A Nonparametric Bayesian Model of Multi-Level Category Learning

Kevin R. Canini
Computer Science Division
University of California
Berkeley, California 94720 USA
kevin@cs.berkeley.edu

Thomas L. Griffiths
Department of Psychology
University of California
Berkeley, California 94720 USA
tom_griffiths@berkeley.edu

Abstract

Categories are often organized into hierarchical taxonomies, that is, tree structures where each node represents a labeled category, and a node's parent and children are, respectively, the category's supertype and subtypes. A natural question is whether it is possible to reconstruct category taxonomies in cases where we are not given explicit information about how categories are related to each other, but only a sample of observations of the members of each category. In this paper, we introduce a nonparametric Bayesian model of multi-level category learning, an extension of the hierarchical Dirichlet process (HDP) that we call the tree-HDP. We demonstrate the ability of the tree-HDP to reconstruct simulated datasets of artificial taxonomies, and show that it produces similar performance to human learners on a taxonomy inference task.

Introduction

Taxonomic structures are a ubiquitous part of the way that people think about the world, appearing in biological phylogenies (Atran 1998), linguistic ontologies (Keil 1979), and many natural kind concepts (Rosch et al. 1976). In a taxonomy, categories are organized into different hierarchical levels, with the higher-level categories representing broader, more inclusive groups of entities, and the lower-level concepts representing narrower, more specific groups. When a category is a direct descendent of another category in a taxonomy, most or all the members of the first are also members of the second. For example, in our intuitive phylogenetic tree for animals, all dogs are mammals, and all border collies are dogs. This taxonomic structure supports efficient inferences about the properties of entities belonging to these categories (Collins and Quillian 1969).

The ubiquity of taxonomies raises a natural question: How can such structures be learned? While we might get some explicit information about the taxonomic relationships between categories, neither children nor artificial systems can rely on such information. Consequently, in this paper we focus on the question of how taxonomies might be learned just from labeled examples of category members. Consider the problem faced by a learner who sees a collection of objects given the labels "animal", "mammal", "dog", and "border collie". The challenge is to induce an appropriate representation for the categories associated with each of these labels, supporting future generalizations, and to determine how these categories are related to one another. For example, our learner would need to identify categories corresponding to "dog" and "border collie", and learn that "border collie" is a kind of "dog". Since the objects can each be associated with multiple labels, and the categories are defined at different levels of abstraction, we refer to this problem as *multi-level category learning*.

The complex relationships between categories make the problem of multi-level category learning quite different from the standard treatment of category learning (or multi-class classification) in cognitive science and machine learning. Most methods for learning categories do not allow complex relationships to exist between those categories. Typically, either categories are treated as independent (for example, by learning conditional distributions over the observed features of the objects separately for each category) or algorithms consider only basic interactions between categories (for example, discriminative methods attempt to discover the boundaries between categories). Multi-level category learning is also different from unsupervised methods for inducing hierarchies, such as hierarchical clustering (Duda, Hart, and Stork 2000; Heller and Ghahramani 2005), structure learning (Kemp and Tenenbaum 2009), learning ontologies (Kemp et al. 2006), or learning hierarchies (Roy et al. 2007; Blei, Griffiths, and Jordan 2010). These unsupervised methods find a way to organize a set of objects into a hierarchical structure, but do so on the basis of the similarity of the objects, rather than using the category labels of those objects.

In this paper, we investigate multi-level category learning in both artificial and natural systems. First, we propose a novel method of learning and representing categories which are organized in taxonomic systems. Our model is a nonparametric Bayesian statistical model which we call the *tree-HDP*. We demonstrate that this model can recover simple taxonomies from just labeled examples of category members. We then turn to natural systems, conducting an experiment studying the performance of human learners in a similar task. A comparison of the model with the experiment results shows that the tree-HDP is able to do just as well—or better than—human learners.

The remainder of the paper is organized as follows. We first present relevant background information, describing the hierarchical Dirichlet process (HDP) and some previous results of using this model to explore aspects of human cat-

egory learning. Next, we turn to the multi-level category learning problem and discuss how ideas from previous work can be extended to solve it. We define the tree-HDP, a nonparametric Bayesian model designed to perform multi-level category learning, and lay out an algorithm for performing statistical inference with the model. Finally, we describe the experiment we conducted with human learners and compare their performance to that of the tree-HDP.

Background

The hierarchical Dirichlet process (HDP) (Teh et al. 2006), on which the tree-HDP is based, was originally introduced in machine learning. More recently, this model has been used to characterize human category learning (Griffiths et al. 2007). In this section we introduce the HDP and summarize its application to category learning.

The Hierarchical Dirichlet Process

A basic problem that arises in clustering a set of observations is determining the number of clusters to use. Recent work in Bayesian statistics has addressed this problem by defining the probability distribution over observations to be a nonparametric mixture model (Neal 1998). These models are based on the Dirichlet process, a stochastic process that defines a distribution on discrete measures. We can describe a discrete measure G in terms of a (possibly infinite) set of atoms θ_k and weights β_k, with $G = \sum_k \beta_k \theta_k$. In a Dirichlet process mixture model (DPMM), the atoms are the parameters of each mixture component and the weights indicate the probability of observations being generated from that component. The Dirichlet process defines a distribution on θ and β, with each $\theta_i \sim H$ being drawn from a *base measure* H and $\boldsymbol{\beta} \sim \text{Stick}(\alpha)$ being drawn from the stick-breaking process (an infinite version of the Dirichlet distribution) with *concentration parameter* α. As α gets larger, it becomes more likely that more mixture components will be used. The DPMM induces a posterior distribution on the number of mixture components used to explain a set of observations.

Some intuitions for the behavior of the Dirichlet process can be obtained by considering the distribution that it induces on partitions of observations. If we integrate out the weights $\boldsymbol{\beta}$ and just consider the discrete variables z_i indicating which component an observation is drawn from, we obtain a simple and intuitive stochastic process. The probability distribution over the component from which the next observation will be drawn is given by

$$P(z_i = k) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{if } k = K + 1, \end{cases} \quad (1)$$

where K is the number of occupied clusters and n_k is the number of observations assigned to cluster k. This stochastic process is known as the *Chinese restaurant process* (CRP), based on a metaphor in which clusters correspond to tables in a restaurant, and observations are customers who take seats at different tables (Aldous 1985). The CRP representation is useful when performing inference in the DPMM, playing a key role in Markov chain Monte Carlo (MCMC) algorithms that are used to sample from the posterior distribution over partitions of the observations (Neal 1998).

The hierarchical Dirichlet process (HDP) extends this approach to allow the same clusters to be used in multiple probability distributions. The basic idea is to model each probability distribution as a DPMM, but to assume that those Dirichlet processes share another Dirichlet process as their base measure. This guarantees that some of the parameters for the mixture components will be shared across distributions, making it possible to capture shared statistical structure. The concentration parameter of the higher-level, shared Dirichlet process determines the extent to which components are likely to be shared across distributions. The HDP has been used to develop nonparametric models that go beyond simple clustering, making it possible to infer the number of topics in a topic model or the number of states in a hidden Markov model (Teh et al. 2006).

Modeling Human Category Learning

A substantial literature in cognitive psychology has explored models of human categorization. Influential models include prototype models, which represent a category by a single central member (Reed 1972), and exemplar models, which store in memory all objects and labels in order to predict the category membership of new objects (Medin and Schaffer 1978; Nosofsky 1986). More recently, researchers have begun to explore representations that are intermediate between these extremes, where each category is represented by a few clusters of objects rather than all exemplars (Love, Medin, and Gureckis 2004; Vanpaemel and Storms 2008). These models have been evaluated through experiments in which people learn a small number of independent categories from labeled examples and then form generalizations about the category membership of new objects.

The literature on human category learning can be linked with the work in machine learning summarized above by observing that many existing models can be expressed in terms of probability density estimation (Ashby and Alfonso-Reese 1995). Prototype models can be shown to be equivalent to parametric density estimation methods, exemplar models are closely related to kernel density estimation, and intermediate representations are similar to mixture models (Rosseel 2002). This connection means that nonparametric models based on the Dirichlet process may also be relevant to aspects of human category learning. Indeed, a model equivalent to the DPMM was suggested by Anderson (1991), and has recently been explored in more detail (Sanborn, Griffiths, and Navarro 2006; Zhu et al. 2010).

The HDP also has interesting links to human category learning, having been suggested as a general framework that subsumes these different approaches (Griffiths et al. 2007). In this framework, each category is represented by a DPMM, and the common base measure provides a way to share mixture components across categories. This approach provides a way to extend existing models to incorporate more complex relationships between categories. Sharing components between categories, rather than treating categories as independent, supports a form of transfer learning in which people can generalize knowledge formed by learning one category to another category. Recent work suggests that a similar kind of transfer learning can be seen when people learn

a set of related categories (Canini, Shashkov, and Griffiths 2010). This ability to share components across distributions is the key to being able to solve the problem of multi-level category learning, which is our focus in the rest of the paper.

Multi-Level Categories and the Tree-HDP

The multi-level category learning problem reduces to being provided with a set of observations drawn from different categories, together with the information that these categories form a taxonomy, and estimating the probability distribution associated with each category and the structure of the taxonomy. In this section, we introduce the tree-HDP model, describe an efficient inference algorithm, and demonstrate that it can be used to solve this problem.

The Tree-HDP Model

The tree-HDP is a generalization of the HDP, described above. In the typical formulation of the HDP, the latent structure G_j of each category j is a draw from a Dirichlet process (DP) with a base measure G_0 that is shared between all the categories. In turn, G_0 is a draw from a higher-level DP with base measure H, a hyperparameter chosen by the modeler. Although the HDP is typically used to model collections of categories arranged in a flat hierarchy, the same statistical definitions can be recursively applied to multiple levels of hierarchy. That is, instead of all the categories inheriting from G_0, some of them can inherit from others.

In practice, each draw from a DP yields a refinement or specialization of its base measure. In the flat HDP, this means that each category is a specialization of the global base measure H, which is typically chosen to give broad, flat coverage over a wide range of parameters. Although the categories will exhibit random fluctuations in their degree of specialization, the use of a flat hierarchy means that each one is, a priori, at the same level of refinement. By contrast, if we push some of the categories down into deeper levels of the tree, they become specializations of their respective parent categories. This is the mechanism that the tree-HDP uses to model taxonomy systems. Intuitively, the tree structure of the HDP is intended to mirror the true hierarchical relationships between the categories.

Formally, we relax the assumption that the random measures G_j are drawn from a Dirichlet process with a common base measure G_0. Instead, we allow the categories to form any valid tree structure with G_0 as the root. We introduce a new set of random variables $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_J\}$ to describe the tree structure, with τ_j denoting the index of category j's parent. If G_j is a child of G_0, then $\tau_j = 0$, and if it is a child of some other category $G_{j'}$, then $\tau_j = j'$. We restrict $\boldsymbol{\tau}$ to form a valid tree structure, i.e., cycles are not allowed.

To specify the full Bayesian probability model of the tree-HDP, it is necessary to choose a prior distribution for the random variables $\boldsymbol{\tau}$. Since the number of nodes is fixed, there are only a finite number of possible tree structures. Any discrete distribution over these tree structures is valid; in this paper, we use a uniform distribution over all trees in order to simplify the inference and reveal the model's underlying strengths and weaknesses. By performing Bayesian inference on the $\boldsymbol{\tau}$ variables along with the other hidden parameters of the HDP, we can infer the posterior distribution over taxonomy structures for any set of observed data.

The tree-HDP takes a different strategy than previous work for combining hierarchical structure with nonparametric Bayesian models, such as the nested CRP (Blei, Griffiths, and Jordan 2010). In the nested CRP, objects are associated with paths in an infinitely deep, infinitely wide tree, and overlapping paths represent similarities between objects at multiple levels of abstraction. In contrast, in the tree-HDP, objects are associated with nodes in a finite tree, and edges represent subset relations between categories. This latter strategy is what makes the tree-HDP natural for modeling multi-level category learning, where an object can have high probability under distributions at multiple levels of the tree.

Inference in the Tree-HDP

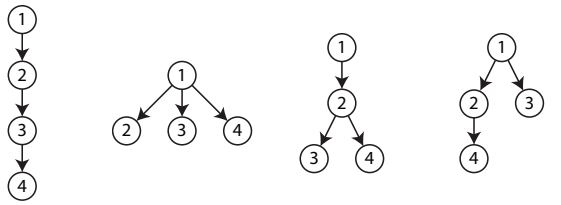
We now give a brief review of the inference procedure described by Teh et al. (2006) for the flat HDP and describe the steps necessary to extend the algorithm for the tree-HDP. Let z_{ji} denote the mixture component associated with x_{ji}, the ith observation from category j. Let m_{jk} denote the number of tables in category j assigned to mixture component k. The weight of mixture component k is denoted by β_{0k} in the global measure G_0 and by β_{jk} in the measure G_j of category j. Note that this differs from the notation of Teh et al. (2006), where the global component weights are called β_k and the category-specific weights are called π_{jk}. The index of the parent of category j is given by τ_j. We relax the assumption that all the categories share a common concentration parameter α_0; instead, the concentration parameter for category j is denoted α_j, and the concentration parameter for the global measure G_0 is denoted α_0 instead of γ.

We use the "posterior sampling by direct assignment" method of Gibbs sampling inference, described in Section 5.3 of Teh et al. (2006). In this method, MCMC inference is performed over the variables $\mathbf{z} = \{z_{ji}\}$, $\mathbf{m} = \{m_{jk}\}$, and $\boldsymbol{\beta}_0 = \{\beta_{0k}\}$. In the flat HDP, the β_{jk} variables can be integrated out because the categories always occupy the leaves of the tree. However, in the tree-HDP, because categories can have other categories as children, the $\boldsymbol{\beta}_j$ variables must be explicitly represented and sampled in the MCMC algorithm.

For notational convenience, we also define the following variables. Let n_{jk} denote the number of observations from category j assigned to mixture component k. Let $v_{jk} = \sum_{j':\tau_{j'}=j} m_{j'k}$ denote the number of tables among the children of category j that are assigned to mixture component k. Each such table constitutes a "virtual" observation from mixture component k in category j, and these counts are necessary for computing many of the quantities used in the inference algorithm. These counts do not appear in the standard HDP formulation because categories don't have other categories as children, making $v_{jk} = 0$ for all categories.

The Gibbs sampling equations are as follows (Teh et al. 2006). The value of $P(z_{ji} = k|\mathbf{z}^{-ji}, \mathbf{m}, \boldsymbol{\beta}, \boldsymbol{\tau})$ is

$$\begin{cases} (n_{jk}^{-ji} + v_{jk} + \alpha_j \beta_{\tau_j k}) f_k^{-x_{ji}}(x_{ji}) & \text{for old } k, \\ \alpha_j \beta_{\tau_j u} f_{k^{\text{new}}}^{-x_{ji}}(x_{ji}) & \text{for } k = k^{\text{new}}. \end{cases} \quad (2)$$



(a) Tree 1 (b) Tree 2 (c) Tree 3 (d) Tree 4

Figure 1: Tree structures used to create the simulated data.

The expression for $P(m_{jk}=m|\mathbf{z},\mathbf{m}^{-jk},\boldsymbol{\beta},\boldsymbol{\tau})$ is

$$\frac{\Gamma(\alpha_j\beta_{\tau_j k})}{\Gamma(\alpha_j\beta_{\tau_j k}+n_{jk}+v_{jk})}s(n_{jk}+v_{jk},m)(\alpha_j\beta_{\tau_j k})^m, \quad (3)$$

where $s(n,m)$ are unsigned Stirling numbers of the first kind. For $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_j$, we have

$$(\beta_{01}\ldots\beta_{0K},\beta_{0u})|\mathbf{z},\mathbf{m},\boldsymbol{\beta}^{-0},\boldsymbol{\tau}\sim\text{Dir}(v_{01}\ldots v_{0K},\alpha_0) \quad (4)$$

$$(\beta_{j1}\ldots\beta_{jK},\beta_{ju})|\mathbf{z},\mathbf{m},\boldsymbol{\beta}^{-j},\boldsymbol{\tau}\sim\text{Dir}(p_{j1}\ldots p_{jK},\alpha_j\beta_{\tau_j u})$$

where Dir is the Dirichlet distribution and p_{jk} is notational shorthand for $n_{jk}+v_{jk}+\alpha_j\beta_{\tau_j k}$. Finally, for sampling the new $\boldsymbol{\tau}$ variables, we have

$$P(\tau_j=t|\mathbf{z},\mathbf{m},\boldsymbol{\beta},\boldsymbol{\tau}^{-j}) \quad (5)$$
$$\propto P(\tau_j=t|\boldsymbol{\tau}^{-j})P(\mathbf{z},\mathbf{m},\boldsymbol{\beta}|\tau_j=t,\boldsymbol{\tau}^{-j}).$$

Because we use a uniform prior, we have $P(\tau_j=t|\boldsymbol{\tau}^{-j})\propto 1$, so

$$P(\tau_j=t|\mathbf{z},\mathbf{m},\boldsymbol{\beta},\boldsymbol{\tau}^{-j})\propto P(\mathbf{z},\mathbf{m},\boldsymbol{\beta}|\tau_j=t,\boldsymbol{\tau}^{-j}) \quad (6)$$
$$=P(\boldsymbol{\beta}_{0\cdot})\textstyle\prod_j P(\boldsymbol{\beta}_{j\cdot}|\boldsymbol{\beta}_{\boldsymbol{\tau}_j\cdot})P(\mathbf{z}_{j\cdot}|\boldsymbol{\beta}_{\tau_j\cdot})P(\mathbf{m}_{j\cdot}|\boldsymbol{\beta}_{\tau_j\cdot},\mathbf{z}_{j\cdot})$$

Now we break down each term in this product:

$$\boldsymbol{\beta}_{j\cdot}|\boldsymbol{\beta}_{\boldsymbol{\tau}_j\cdot}\sim\text{Dir}(\alpha_j\beta_{\tau_j 1},\ldots,\alpha_j\beta_{\tau_j K},\alpha_j\beta_{\tau_j u}),$$

$$P(\mathbf{z}_{j\cdot}|\boldsymbol{\beta}_{\tau_j\cdot})=\textstyle\prod_k\frac{\Gamma(n_{jk}+\alpha_j\beta_{\tau_j k})}{\Gamma(\alpha_j\beta_{\tau_j k})}, \quad (7)$$

$$P(\mathbf{m}_{j\cdot}|\boldsymbol{\beta}_{\tau_j\cdot},\mathbf{z}_{j\cdot})=\textstyle\prod_k\frac{\Gamma(\alpha_j\beta_{\tau_j k})s(n_{jk},m_{jk})(\alpha_j\beta_{\tau_j k})^{m_{jk}}}{\Gamma(\alpha_j\beta_{\tau_j k}+n_{jk})}.$$

Combining these terms and dropping out the terms that don't depend on t, we have

$$P(\tau_j=t|\mathbf{z},\mathbf{m},\boldsymbol{\beta},\boldsymbol{\tau}^{-j}) \quad (8)$$
$$\propto P(\boldsymbol{\beta}_{j\cdot}|\boldsymbol{\beta}_{t\cdot})P(\mathbf{z}_{j\cdot}|\boldsymbol{\beta}_{t\cdot})P(\mathbf{m}_{j\cdot}|\boldsymbol{\beta}_{t\cdot},\mathbf{z}_{j\cdot})$$
$$\propto(\beta_{ju})^{\alpha_j\beta_{tu}}\textstyle\prod_k(\beta_{jk})^{\alpha_j\beta_{tk}}(\alpha_j\beta_{tk})^{m_{jk}}.$$

Since τ_j can only take on a finite number of values, we can compute the normalization factor by summing over t.

Evaluation on Recovering Taxonomies

To verify the ability of the tree-HDP to reconstruct taxonomies, we used it to infer some small taxonomies with simulated data. We built four different hierarchies of four categories each (see Figure 1) and used the HDP generative model conditioned on these taxonomy structures to sample 1000 observations from each category. Each category was modeled as mixture of Gaussian distributions with two independent dimensions. To make inference more tractable, we used a conjugate base measure (Normal-scaled inverse gamma) on each dimension, which had hyperparameters of $\lambda=0$, $\nu=0.01$, $\alpha=3$, and $\beta=1$. λ and ν control the prior distribution over the mixture components' locations, while α and β control the prior distribution over the mixture components' variance parameters. The Dirichlet process concentration parameter α_j was set to 10 for every category.

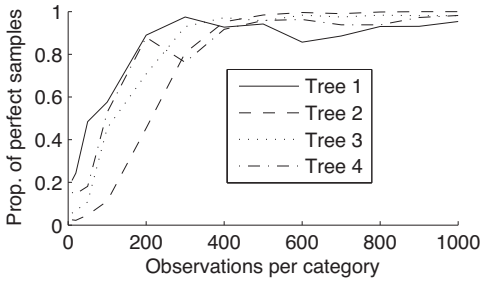


Figure 2: Inference results for the tree-HDP model on the simulated datasets. Curves show the proportion of MCMC samples which perfectly reconstruct the correct tree. The four tree structures are depicted in Figure 1.

For each of the four taxonomies, we ran the tree-HDP inference algorithm on the generated data, where the number of observations given to the model ranged from 10 to 1000. The base measure hyperparameters were matched to those used to generate the data. The Gibbs sampling procedure was run for 51,000 iterations in total, with samples being taken every 10 iterations after a burn-in period of 1000 iterations. The tree structure was initialized to a flat hierarchy (all τ_j variables set to 0), and sampling of the τ_j variables was not performed until halfway through the burn-in period to allow the model to find good clusterings of the observations before constraining the probabilities of the clusterings by committing to a deep tree structure.

The results of the inference are summarized in Figure 2. The model's performance was similar for all four trees. The proportion of samples which perfectly reconstructed the correct tree rose from less than 20% with 10 observations per category, to close to 100% with 400 observations per category. These results show that for these data, the method is consistent: with enough observations, it converges to the correct hidden taxonomy structure.

Experiment with Human Learners

In the previous section, we evaluated the performance of the tree-HDP on simulated data. Since we know that people are good at solving this problem, we were interested in comparing the model's performance to that of human learners. We conducted an experiment where the model and human learners performed a multi-level category learning task.

Method

Participants We recruited 95 participants from Amazon Mechanical Turk who received $1.00 as compensation. In addition, 95 undergraduate students participated for course credit. No significant differences were found between the two participant pools, so their data were pooled for analysis.

Stimuli We constructed an artificial taxonomy of 14 categories (see Figure 3(a), with the black arrows comprising the edges of the hierarchy). The categories were composed

of a total of 8 clusters of visual stimuli from (Sanborn, Griffiths, and Shiffrin 2010). The appearance of each stimulus is controlled by six continuous-valued parameters, and each cluster was defined by a multivariate Gaussian distribution over the values of these six parameters. Each of the eight categories at the leaves of the trees contained observations from only a single cluster; each of the four categories in the middle level contained observations from the two clusters of its descendants, and each of the two top-level categories contained observations from the four clusters below it.

Procedure Participants were first shown four examples of each category, for a total of 56 observations. Within each category, the examples were equally distributed among the category's clusters. An example of one round of training observations is shown in Figure 3(a). These observations were grouped on the computer screen according to their category labels, and they remained on-screen for the duration of the experiment. After the 56 observations were presented, the participants completed a test in which 28 stimuli were presented along with an attached category label. The participants were asked whether or not each category label was correct, and their responses were recorded. If they answered at least 26 of these trials correctly, they proceeded to the final test session; otherwise they repeated another round of observations and another 28-trial test session. Once the participants reached the performance criterion, they were asked to reconstruct the taxonomy corresponding to the categories they had just learned. They accomplished this by arranging 14 labels on the screen which contained the names of the categories, and then drawing arrows between them to indicate the parent-child hierarchical relationships. All participants completed this same task on a known taxonomy of fruit and vegetable categories beforehand to confirm that they properly understood the instructions.

Results

Out of the 190 participants, 78 (41%) perfectly reconstructed the correct tree structure. The total number of incorrect edges among all participants was 1401, an average of 7.37 errors per subject. Figure 3(a) depicts the aggregate responses from all the subjects. Black arrows, corresponding to edges in the correct taxonomy, were chosen 61–72% of the time, while gray arrows were chosen 5–15% of the time. The gray arrows are those that are not in the correct taxonomy but appeared with statistically significant frequency.[1]

These results show that human learners are able to accurately reconstruct taxonomy structures from a limited number of examples. The types of errors made are very systematic and give insight into the mental representations that people use to solve the problem. With one exception, all of the incorrect edges either point in the wrong direction (towards a category's ancestor) or point in the right direction but skip a level. The first type of error can be explained by people not adequately understanding the meaning of the arrows they were drawing; perhaps they had accidentally reversed the meaning of parent-child relationships. The second type of error shows that people sometimes do not classify categories at the lowest possible level, but occasionally produce taxonomies which are "flatter" than they could be otherwise.

[1] $p < 0.05$ according to an exact binomial test with success probability equal to the overall average frequency of edges across all human results.

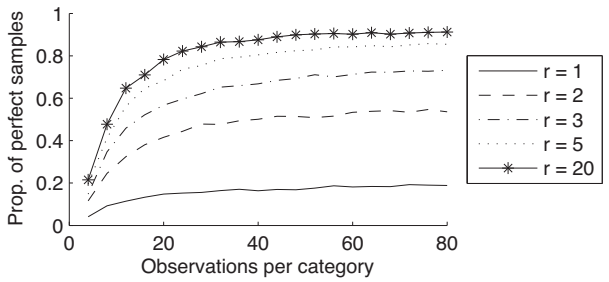


Figure 4: Performance of the tree-HDP on the experimental data. Results are shown for five different settings of r, the inverse temperature of the sampling distribution over tree structures. Each curve shows the proportion of samples which perfectly reconstruct the correct tree structure.

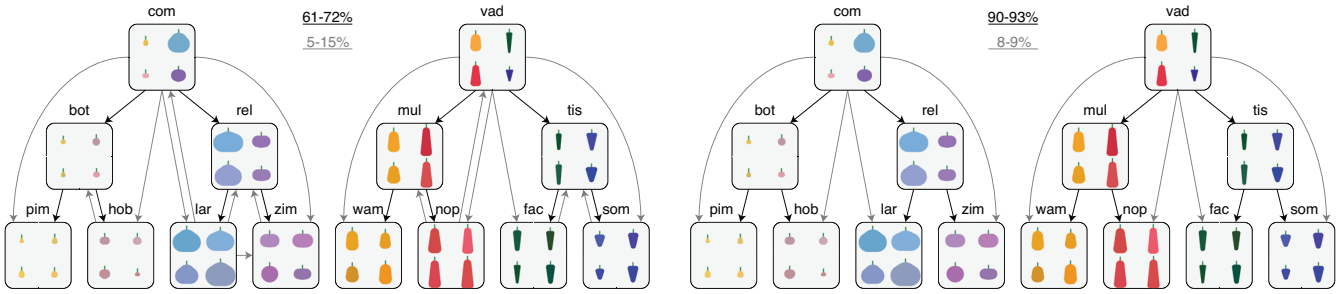
Modeling

We simulated the tree-HDP model on the same task that the human learners completed. In order to explore the variability of the model's performance, we trained it on a range of 4–80 observations per category. The human learners observed 5.3 examples per category, on average. As with the simulated data, the model represented each category as a mixture of Gaussian distributions with six independent dimensions. The hyperparameters of the the conjugate base measure (again, Normal-scaled inverse gamma) were fit to the training data, with parameters $\lambda = 0$, $\nu = 0.01$, $\alpha = 1.6$, and $\beta = 6.3$. The Dirichlet process concentration parameters were inferred from the data separately for each category, using a Gamma$(1, 0.01)$ prior distribution. To focus the posterior more on high-probability trees, we ran versions of the model at "temperatures" of $1/r$ for values of r between 1 and 20, corresponding to raising the Gibbs sampling distribution for τ_j in the MCMC algorithm to the power of r.

The model results are shown in Figure 4. In general, more observations per category and higher values of r both led to better performance. The model's performance covers a wide range of values (4–91%) depending on these two parameters. The aggregated samples from one version of the model (with 8 observations per category and $r = 3$) is shown in Figure 3(b). In general, the model very accurately reconstructed the taxonomy, and interestingly, the mistakes it makes are very similar to those of the human learners. Ignoring the "backwards" edges produced by the human learners, there is only a single difference in the significant mistakes made by the model and the people: the edge from "lar" to "zim". The correlation between the edge frequencies in people's reconstructed hierarchies and the model's samples was 0.988.

Conclusion

Learning the conceptual structures that characterize our world requires being able to induce relationships between categories from examples of their members. We have pre-



(a) Human learners (b) Tree-HDP model

Figure 3: Results for Experiment 1 for the (a) human learners, and the (b) tree-HDP model. Black arrows were chosen by a majority of participants (or model samples), and gray arrows were chosen by a small but statistically significant number of participants (or samples).

sented a nonparametric Bayesian model that can be used to solve the problem of multi-level category learning and shown that its performance is similar to that of humans. This analysis helps explain how it is possible for learners to induce taxonomies from only labeled observations, and provides a new tool for learning categories in contexts where the assumption that categories are independent of one another is invalid. In future work, we hope to extend this analysis to incorporate direct statements of the relations between categories, as might be provided in verbal instruction or found through text mining, and consider how our approach can be extended to more complex conceptual structures.

Acknowledgements

This work was supported by grants IIS-0845410 from the National Science Foundation and FA-9550-10-1-0232 from the Air Force Office of Scientific Research.

References

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

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

Ashby, F. G., and Alfonso-Reese, L. 1995. Categorization as probability density estimation. *Journal of Math. Psych.* 39:216–233.

Atran, S. 1998. Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences* 21:547–609.

Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57(2):1–30.

Canini, K. R.; Shashkov, M. M.; and Griffiths, T. L. 2010. Modeling transfer learning in human categorization with the hierarchical Dirichlet process. In *Proc. ICML-27*.

Collins, A., and Quillian, M. 1969. Retrieval time from semantic memory. *Verbal Learning & Verbal Behaviour* 8:240–247.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2000. *Pattern classification*. New York: Wiley.

Griffiths, T. L.; Canini, K. R.; Sanborn, A. N.; and Navarro, D. J. 2007. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proc. CogSci-29*.

Heller, K. A., and Ghahramani, Z. 2005. Bayesian hierarchical clustering. In *Proc. ICML-22*.

Keil, F. C. 1979. *Semantic and conceptual development: An ontological perspective*. Cambridge: Harvard Univ. Press.

Kemp, C., and Tenenbaum, J. B. 2009. Structured statistical models of inductive reasoning. *Psych. Review* 116(1):20–58.

Kemp, C.; Tenenbaum, J. B.; Griffiths, T. L.; Yamada, T.; and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. In *Proc. AAAI-21*.

Love, B.; Medin, D.; and Gureckis, T. 2004. SUSTAIN: A network model of category learning. *Psych. Review* 111(2):309–332.

Medin, D. L., and Schaffer, M. M. 1978. Context theory of classification learning. *Psych. Review* 85(3):207–238.

Neal, R. M. 1998. Markov chain sampling methods for dirichlet proces mixture models. Technical Report 9815, Department of Statistics, University of Toronto.

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

Reed, S. K. 1972. Pattern recognition and categorization. *Cognitive Psych.* 3:393–407.

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

Rosseel, Y. 2002. Mixture models of categorization. *Journal of Math. Psych.* 46:178–210.

Roy, D.; Kemp, C.; Mansinghka, V.; and Tenenbaum, J. 2007. Learning annotated hierarchies from relational data. In *Proc. NIPS-19*.

Sanborn, A. N.; Griffiths, T. L.; and Navarro, D. J. 2006. A more rational model of categorization. In *Proc. CogSci-28*.

Sanborn, A. N.; Griffiths, T. L.; and Shiffrin, R. M. 2010. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psych.* 60(2):63–106.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476):1566–1581.

Vanpaemel, W., and Storms, G. 2008. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review* 15(4):732–749.

Zhu, X.; Gibson, B. R.; Jun, K.-S.; Rogers, T. T.; Harrison, J.; and Kalish, C. 2010. Cognitive models of test-item effects in human category learning. In *Proc. ICML-27*.