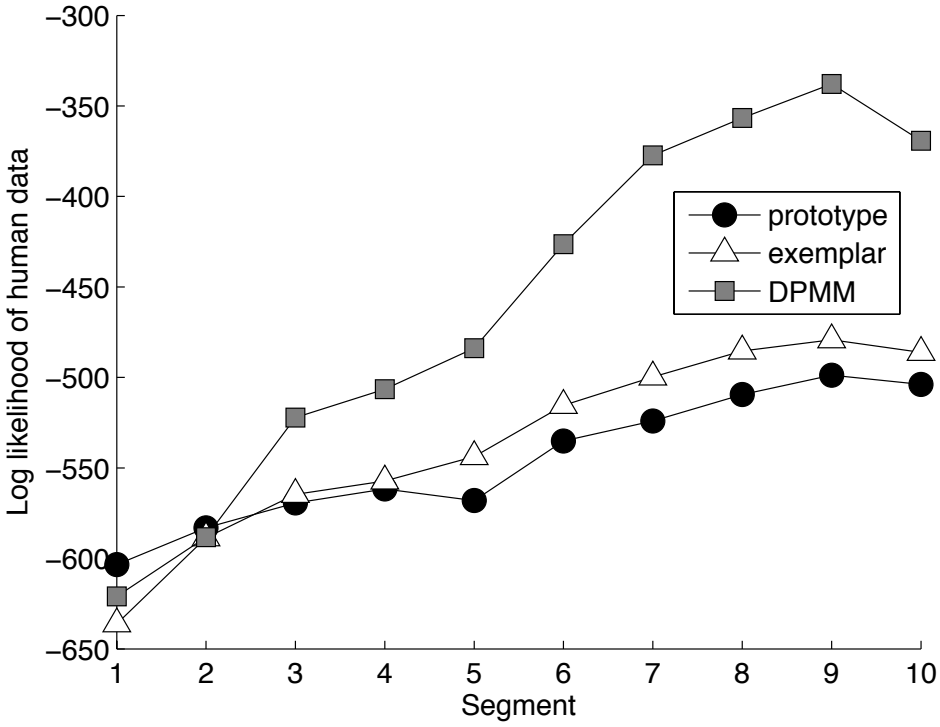

Figure 16: Log likelihood of human data for Smith & Minda 1998, Experiment 3:NLS, with respect to each of the three models.

4.1.4 Experiment 4

Smith and Minda decided to replicate some of their previous experiments using different stimuli. While Experiments 1, 2, and 3 exposed subjects to nonsense words, Experiment 4 instead used line drawings of bug-like creatures.

There were two sets of category structures: 4-dimensional not linearly separable (identical to those in Experiment 3:NLS), and 6-dimensional not linearly separable (identical to those in Experiments 1:NLS and 2:NLS). The graphical depiction of the stimuli is shown in Figures 17 and 18 for Experiment 4:4D, and in Figures 19 and 20 for Experiment 4:6D. Here, each binary-valued dimension of a stimulus corresponds to one of two values for a feature of the line drawing, e.g., eye type, body size, and antenna shape.

Subjects were exposed to 70 blocks of the 8 stimuli in Experiment 4:4D and 40 blocks of the 14 stimuli in Experiment 4:6D. The responses were aggregated into 10 segments of 7 blocks each for Experiment 4:4D and 10 segments of 4 blocks each for Experiment 4:6D. The average responses are shown in Figure 21 (a) and 23 (a) for Experiment 4:4D and Experiment 4:6D, respectively.

Modeling procedure

The same modeling procedure was followed for Experiment 4 as for Experiments 1-3.

Results

The response rates of the prototype, exemplar, and DPMM models are shown in Figures 21 (b), (c), and (d), respectively, for Experiment 4:4D. Figure 22 shows the log-likelihood of the human data under each model across time. In this experiment, Smith and Minda found a significant advantage for the exemplar model throughout all stages of learning. My results partially recreate this, showing a slight advantage for the exemplar model through most stages of learning. The DPMM fit the human

Figure 17: The Category A stimuli for Experiment 4:4D.

Figure 18: The Category B stimuli for Experiment 4:4D.

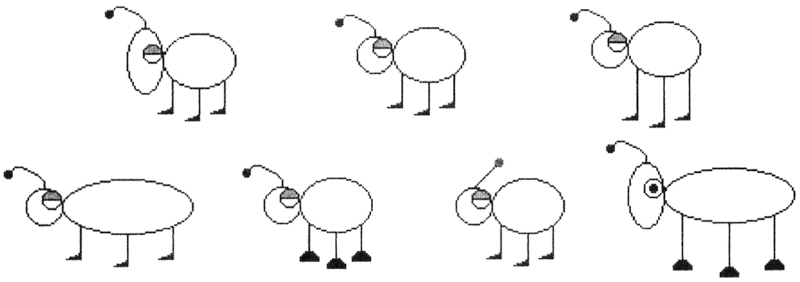

Figure 19: The Category A stimuli for Experiment 4:6D.

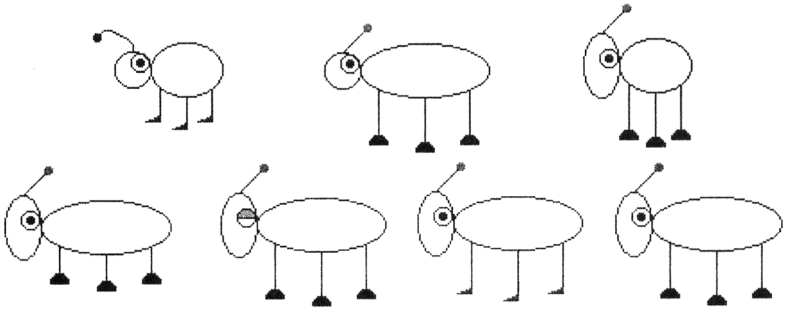

Figure 20: The Category B stimuli for Experiment 4:6D.

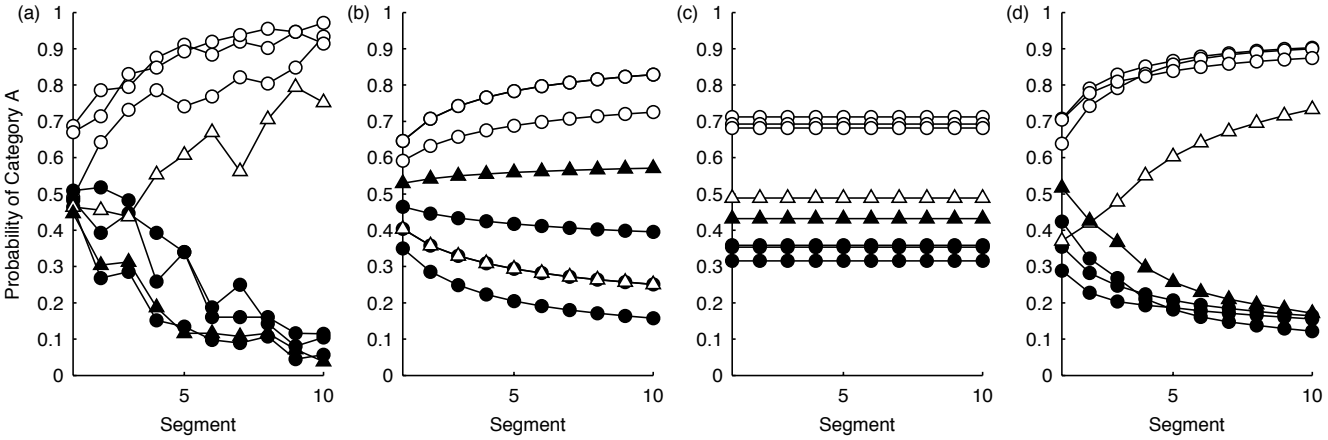

Figure 21: Human data and model predictions for Smith & Minda 1998, Experiment 4:4D. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the distractor stimuli (1011 and 1000, respectively).

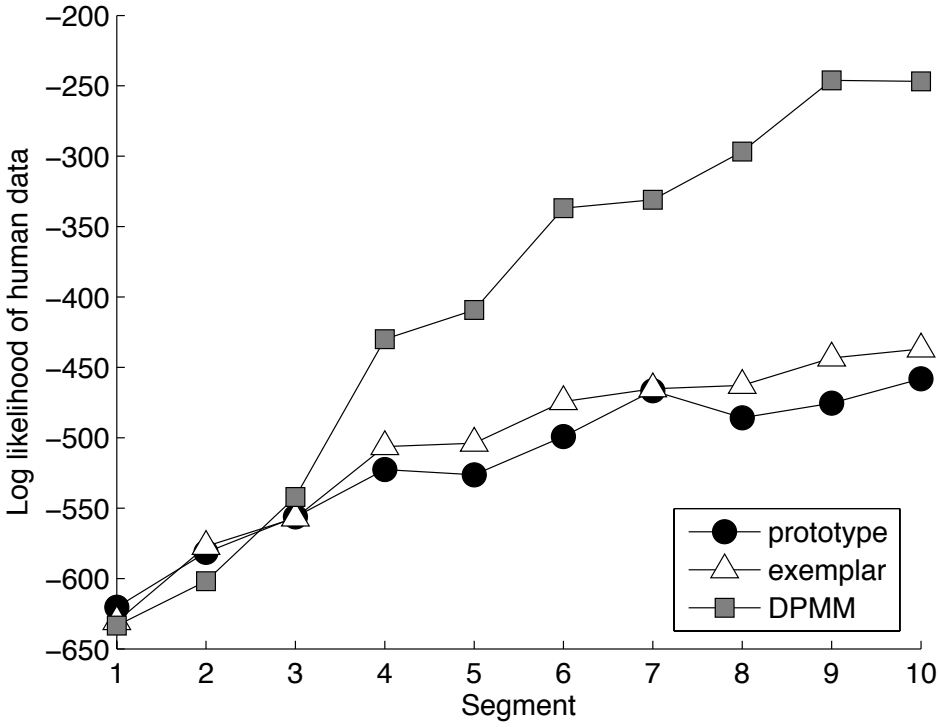

Figure 22: Log likelihood of human data for Smith & Minda 1998, Experiment 4:4D, with respect to each of the three models.

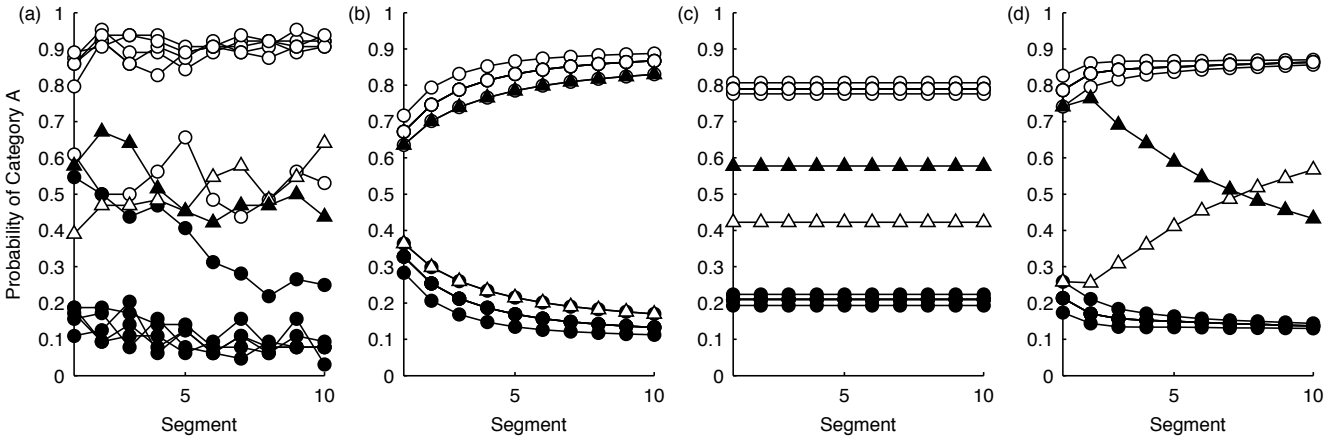

Figure 23: Human data and model predictions for Smith & Minda 1998, Experiment 4:6D. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure (111101 and 000100, respectively).

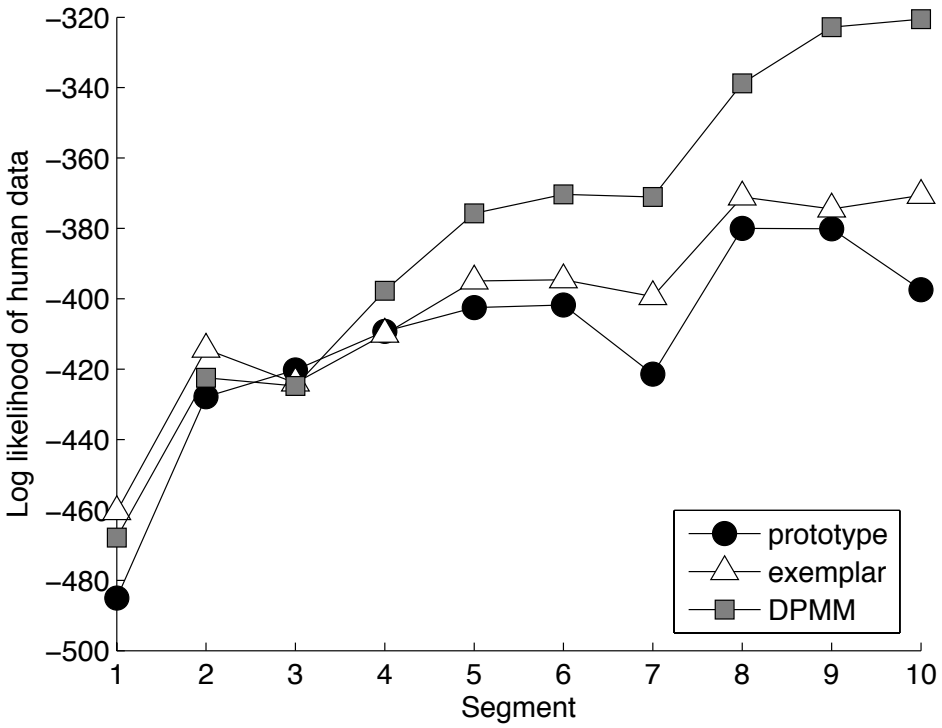

Figure 24: Log likelihood of human data for Smith & Minda 1998, Experiment 4:6D, with respect to each of the three models.

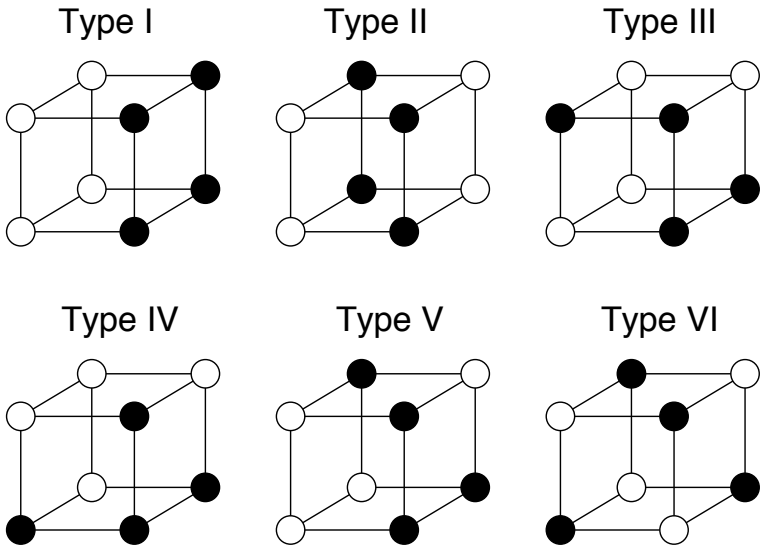

Figure 25: The six types of category structures used in Nosofsky et al. 1994.

data significantly better overall, however. As in Experiments 1-3, this is presumably due to the crossover effect of the Category A distractor stimulus.

The response rates for Experiment 4:6D are shown in Figure 23, and the log-likelihood scores are presented in Figure 24. The comparison of the prototype and exemplar models' performance in this experiment is comparable to the findings of Smith and Minda. The DPMM again explains human performance significantly better overall.

4.2 Nosofsky et al. 1994

The Nosofsky et al. 1994 experiment is a replication and extension of Shepard, Hovland, and Jenkins (1961). The goal of the experiment was to determine the relative performance of three existing categorization models (ALCOVE [7], RMC [2, 3], and the configural-cue model [6]) on the well-known task of learning category structures defined on stimuli with three binary-valued features. Modulo reflection, rotation, and inversion, it is possible to define six different 2-category structures, shown in Figure 25. It has been shown previously [16] and confirmed by Nosofsky et

al. that people are able to learn categories of Type I most easily, followed by Type II, then Types III, IV, and V, with Type VI structures being the most difficult to learn. The key result of this experiment is that models excel at explaining human performance when they include a way for certain dimensions of the stimuli to receive preference when calculating psychological distances.

Although the basic cluster density function given by Equation (5) doesn't include weighting coefficients for the different dimensions, it allows stimuli to be clustered together which share many common features. So we would expect that the DPMM should tend to create a separate cluster for each contiguous group of stimuli and more quickly learn category structures with fewer clusters. This intuition is in congruence with the difficulty displayed by human learners.

For each of the six category structure types, the subjects were presented with a series of stimuli in the form of a simple drawing that assumed three binary-valued features: shape, color, and size. Each stimulus was either a square or a triangle, black or white, and large or small. Each of the six types of category structures in Figure 25 were tested. The subjects were presented with a random permutation of the 8 stimuli and asked to identify each as belonging to either Category A or Category B, receiving feedback after each stimulus. This block of 8 stimuli was repeated 50 times for each subject, and the average training error for each category structure type and segment of 2 blocks was recorded. The authors found that most training errors had dropped to zero after 16 segments of 2 blocks, so only these segments were used to compare model fits. The average training errors are presented in Figure 26 (a).

Modeling procedure

The three models were exposed to the same data as the human subjects and used to categorize each stimulus after each segment of 2 blocks. The cluster probability distributions were identical to those used in the Smith and Minda experiments (see

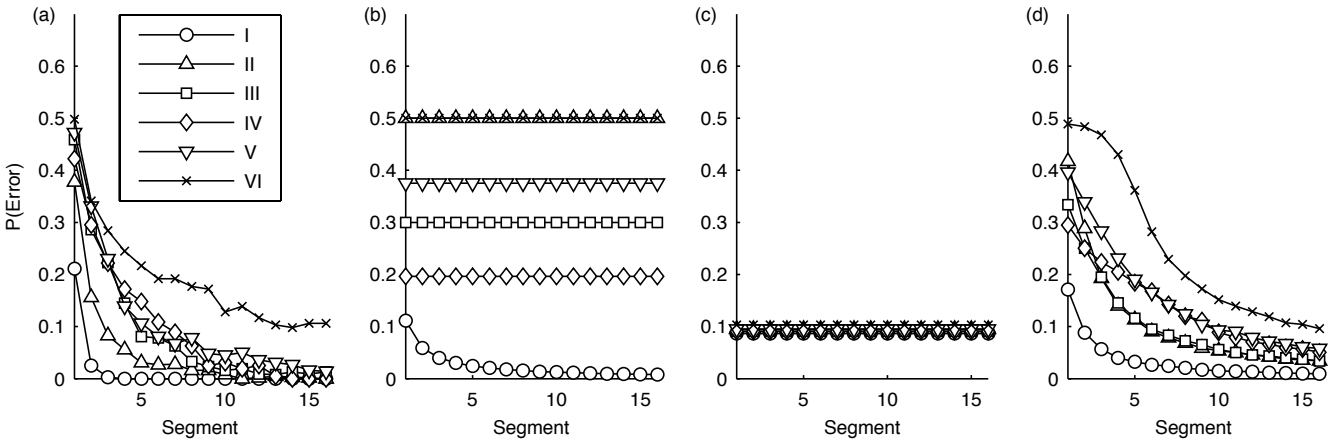

Figure 26: Average number of errors per segment of human data and model predictions for Nosofsky et al. 1994. (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM.

Equations (5) and (6)). Again, a guessing-rate parameter Γ was used, but not a response-scaling parameter γ (see Equation (7)).

Rather than fitting the parameters $\beta_0 = \beta_1$ and Γ to each subject individually, the procedure used by Nosofsky et al. in [11] was followed, where the parameters of each model were fixed across all subjects and category types.

Results

The response rates of the prototype, exemplar, and DPMM models are shown in Figures 26 (b), (c), and (d), respectively. Table 5 shows the total sum-squared-error between the human error rates and the model error rates for all category structure types.

This experiment highlights the main weakness of a prototype-based model: in type II and type VI category structures, the two categories have identical prototypes, and so the model is unable to do any better than random guessing in these situations. Only type I and type IV category structures are sufficiently differentiated for the prototype model to perform well. The exemplar model, on the other hand, can perform as objectively well as necessary. Unfortunately, it is unable to learn from

Model	SSE
Prototype	7.721
Exemplar	1.328
DPMM	0.347

Table 5: The sum-squared-error (SSE) of the best-fitting model of each type. SSE is computed across all six category structure types and all 16 training segments.

repeated exposure and is constrained to a flat error curve. The DPMM interpolates between a prototype-style representation and an exemplar-style representation and explains human performance much better than the other two models.

Nosofsky et al. report SSE values below 0.25 for all the models they implemented, with the RMC achieving 0.182 in particular. The SSE value of the DPMM comes impressively close to this, considering it has only 2 free parameters, while the RMC, as implemented by Nosofsky et al., has 4.

5 Conclusion

There is a long history of various algorithms attempting to model the dynamics of human categorization. Most can be described as some adaptation of the basic exemplar and prototype models. Since these two models have unique strengths and weaknesses and can be interpreted as opposite ends of a spectrum, much attention has been given to finding new models that interpolate between them. In particular, the Varying Abstraction Model [21] and Mixture Model of Categorization [14] allow categories to be represented as a combination of discrete clusters. The Rational Model of Categorization (RMC) [2, 3] provides an efficient algorithm for automatically determining cluster memberships, but it suffers from a number of problems. With Neal's realization that the RMC's underlying model is equivalent to that of the Dirichlet process mixture model (DPMM), we are able to implement an algorithm for sampling from this model that is both efficient and asymptotically optimal. The DPMM's ability to automatically interpolate between prototype and exemplar-style models as the data warrants is the key feature that allows it to explain human performance so well.

References

[1] D. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin, 1985.

[2] John R. Anderson. *The adaptive character of thought*. Erlbaum, Hillsdale, NJ, 1990.

[3] John R. Anderson. The adaptive nature of human categorization. *Psycholgical Review*, 98(3):409–429, 1991.

[4] F. Gregory Ashby and Leola A. Alfonso-Reese. Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:216–233, 1995.

[5] F. Gregory Ashby and Ralph E. Gott. Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1):33–53, January 1988.

[6] Mark A. Gluck and Gordon H. Bower. Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, 27(2):166–195, April 1988.

[7] John K. Kruschke. Alcove: An exemplar-based connectionist model of category learning. *Psycholgical Review*, 99(1):22–44, January 1992.

[8] Douglas L. Medin and Marguerite M. Schaffer. Context theory of classification learning. *Psychological Review*, 85(3):207–238, 1978.

[9] Douglas L. Medin and Paula J. Schwanenflugel. Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 7(5):355–368, 1981.

[10] Radford M. Neal. Markov chain sampling methods for dirichlet proces mixture models. Technical Report 9815, Department of Statistics, University of Toronto, September 1998.

[11] Robert M. Nosofsky, Mark A. Gluck, Thomas J. Palmeri, Stephen C. McKinley, and Paul Glauthier. Comparing models of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory and Cognition*, 22(3):352–369, 1994.

[12] Robert M. Nosofsky, Thomas J. Palmeri, and Stephen C. McKinley. Rule-plus-exception model of classification learning. *Psycholgical Review*, 101(1):53–79, 1994.

[13] M. I. Posner and S. W. Keele. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77:353–363, 1968.

[14] Yves Rosseel. Mixture models of categorization. *Journal of Mathematical Psychology*, 46:178–210, 2002.

[15] Adam N. Sanborn, Thomas L. Griffiths, and Daniel J. Navarro. A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006.

[16] R. N. Shepard, C. I. Hovland, and H. M. Jenkins. Learning and memorization of classifications. *Psychological Monographs*, 75, 1961. 13, Whole No. 517.

[17] Roger N. Shepard. Towards a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, September 1987.

[18] J. David Smith and John Paul Minda. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(6):1411–1436, 1998.

[19] Yee Whye Teh. Nonparametric bayesian mixture models - release 1. `http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html`.

[20] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2004.

[21] W. Vanpaemel, G. Storms, and B. Ons. A varying abstraction model for categorization. In B. Bara, L. Barsalou, and M. Bucciarelli, editors, *Proceedings of the 27th annual conference of the Cognitive Science Society*, pages 2277–2282, Mahwah, NJ, 2005. Lawrence Erlbaum.