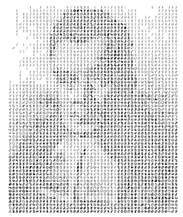
The Hierarchical Dirichlet Process as a Model of Human Category Learning

Kevin R. Canini
kevin@cs.berkeley.edu

Thomas L. Griffiths
tom_griffiths@berkeley.edu

University of California, Berkeley



Introduction

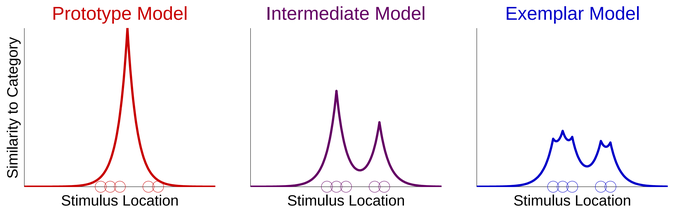
Category learning is the inference of category structures from a set of stimuli each labeled as belonging to one of the categories. Human category learning is of central importance to many ideas in psychology. One of the most basic traits of human cognition is the ability to group objects in the world into cohesive categories. Despite the apparent simplicity of this problem, people display quite complex behaviors in categorization settings. Their high degree of insightfulness and intelligence has been verified both anecdotally and through laboratory experiments. Some features of human-level categorization are: selectively attending to salient features, generalizing categories to novel objects, forming new categories to explain surprising data, generating hierarchical category structures, estimating confidence by deciding how loosely to generalize, and transferring knowledge between tasks about what types of objects tend to be categorized together.

Classical models of human category learning have not been able to explain all of these phenomena. The hierarchical Dirichlet process (HDP) is a framework that can be used to specify rich models of human categorization, both subsuming many existing models and containing promising new ones as well.

Representing Categories

Psychological models of categorization fall into three general categories:

- Prototype models (Reed, 1972) represent a category by a single object, the *prototype*. The strength of a new stimulus' membership in the category is measured by its similarity to the prototype.
- Exemplar models (Medin & Schaffer, 1978) represent a category by memorizing every instance of it, the *exemplars*. The strength of a new stimulus' membership in the category is measured by its average similarity to the exemplars.
- Intermediate models (Anderson, 1990; Rosseel 2002; Vanpaemel et al., 2005) represent a category by clustering its instances and computing the strength of a new stimulus' membership in the category by its average similarity to the cluster centers. This is equivalent to a prototype model when only a single cluster is used, and to an exemplar model when every object is in its own cluster. The most interesting cases are between these extremes.



In theory, intermediate models must consider all partitions of a category's instances. In previous work, this limitation has been skirted by using suboptimal, greedy clustering algorithms and/or assuming that the number of clusters is fixed ahead of time. Due to efficient sampling algorithms such as Markov chain Monte Carlo (MCMC), more sophisticated models based on Bayesian inference, such as the HDP, can now be tested.

Dirichlet Process Mixture Models

Dirichlet process mixtures models (DPMMs) (Antoniak, 1974) probabilistically partition the objects in a category into clusters according to the prior probability distribution

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

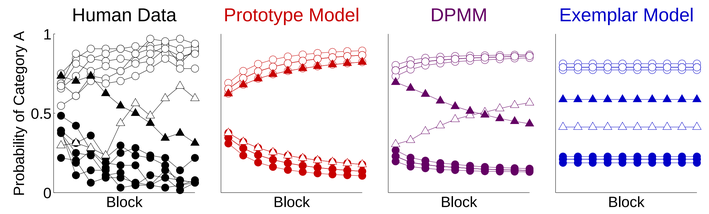
where N is the total number of objects, K is the number of clusters, M_k is the size of cluster k, and α is a parameter that governs how much the model favors a few large clusters or many small clusters. The DPMM has the flexibility to move between prototype- and exemplar-style category representations as warranted by the data. The DPMM is equivalent to the Rational Model of Categorization introduced by Anderson (1990). We use a Gibbs sampling algorithm to cluster a category's objects rather than the greedy algorithm used by Anderson.

The Prototype-Exemplar Transition

Smith and Minda (1998) showed that neither prototype nor exemplar models are strictly better at modeling human categorization, with people transitioning between these two styles during a single experiment. The following categories were learned by human subjects (note the distractor stimuli at the ends of the lists):

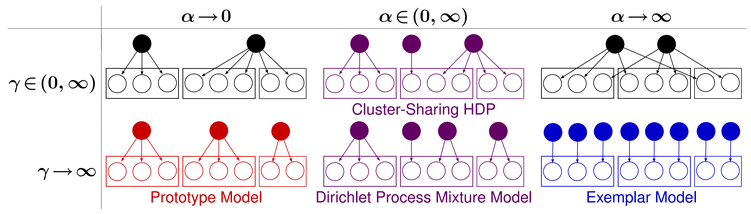
	Stimuli from Smith & Minda, 1998, Experiment 2-NLS
A	000000, 100000, 010000, 001000, 000010, 000001, 111101
B	111111, 011111, 101111, 110111, 111011, 111110, 000100

The transition from prototype-style learning to exemplar-style learning is well-replicated by modeling each category with an independent DPMM (Griffiths, Canini, et al., 2007). The DPMM is the only model of the three that explains the crossover of the distractor stimuli.



Hierarchical Dirichlet Processes

Hierarchical Dirichlet processes (HDPs; Teh, Jordan, et al., 2004) allow categories to pool their objects together into shared clusters, thereby sharing statistical strength. Through various settings of the HDP's hyperparameters α (which controls how large clusters within a category tend to be) and γ (which controls how likely categories are to share clusters), a number of models can be derived, including the prototype, exemplar, and DPMM models (Griffiths, Sanborn, et al. 2007).



Most interestingly, the "cluster-sharing" HDP model promises to exhibit previously unexplained traits of human learning.

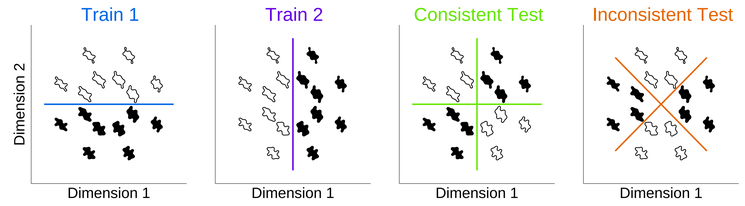
Transfer Learning

During the course of category learning, people inevitably gain information about what types of objects tend to be categorized together. This knowledge influences future episodes of category learning by creating assumptions about the way objects will be grouped together.

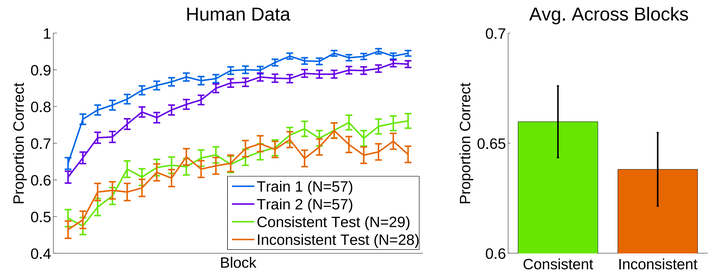
While learning the category of striped objects, the striped cats form a natural cluster. This induces a prior belief that the striped cats will tend to be members or non-members of other categories together, thereby reducing the time it takes to learn the category of all cats.

Experiments

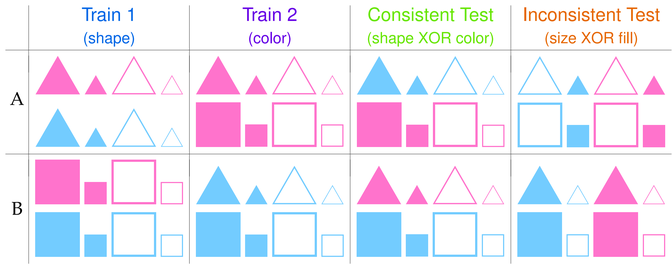
We are conducting experiments to quantify the transfer learning effect described above. In both experiments, subjects learn to categorize 16 stimuli in three different ways. The last task is either *consistent* or *inconsistent* with the first two, as depicted below.



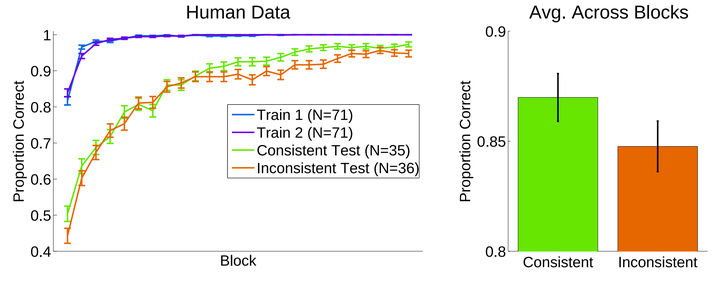
We counterbalance the dimensions used for Train 1 and Train 2, so that the same Test condition can be either consistent or inconsistent. Preliminary results are shown below.



Interestingly, the interference between the Train 1 and Train 2 sessions seems greater than between the Training and Test sessions. Since many subjects became fatigued part-way through Experiment 1, Experiment 2 is designed to be easier to complete.



Preliminary results are shown below.



Acknowledgements

This work was supported by grant number FA9550-07-1-0351 from the Air Force Office of Scientific Research.