A Rational Analysis of Rule-based Concept Learning

Noah D. Goodman[1] (ndg@mit.edu), **Thomas Griffiths**[2] (tom_griffiths@berkeley.edu),
Jacob Feldman[3] (jacob@ruccs.rutgers.edu), **Joshua B. Tenenbaum**[1] (jbt@mit.edu)
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Psychology, University of California, Berkeley
[3]Department of Psychology, Center for Cognitive Science, Rutgers University

Abstract

We propose a new model of human concept learning that provides a rational analysis for learning of feature-based concepts. This model is built upon Bayesian inference for a grammatically structured hypothesis space—a "concept language" of logical rules. We compare the model predictions to human generalization judgments in two well-known category learning experiments, and find good agreement for both average and individual participants' generalizations.

Keywords: concept learning; categorization; Bayesian induction; probabilistic grammar; rules.

Introduction

Concepts are a topic of perennial interest to psychology, particularly concepts which identify kinds of things. Such concepts are mental representations which enable one to discriminate between objects that satisfy the concept and those which do not. Given their discriminative use, a natural hypothesis is that concepts are simply rules for classifying objects based on features. Indeed, the "classical" theory of concepts (see Smith and Medin, 1981) takes this viewpoint, suggesting that a concept can be expressed as a simple feature-based rule: a conjunction of features that are necessary and jointly sufficient for membership. Early models based on this approach failed to account for many aspects of human categorization behavior, especially the graded use of concepts (Mervis and Rosch, 1981). Attention consequently turned to models with a more statistical nature: similarity to prototypes or to exemplars (Medin and Schaffer, 1978; Kruschke, 1992; Love et al., 2004). The statistical nature of many of these models has made them amenable to a rational analysis (Anderson, 1990), which attempts to explain why people do what they do, complementing (often apparently ad-hoc) process-level accounts.

Despite the success of similarity-based models, recently renewed interest has led to more sophisticated rule-based models. Among the reasons for this reconsideration are the inability of similarity-based models to provide a method for concept combination, common reports by participants that they "feel as if" they are using a rule, and the unrealistic memory demands of most similarity-based models. The RULEX model (Nosofsky et al., 1994), for instance, treats concepts as conjunctive rules plus exceptions, learned by a heuristic search process, and has some of the best fits to human experimental data—particularly for the judgments of individual participants. Parallel motivation for reexamining the role of logical structures in human concept representation comes from evidence that the difficulty of learning a new concept is well predicted by its logical complexity (Feldman, 2000).

However, existing rule-based models are primarily heuristic—no rational analysis has been provided, and they have not been tied to statistical approaches to induction. A rational analysis for rule-based models might assume that concepts are (represented as) rules, and ask what degree of belief a rational agent should accord to each rule, given some observed examples. We answer this question by formulating the hypothesis space of rules as words in a "concept language" generated by a context-free grammar. Considering the probability of productions in this grammar leads to a prior probability for words in the language, and the logical form of these words motivates an expression for the probability of observed examples given a rule. The methods of Bayesian analysis then lead to the Rational Rules model of concept learning. This grammatical approach to induction has benefits for Bayesian rational analysis: it compactly specifies an infinite, and flexible, hypothesis space of structured rules and a prior that decreases with complexity. The Rational Rules model thus makes contributions to both rule-based concept modeling and rational statistical learning models: to the former it provides a rational analysis, and to the latter it provides the grammar-based approach. Across a range of experimental tasks, this new model achieves comparable fits to the best rule-based models in the literature, but with fewer free parameters and arbitrary processing assumptions.

An Analysis of Concepts

A general approach to the rational analysis of inductive learning problems has emerged in recent years (Anderson, 1990; Tenenbaum, 1999; Chater and Oaksford, 1999). Under this approach a space of hypotheses is posited, and beliefs are assigned using Bayesian statistics—a coherent framework that combines data and a priori knowledge to give posterior degrees of belief. Uses of this approach, for instance in causal induction (Griffiths and Tenenbaum, 2005) and word learning (Xu and Tenenbaum, 2005), have successfully predicted human generalization behavior in a range of tasks.

In our case, we wish to establish a hypothesis space of rules, and analyze the behavior of a rational agent trying to learn those rules from labeled examples. Thus the learning problem is to determine $P(F|\mathbf{E},\ell(\mathbf{E}))$, where F ranges over rules, $\mathbf{E}$ is the set of observed example objects (possibly with repeats) and $\ell(\mathbf{E})$ are the observed labels. (Throughout this section we consider a single labeled concept, thus $\ell(x) \in \{0,1\}$ indicates whether x is an example or a non-example of the concept.) This quantity may be expressed

(through Bayes' formula):

$$P(F|\mathbf{E},\ell(\mathbf{E})) \propto P(F)P(\mathbf{E},\ell(\mathbf{E})|F) \tag{1}$$

To use this relationship we will need, in addition to a hypothesis space, the prior probability, $P(F)$, and a likelihood function, $P(\mathbf{E},\ell(\mathbf{E})|F)$.

Concept Representation

The hypothesis space of rules is given by well formed formulae of a concept language, which is specified by a context-free grammar over an alphabet of terminal symbols. In our case the terminal symbols are logical connectives ($\neg,\wedge,\vee$), grouping symbols, and a set of feature predicates. The feature predicates are formed from functions $f_i(x)$, which report the value of a physical feature, and the operator $=c$, which represents comparison with constant c: each feature predicate is of the form $f_i(x)=c$ (read "the i^{th} feature of object x has value c"). For brevity we consider here only Boolean features: $f_i(x) \in \{0,1\}$ and $c \in \{0,1\}$. (The extension to continuous features may be made in a straightforward manner, replacing the equality operator by inequality.)

The DNF grammar (Fig. 1) generates formulae in disjunctive normal form (ie. disjunctions of conjunctions of features). Each concept in this language consists of a set of "definitions", the B non-terminals, and each definition consists of necessary and sufficient features (the P non-terminals, which become feature predicates). For instance, from the start symbol S, we might generate two "definitions", $(B)\vee(B)$; from these we perhaps reach $(P\wedge P)\vee(P)$; each P is specialized to a feature: $(D_2\wedge D_1)\vee(D_4)$; and, finally $(f_2(x)=1\wedge f_1(x)=0)\vee(f_4(x)=0)$. We will focus on the DNF grammar, a natural interpretation of the classical theory of concepts, for the present exercise, but it is by no means the only (or most interesting) possibility. For instance, one may formulate a grammar inspired by the representations of Nosofsky et al. (1994), which generates a rule plus exceptions, or a grammar of implications, inspired by Feldman (2006), which represents causal structure.

The concept language generates well formed formulae of first-order logic, which allows us to recursively evaluate a formula on a given object (which is the usual approach in mathematical logic (Enderton, 1972)). Briefly, each term involving a logical connective can be evaluated in terms of its constituents, and, presuming that we know the feature values for the object x, we can evaluate all of the feature predicates. This assigns a truth value $F(x)$ to formula F for each object x.

A Syntactic Prior

By supplementing the grammar with a probability for each production we get a prior over the formulae of the language. The probability of a given parse (sequence of productions) is:

$$P(T|\mathcal{G},\tau) = \prod_{s\in T}\tau(s), \tag{2}$$

$$\begin{array}{lcl} S & \rightarrow & (B)\vee S \\ S & \rightarrow & (B) \\ B & \rightarrow & B\wedge P \\ B & \rightarrow & P \\ P & \rightarrow & D_1 \\ & \vdots & \\ P & \rightarrow & D_N \\ D_1 & \rightarrow & f_1(x)=1 \\ D_1 & \rightarrow & f_1(x)=0 \\ & \vdots & \\ D_N & \rightarrow & f_N(x)=1 \\ D_N & \rightarrow & f_N(x)=0 \end{array}$$

Figure 1: The DNF Grammar. S is the start symbol, and B,P,D_i the other non-terminals. $f_i(x)$ is the value of the i^{th} feature.

where $\mathcal{G}$ is the grammar, $s\in T$ are the productions of parse T, and $\tau(s)$ their probability. ($\tau(s)$ sums to one over the productions of each non-terminal.) The DNF grammar is a unique production grammar—there is a single production for each well-formed formula—so the probability of a formula is also given by Eq. 2. (We write F below for both the formula and its parse.) Note that this prior captures a simplicity bias: syntactically shorter formulae have smaller parse trees, thus higher prior probability.

We have no a priori reason to prefer one set of values for τ to another, so we assume a uniform prior over the possible values of τ. The probability of a parse becomes:

$$\begin{aligned} P(T|\mathcal{G}) &= \int P(\tau)\prod_{s\in F}\tau(s)d\tau \\ &= \int \prod_{s\in F}\tau(s)d\tau \\ &= \prod_{Y\in N}\beta(\text{Count}(Y,F)+1), \end{aligned} \tag{3}$$

where $\beta(\cdot)$ is the normalizing constant of the Dirichlet distribution, and $\text{Count}(Y,F)$ is the vector of counts of the productions for non-terminal symbol Y in the unique parse of F.

Likelihood: Evaluation and Outliers

To derive a likelihood function, we begin by making the *weak sampling* assumption, that the set of observed examples is independent of the concept:

$$P(\mathbf{E},\ell(\mathbf{E})|F) = P(\ell(\mathbf{E})|F,\mathbf{E})P(\mathbf{E}). \tag{4}$$

The term $P(\mathbf{E})$ will cancel from our calculations when all feature values are observed for all objects. Next we assume that the label is true exactly when an object satisfies the hypothesized formula. Thus, if we knew that the observed labels were correct, and we required an explanation for each observation, this likelihood would reduce to evaluation of the formula for each example (logical *true* is interpreted as probability 1, etc.):

$$P(\ell(\mathbf{E})|F,\mathbf{E}) = \bigwedge_{x\in\mathbf{E}}\ell(x)\leftrightarrow F(x). \tag{5}$$

However, to allow concepts that explain only some of the observations, we assume that there is a probability e^{-b} that any given observation is an outlier (ie. an unexplainable observation which should be excluded from induction). Writing S for the set of examples which are *not* outliers, the likelihood becomes:

$$\begin{aligned} P(\ell(\mathbf{E})|F,\mathbf{E}) &= \sum_{S\subseteq\mathbf{E}} (1-e^{-b})^{|S|}(e^{-b})^{|\mathbf{E}|-|S|} \bigwedge_{x\in S} \ell(x)\leftrightarrow F(x) \\ &= \sum_{S\subseteq\{x\in\mathbf{E}|\ell(x)\leftrightarrow F(x)\}} (1-e^{-b})^{|S|}(e^{-b})^{|\mathbf{E}|-|S|} \\ &= e^{-b|\{x\in\mathbf{E}|\neg(\ell(x)\leftrightarrow F(x))\}|}, \end{aligned} \tag{6}$$

where the second step follows from the Binomial Theorem.

The Rational Rules Model

The above likelihood and prior, combined using Bayes' rule, constitute a model of concept learning, which we call the Rational Rules model (RR_{DNF}, to indicate the grammar). The posterior probability for this model is:

$$P(F|\mathbf{E},\ell(\mathbf{E})) \propto \left(\prod_{Y\in N} \beta(|\{y\in F\}|+1)\right) e^{-b|\{x\in\mathbf{E}|\neg(\ell(x)\leftrightarrow F(x))\}|} \tag{7}$$

There is a trade-off in this posterior between explanatory completeness and conceptual parsimony. Though some examples may be ignored as outliers, concepts that explain more data are preferred by the likelihood function. On the other hand, simpler (i.e. syntactically shorter) formulae are preferred by the prior.

Using this posterior belief function, the generalization probability that a test object "t is an L" is:

$$P(\ell(t)|\mathbf{E},\ell(\mathbf{E})) = \sum_F F(t)P(F|\mathbf{E},\ell(\mathbf{E})). \tag{8}$$

This generalization probability represents inference by the "ideal learner". Initially we assume, consistent with standard practices of Bayesian modeling, that the average of the human population matches this ideal. At the end of the next section we consider the relationship between the ideal learner and individual human learners.

Comparison with Human Category Learning

In the preceding sections we have presented a rational analysis of concept learning when concepts are represented in a conceptual language of propositional rules. In this section we explore the extent to which this rational analysis can explain human learning. We will consider two experiments from the concept learning literature that have often been used as initial tests for modeling efforts.

In the experiments considered below participants were required to distinguish between two categories, A and B, which were mutually exclusive. For simplicity in fitting the model we assume that the population is an even mixture of people who take A to be the main category, and B the contrast category, with vice versa. Since these experiments have similar numbers of A and B examples, this is probably a reasonable initial assumption.

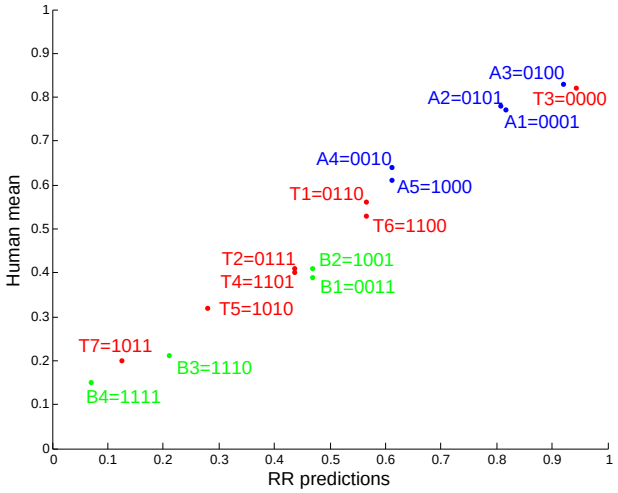

Figure 2: The category structure of Medin and Schaffer (1978), plotted on model vs. human results. Each example has four binary features; A1-A5 are examples of category A, B1-B4 of B, and T1-T7 are transfer objects. The human mean generalization rates of Nosofsky et al. (1994) are plotted against the predictions of the RR_{DNF} model (b=1). The model accounts for R^2=0.98 of the variance in human data.

Participants were trained on the category structures in a blocked-learning paradigm: each example in the training set was presented once per block, and blocks were presented until the training set could be classified accurately (relative to a predetermined threshold). It is often the case that different effects occur as training proceeds, and these effects can be tricky to capture in a rational model. However, it is apparent from Eq. 6 that the Rational Rules model with outlier parameter b and N identical blocks of examples is equivalent to the model with one block and parameter $N\cdot b$. This makes intuitive sense: the more often an example is seen, the less likely it is to be an outlier. Thus we may roughly model the course of human learning by varying the b parameter – effectively assuming a constant outlier probability while increasing the number of trials.

The model was approximated by Monte Carlo simulation (30,000 samples for each run). Except where otherwise noted we have coarsely optimized over b by taking the best fitting result from among b=$\{1,2,\ldots,8\}$.

Prototype Enhancement and Typicality Effects

The second experiment of Medin and Schaffer (1978), among the first studies of ill-defined categories, used the category structure shown on Fig. 2 (we consider the human data from the Nosofsky et al. (1994) replication of this experiment, which counter-balanced physical feature assignments). This experiment is a common first test of the ability of a model

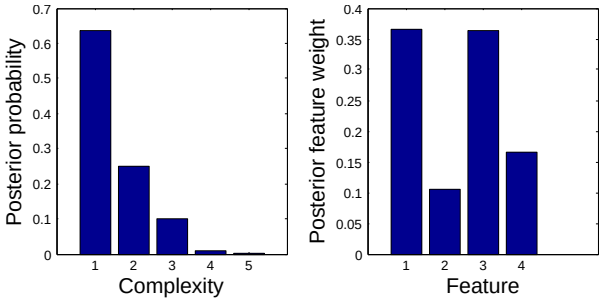

Figure 3: Posterior complexity and feature weight distributions for RR_{DNF} on the category of Medin and Schaffer (1978) (see Fig. 2). The model focuses on simple rules along dimensions 1 and 3.

to predict the pattern of generalizations on novel stimuli, and demonstrates two important effects: prototype enhancement (Posner and Keele, 1968), and (a certain flavor of) typicality.

The overall fit of the Rational Rules model (Fig. 2) is good: R^2=0.98. Other models of concept learning are also able to fit this data quite well: for instance R^2=0.98 for RULEX, and R^2=0.96 for the context model (Medin and Schaffer, 1978). However, RR_{DNF} has only a single parameter (the outlier parameter), while each of these models has four or more free parameters; indeed, the full RULEX model has nine free parameters (whose interpretation is not entirely clear).

In Fig. 3 we have plotted the posterior probability mass that the RR_{DNF} model places on all formulae of a given complexity and the posterior feature weights—the expected importance of each feature. (Complexity is measured by Boolean complexity: the number of feature symbols in the formula). We see that the RR_{DNF} model solves this concept learning problem, as human learners do, by placing most of its weight on simple formulae along features 1 and 3. It has been noted before (Navarro, 2006) that selective attention effects, like this one, emerge naturally from the Bayesian framework.

The object T3=0000 is the prototype of category A, in the sense that most of the examples of category A are similar to this object (differ in only one feature) while most of the examples of category B are dissimilar. Though it never occurs in the training set, the importance of this prototype is reflected in the human transfer judgments: T3 is, by far, the most likely transfer object to be classified as category A. The Rational Rules model predicts this prototype enhancement. The simple formulae $f_1(x)$=0 and $f_3(x)$=0 each have high posterior probability, these agree on the categorization of T3 and so combine (together with many lower probability formulae) to enhance the probability that T3 is in category A.

The degree of typicality, or recognition rate for training examples, is often taken as a useful proxy for category centrality (Mervis and Rosch, 1981) because it correlates with many of the same experimental measures (such as reaction time). Indeed we see greater typicality for the prototype of category B, the object B4=1111, than for other training examples: though presented equally often it is classed into category B far more often. The Rational Rules model also predicts this typicality effect, in a manner similar to prototype enhancement.

Table 1: The category structure of Medin et al. (1982), with initial and final block mean human responses of McKinley and Nosofsky (1993), and the predictions of RR_{DNF} at b=1 and b=8.

	Object	Human: Initial	Final	RR_{DNF}: b=1	b=8
A1	1111	0.64	0.96	0.84	1
A2	0111	0.64	0.93	0.54	1
A3	1100	0.66	1	0.84	1
A4	1000	0.55	0.96	0.54	0.99
B1	1010	0.57	0.02	0.46	0
B2	0010	0.43	0	0.16	0
B3	0101	0.46	0.05	0.46	0.01
B4	0001	0.34	0	0.16	0
T1	0000	0.46	0.66	0.2	0.56
T2	0011	0.41	0.64	0.2	0.55
T3	0100	0.52	0.64	0.5	0.57
T4	1011	0.5	0.66	0.5	0.56
T5	1110	0.73	0.36	0.8	0.45
T6	1101	0.59	0.36	0.8	0.44
T7	0110	0.39	0.27	0.5	0.44
T8	1001	0.46	0.3	0.5	0.43

Correlated Dimensions

Medin et al. (1982) studied the category structure shown in Table 1. This structure affords two strategies: the first two features are individually diagnostic of category membership, but not perfectly so, while the correlation between the third and fourth features is perfectly diagnostic. It was found that human learners relied on the perfectly diagnostic, but more complicated, correlated features. McKinley and Nosofsky (1993) replicated this result, studying both early and late learning by eliciting transfer judgments after initial and final training blocks. They found that human participants relied primarily on the individually diagnostic dimensions in the initial stage of learning, and transitioned to the correlated features later in learning. The RR_{DNF} model explains most of the variance in human judgments in the final stage of learning, R^2=0.95 when b=8; see Fig. 4. Correlation with human judgments after one training block is also respectable: R^2=0.69 when b=1. Thus, the course of human learning in this experiment can be modeled quite well, as predicted, by varying the outlier parameter b. By comparison RULEX best fits are R^2=0.99 for final, and R^2=0.67 for initial learning block, but with several extra free parameters.

Individual Generalization Patterns

Nosofsky et al. (1994) investigated the pattern of generalizations (i.e. the sequence of responses to transfer questions) made by individual participants, in addition to the group averages. One may wonder whether it is necessary to consider these individual differences. As noted in Nosofsky and Palmeri (1998), even the best binomial model (a model that predicts the averaged data perfectly, but assumes that all individuals behave identically) does very poorly at predicting individual generalization patterns (in this case R^2=0.24).

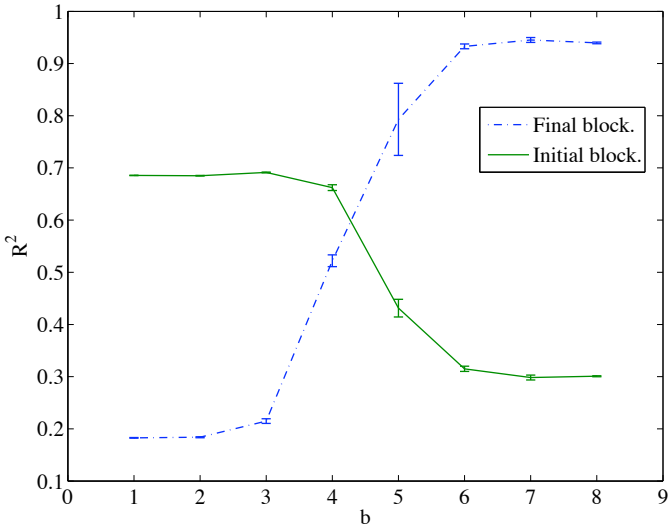

Figure 4: Comparison of human judgments with RR_{DNF} model predictions. Fit of RR_{DNF} model to the initial and final block mean human generalization pattern for several values of the outlier parameter, see Table 1. The model fits the early learning data well at small values of b, and the late learning data well for larger values.

To model individual generalizations using the Rational Rules model we make an additional assumption about how individuals approximate the full posterior distribution over formulae. The simplest assumption (similar in spirit to RULEX) is that each participant uses a single formula, sampled randomly from the posterior. Note that this predicts the same group averages, for large enough populations, but provides a psychological model more similar to existing processoreinted models of rule use. This interpretation of the RR_{DNF} model explains $R^2 = 0.83$ of the variance in human generalization for the 36 individual generalization patterns reported in Nosofsky et al. (1994). However, it is troubling to reintroduce deterministic rule use in this way, considering evidence of graded category centrality effects even within individuals. A similar, but more flexible, possibility is that each individual considers a small number of rules, and weights them appropriately. This is similar to the idealized cognitive models suggested by Lakoff (1987), and explains how individuals could exhibit graded behavior, if weight is spread among several rules, or almost deterministic behavior, if a single rule predominates. A model in which individuals take two or three samples from the RR_{DNF} posterior still fits the data of Nosofsky et al. (1994) well: $R^2 = 0.87$. RULEX does similarly well, $R^2 = 0.86$, but again uses several additional parameters. As with RULEX, the qualitative match of the Rational Rules model to human judgments is quite good, as show in Fig. 5.

Discussion and Conclusion

We have suggested an approach for analyzing human concept learning: assume that concepts are represented in a concept language, propose a specific grammar and inductive semantics for this language, then describe rational inference from examples to words of the language. Carrying out this scheme using a grammar for DNF formulae, we derived the Rational Rules (RR_{DNF}) model of concept learning. This model was shown to predict human judgments in several key category learning experiments, and to do so with only one, readily interpretable, parameter. The model was also used to predict the generalization judgments of individual learners, extending the usual reach of rational analysis.

To model individual judgments we assumed that each learner arrives at a small set of rules, chosen and weighted in accord with the complete rational model. This raises a question: how are individuals able to find such samples? One answer, which has been given by the RULEX model, is that learners apply a variety of hypothesis testing and generation heuristics. To understand how such heuristics are related to rational modeling, particularly to the RR_{DNF} model, investigation is needed into on-line algorithms which can approximate the Bayesian posterior while maintaining only a small number of hypotheses—that is, "rational process models". A useful initial exploration has been made by Sanborn et al. (2006), who focused on similarity-based models. (We have, in fact, verified some of the simulation results presented in this paper using a similar on-line sequential Monte Carlo algorithm, based upon Chopin (2002), but have not evaluated this algorithm for psychological relevance.)

As noted in Love et al. (2004), many of the ideas underlying SUSTAIN, and other modern descendants of exemplar models, are similar to those underlying RULEX. In particular, each attempts to provide a rich set of possible concepts, while controlling their complexity in any given instance. The grammar-based approach introduced here has a similar goal, but at a different descriptive level, and thus complements much of the earlier modeling work.

Unlike many approaches, the basic idea of grammar-based induction can be easily extended to new situations, such as role-governed or adjective-like concepts. Indeed, an important direction for future work concerns the representations of other types of concepts by extending the concept language to a larger fragment of first-order, or higher-order, logic.

The proposal that concepts are represented by words in a concept language is not new in cognitive science—indeed this is a principal component of the language of thought hypothesis (Fodor, 1975). Nor is the idea that cognition can be analyzed by considering an optimally rational agent new: ideal observers have been prominent in vision research (Geisler, 2003) and cognitive psychology (Anderson, 1990; Chater and Oaksford, 1999). However, the combination of these ideas leads to an exciting, and neglected, project: rational analysis of the language of thought. We have shown in this paper that

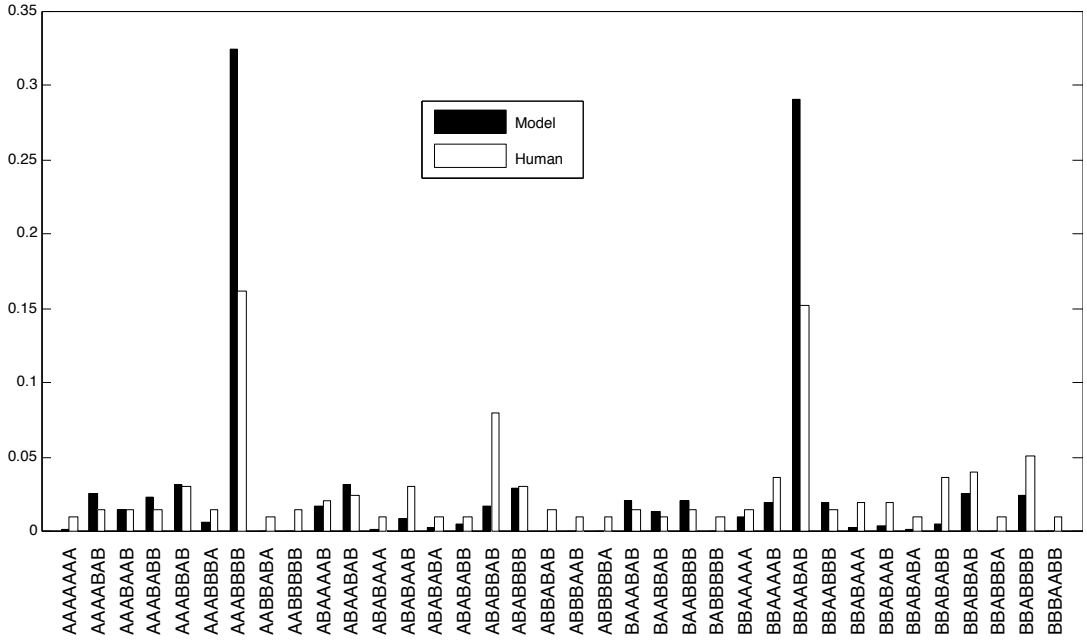

Figure 5: Individual generalization patterns: the portion of participants responding with the indicated categorizations for the seven transfer stimuli of Fig. 2. Human data from Nosofsky et al. (1994), Experiment 1. The model values assume 2 samples for each simulated participant, b=4.

rigorous results are possible in this program, and that they can provide accurate models of basic cognitive processes.

References

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

Chater and Oaksford (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, 3(2):57–65.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

Enderton, H. B. (1972). *A mathematical introduction to logic*. Academic Press, New York.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407:630–633.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology*, 50:339–368.

Fodor, J. A. (1975). *The language of thought*. Harvard University Press: Cambridge, MA.

Geisler, W. S. (2003). Ideal observer analysis. In Chalupa, L. and Werner, J., editors, *The Visual Neurosciences*, pages 825–837. MIT press.

Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51:285–386.

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

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press, Chicago.

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

McKinley, S. C. and Nosofsky, R. M. (1993). Attention learning in models of classification. (Cited in Nosofsky, Palmeri, and McKinley, 1994).

Medin, D. L., Altom, M. W., Edelson, S. M., and Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8:37–50.

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

Mervis, C. B. and Rosch, E. H. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32:89–115.

Navarro, D. J. (2006). From natural kinds to complex categories. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Nosofsky, R. M. and Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, 5:345–369.

Nosofsky, R. M., Palmeri, T. J., and McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101(1):53–79.

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

Sanborn, A., Griffiths, T., and Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

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

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in Neural information Processing Systems 11*.

Xu, F. and Tenenbaum, J. B. (2005). Word learning as Bayesian inference: Evidence from preschoolers. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*.