Grammar-Based Concept Learning with Numbers and Animals

9.66 Final Project

Leon Lin

Goodman et al. (2008) develop a model for concept learning through Bayesian inference based on the representation of concepts as logical formulae generated by a grammar. They show that the model explains human performance in learning various artificial concepts. In this project we apply largely the same model to the domains of numbers and animals. The grammar proves to be highly expressive, allowing for an exploration of the tradeoff between the complexity of a hypothesis and how well it fits the input data. We also compare the model's behavior to human intuitions, with mixed results.

Human minds have a striking ability to learn general concepts from just a small number of examples. Two or perhaps just one example of, say, a zebra or a wheel suffices to teach a child those respective categories.

In this project we build a general-purpose Bayesian rule-based concept learner (`https://github.com/leonxlin/concept-learner`) and compare it to human intuitions in three different domains: numbers, animals, and a simple artificial domain.

A grammar for concepts

In this project we will model the problem of concept learning much as Goodman, Tenenbaum, Feldman, & Griffiths (2008) do. We provide a condensed exposition of their framework, slightly modified to employ "strong sampling," following one of their suggestions.

Our concept learner lives in a *world* consisting of a set W of *objects*, each of which may exhibit some subset of the *features* in the set of possible features. The learner is aware of exactly which features each object exhibits.

It would be interesting to consider worlds with more structure — with nonbinary features or relationships between objects — but this setup turns out to be enough for the simple data sets that we will explore.

A subset $C \subset W$ of the objects belong to a *concept*, which the learner seeks to learn. A few objects are selected randomly from C (this is strong sampling), and perhaps a few are also selected from the complement of C. These few objects and an indication of whether each belongs to C are given to the learner (this is the *observed data*, or input). The learner receives incorrect information about whether a particular object is in C with probability e^{-b}, for some *outlier parameter* b, as in Goodman et al. (2008).

Given a few positive examples of a concept, there are potentially many possible larger sets C that contain those examples. How is the learner to choose from among them?

$$
\begin{aligned}
S &\to x \in C \Leftrightarrow \text{Disj} \\
\text{Disj} &\to (\text{Conj}) \vee \text{Disj} \\
\text{Disj} &\to \text{False} \\
\text{Conj} &\to P \wedge \text{Conj} \\
\text{Conj} &\to \text{True} \\
P &\to \text{Prim} \\
P &\to \neg\text{Prim} \\
\text{Prim} &\to f_1(x) \\
\text{Prim} &\to f_2(x) \\
&\vdots \\
\text{Prim} &\to f_N(x)
\end{aligned}
$$

Figure 1. A modified version of the production rules of the grammar used in Goodman et al. (2008) for formulae in disjunctive normal form when there are N features.

Feldman (2000) found that the most natural concepts are the simple ones, that is, those with short representations as boolean propositional formulae in terms of feature predicates f_i, where $f_i(x)$ is true if and only if object x exhibits feature i. Such a formula might look like

$$x \in C \Leftrightarrow (f_1(x) \wedge f_3(x)) \vee f_2(x).$$

Again following Goodman et al. (2008), we will consider formulae in *disjunctive normal form*, such as the one above. These formulae are disjunctions of conjunctions of terms of the form $f_i(x)$ or $\neg f_i(x)$. A grammar for such formulae is given in figure 1.[1] With the standard semantics, a formula

[1] Actually this grammar generates formulae like $x \in C \Leftrightarrow (f_1(x) \wedge f_3(x) \wedge \text{True}) \vee (f_2(x) \wedge \text{True}) \vee \text{False}$, which are equivalent.

generated by this grammar picks out a specific set of objects — a *hypothesis* for the concept.

Any partition of the objects into concepts and non-concepts is specified by some formula, so long as objects with the same features are categorized in the same way. Two formulae may define the same concept.

Prior, likelihood, and posterior

If the representation of learned concepts in the mind is something like these formulae, then a rational learner is interested in the most likely formulae given the observed data (and the world and the grammar, which we henceforth assume are always given). Bayes tells us that for any observed data D and a formula F,

$$P(F \mid D) \propto P(F)P(D \mid F).$$

Since we are using strong sampling, the likelihood that we observe D, supposing that F accurately describes the concept, obeys

$$P(D \mid F) \propto \binom{|C_F|}{|D \cap C_F|}^{-1} \binom{|W - C_F|}{|D \cap (W - C_F)|}^{-1} e^{-b|\operatorname{diff}(D,F)|},$$

where C_F is the set of objects belonging to the concept according to F; D, when treated as a set (in a slight abuse of notation), is the set of objects in the observed data; and $\operatorname{diff}(D, F)$ is the set of objects that D and F disagree about, that is, those $x \in D$ for which D says $x \in C$ but F says $x \notin C$, or vice versa.[2]

To apply Bayesian inference, we will also need a prior probability distribution $P(F)$ on the formulae. Remember that we want a prior that penalizes complicated formulae.

One way to get such a prior is to assign a probability distribution on the production rules that apply to each nonterminal symbol in the grammar. (Thus for instance we might have for the nonterminal symbol Conj weights of 0.3 and 0.7 on the rules 'Conj $\to P \wedge$ Conj' and 'Conj $\to$ True' respectively.) If τ is a such a set of *production probabilities*, then the probability $P(F \mid \tau)$ of a formula F is the product of the probabilities of each rule used in its derivation. For example, if there is a uniform probability distribution on the rules for expanding each nonterminal symbol, then picking a random formula from the prior is equivalent to starting with S and picking rules uniformly at random (from the applicable ones) until all symbols are terminal.

To avoid picking a particular set of production probabilities, we can integrate over them to obtain a prior showing no bias to any one of them. Goodman et al. (2008) derive this prior. It is

$$P(F) = \prod_Y \frac{\beta(\mathbf{C}_Y(F) + \mathbf{1})}{\beta(\mathbf{1})},$$

where F is a formula; the product is over nonterminal symbols Y of the grammar; β is the multinomial beta function

$$\beta(c_1, \ldots, c_n) = \frac{\prod_i \Gamma(c_i)}{\Gamma(\sum_i c_i)};$$

$\mathbf{C}_Y(F)$ is a vector giving the number of times each production rule applicable to Y is applied in the unique derivation of F (the order of the coordinates in the vector doesn't matter); and $\mathbf{1}$ is a vector of ones having the same length as $\mathbf{C}_Y(F)$, that is, having as many coordinates as there are rules applicable to Y.

Goodman et al. (2008) observe that this prior favors not only simple formulae but also the reuse of rules (and hence also the reuse of features predicates f_i).

The likelihood and the prior enable us to sample from the posterior $P(F \mid D)$, as is detailed in the next section. We assume that the learner assesses the probability that some object x is in C (the *generalization probability*) by consulting a few likely hypotheses drawn from the posterior (and weighting their judgments of x according to their respective posterior probabilities). Goodman et al. (2008) defend this choice — *hypothesis sampling* — but do not say exactly how many hypotheses (formulae) should be considered. The number ϕ of formulae to take from the top of the list may be used as another free parameter (along with b), but in this project we will always take $\phi = 10$.

Method

Given D we sample from the posterior distribution on formulae using the Metropolis-Hastings algorithm. Code implementing this algorithm is available at `https://github.com/leonxlin/concept-learner`.

We use the same Markov chain on formulae as Goodman et al. (2008) do. Given a formula F, a proposal formula F' is chosen by randomly selecting a nonterminal node of F and regenerating its subtree in the unique parse tree of F using a uniform set τ of production probabilities as described above. If $F = F'$ the process is repeated. Let $Q(F' \mid F)$ be the probability that a particular F' is proposed given F. With a counting argument one can show that

$$\frac{Q(F \mid F')}{Q(F' \mid F)} = \frac{|F|\, P(F \mid \tau)}{|F'|\, P(F' \mid \tau)},$$

where $|F|$ is the number of nonterminal nodes in the parse tree of F. Then accepting (transitioning from F to) F' with probability equal to the minimum of 1 and

$$\frac{P(F' \mid D)}{P(F \mid D)} \cdot \frac{Q(F \mid F')}{Q(F' \mid F)} = \frac{P(D \mid F')P(F')}{P(D \mid F)P(F)} \cdot \frac{|F|\, P(F \mid \tau)}{|F'|\, P(F' \mid \tau)}$$

makes the Markov chain converge to the posterior distribution $P(F \mid D)$.

[2] We assume $e^{-b} \ll 1$ so that we can ignore the factors $1 - e^{-b}$.

Object	Features	Human	Model, $b = 1.3$	$b = 5$
A1	0001	0.77	0.79	1.00
A2	0101	0.78	0.79	1.00
A3	0100	0.83	0.87	1.00
A4	0010	0.64	0.57	1.00
A5	1000	0.61	0.54	0.95
B1	0011	0.39	0.49	0.05
B2	1001	0.41	0.46	0.00
B3	1110	0.21	0.26	0.00
B4	1111	0.15	0.18	0.00
T1	0110	0.56	0.57	0.13
T2	0111	0.41	0.49	0.05
T3	0000	0.82	0.87	1.00
T4	1101	0.40	0.46	0.00
T5	1010	0.32	0.26	0.87
T6	1100	0.53	0.54	0.08
T7	1011	0.20	0.18	0.00

Figure 2. Objects and features in a world proposed by Medin & Schaffer (1978), with human data (recognition rates) from Nosofsky et al. (1994) and the generalization probabilities of our model. There are four possible features, and the features each object exhibits are indicated with binary strings (1 for yes and 0 for no). The objects A1–A5 are observed to be examples of the concept; the objects B1–B4 are observed to not be examples of the concept.

A four-feature world

As a sanity check we apply the model to a certain world of objects proposed by Medin & Schaffer (1978) and compare it to responses from humans gathered by Nosofsky et al. (1994), though they do not assume strong sampling.

The world consists of 16 objects, exhibiting all possible subsets of four features. Five of the objects are given as positive examples of a concept; four are given as negative examples.

The results of 30000 samples with $b = 1.3$ and 30000 samples with $b = 5$ are shown in figure 2. The fit between our model with $b = 1.3$ and the human data was $R^2 = 0.95$. The top $\phi = 10$ formulas sampled with $b = 1.3$ were as follows.

1. $x \in C \Leftrightarrow (\neg f_1(x) \wedge \text{True}) \vee \text{False}$
2. $x \in C \Leftrightarrow (\neg f_3(x) \wedge \text{True}) \vee \text{False}$
3. $x \in C \Leftrightarrow \text{False}$
4. $x \in C \Leftrightarrow (\text{True}) \vee \text{False}$
5. $x \in C \Leftrightarrow (\neg f_4(x) \wedge \text{True}) \vee \text{False}$
6. $x \in C \Leftrightarrow (\neg f_1(x) \wedge \neg f_1(x) \wedge \text{True}) \vee \text{False}$
7. $x \in C \Leftrightarrow (\text{True}) \vee (\text{True}) \vee \text{False}$
8. $x \in C \Leftrightarrow (\neg f_3(x) \wedge \neg f_3(x) \wedge \text{True}) \vee \text{False}$
9. $x \in C \Leftrightarrow (\text{True}) \vee (\text{True}) \vee (\text{True}) \vee \text{False}$
10. $x \in C \Leftrightarrow (\neg f_1(x) \wedge \neg f_3(x) \wedge \text{True}) \vee \text{False}$

The low outlier parameter, i.e., high tolerance for error in the observed data, combined with the prior's bias toward simple formulas allows such formulae as 'False' and '(True)∨False' to make the top ten. The nature of our grammar allows a high degree of redundancy both between and within formulae. Though the plausibility of some of these formulae as mental representations of concepts is dubious, the model seems to correlate well with human behavior.

With $b = 5$, the balance between fit and rule complexity tips toward the former, and for most of the top ten formulas the condition on $x \in C$ was a variation on

$$(\neg f_1(x) \wedge \neg f_3(x) \wedge \text{True}) \vee (\neg f_2(x) \wedge \neg f_4(x) \wedge \text{True}) \vee \text{False}.$$

A world of numbers

Tenenbaum (2000) found that a Bayesian model predicted human learning of categories of integers from 1 to 100. The model's hypothesis space included categories of multiples such as $\{2, 4, \ldots, 100\}$ and $\{9, 18, \ldots, 99\}$ and categories of consecutive numbers of such as $\{33, 34, \ldots, 41\}$, among other possible categories.

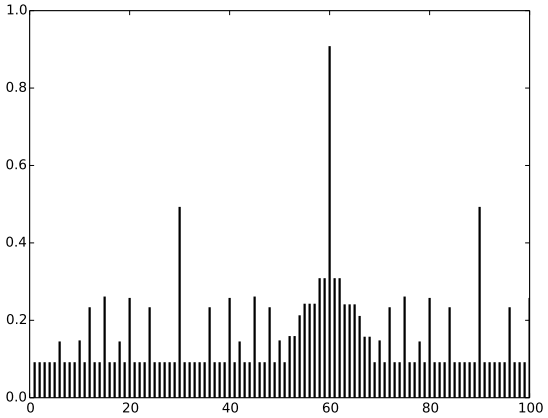
Accordingly we consider the following set of 62 features: for $i = 2, \ldots, 30$ let $f_i(x)$ be true when x is a multiple of i and for $i = 3, 6, \ldots, 99$ let $g_i(x)$ be true when $|x - i| \leq 5$. (These particular values of i are not special; they are chosen so that the total number of features is manageable.) These features, together with our grammar, allow us to express a wide variety of concepts, such as 'multiples of 3' and 'multiples of 20 and odd numbers greater than 90.'

Sampling formulae from the posterior with various observed data (e.g., $\{62, 52, 57, 55\}$), we find that the most likely formulae are mostly equivalent to 'True'. This happens because in the grammar, the large number of features causes any particular nontrivial conjunction to be quite unlikely in the prior, which causes the posterior to strongly favor formulae that don't involve any features. Therefore we will use the modified grammar in figure 3, which induces a prior much less heavily weighted toward formulae like '(True) ∨ (True) ∨ False'.

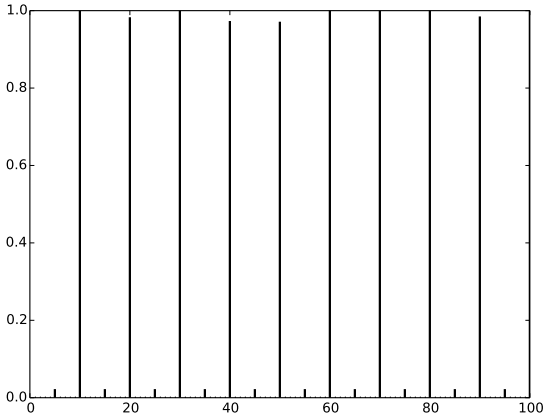
A clear size principle emerges with the change of grammar.[3] For example, if the observed data tells the learner only that $60 \in C$, then the top hypothesis is that C consists of multiples of 30: the formula $f_{30}(x)$ is only true for three numbers, whereas the formula $f_{10}(x)$, for instance, is true for ten numbers; this gives $f_{30}(x)$ a boost over $f_{10}(x)$ in the posterior.

[3] In this section, each finding comes uses 20000–30000 samples with $b = 1.3$. The number of samples had to be reduced to 20000 sometimes because our implementation, which memoizes formulas, tended to result in memory errors with 30000 samples on some imputs.

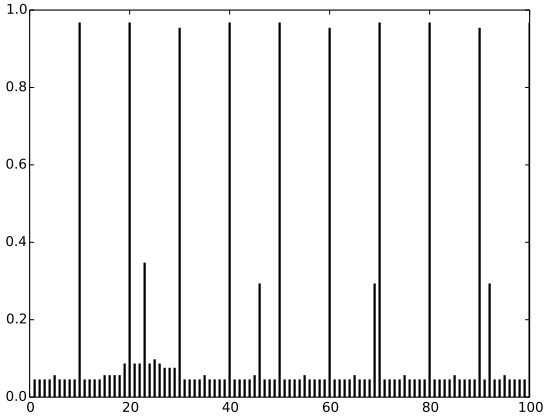
Observed: $60 \in C$



Observed: $\{60, 80, 10, 30\} \subset C$



Observed: $\{60, 80, 10, 30, 23\} \subset C$



Observed: $\{60, 80, 10, 30, 23, 25, 21\} \subset C$

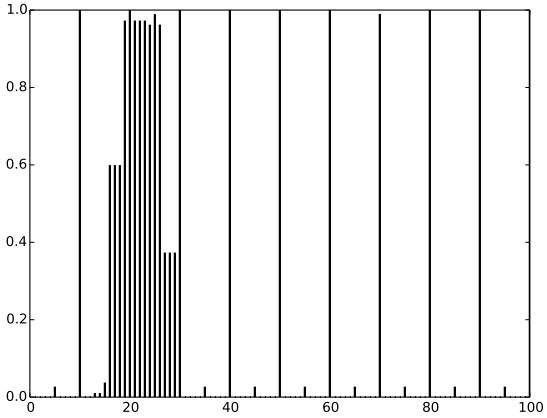


Figure 4. Generalization probabilities in the numbers world, according to the model.

However, the size principle, a consequence of the likelihood, is offset by the prior's aversion to complicated formulae. The formula $f_{30}(x) \wedge g_{60}(x)$ defines an even more exclusive club of just one, but it is beat by $f_{30}(x)$ in the posterior because its derivation has more steps.

The addition of evidence can push the posterior back in the direction of more complicated formulae. For example, with the evidence that $\{58, 59, 61, 62\} \subset C$, the top three formulae are

$$g_{63}(x) \quad g_{57}(x) \quad g_{63}(x) \wedge g_{57}(x).$$

The third formula here defines a narrower cluster of numbers. If new evidence falls outside this hypothesis, the model adjusts itself: told that $\{58, 59, 61, 62, 64, 65, 67, 68\} \subset C$, the model chooses

$$g_{63}(x) \quad \neg f_3(x) \wedge g_{63}(x) \quad g_{63}(x) \wedge \neg f_3(x)$$

as its top three formulae.

Figure 4 further illustrates some of the same principles. We did not gather human data, though it would be interesting to see whether the model matches humans learning of compound rules. On those input sets for which Tenenbaum (2000) provides human data, our model does not track the human responses as well as the original model does, but some of the qualitative effects are still present.

A world of animals

In this section we consider a world of 33 animals and 102 features provided by Kemp & Tenenbaum (2008), who give a graphical model that predicts the tree structure of animal categories. A tree structure strongly suggests a prior over concept hypotheses that favors concepts that correspond to subtrees of the hierarchy. The size principle then predicts

$$
\begin{aligned}
S &\rightarrow x \in C \Leftrightarrow \text{Disj} \\
\text{Disj} &\rightarrow (\text{Conj}) \vee \text{Disj} \\
\text{Disj} &\rightarrow \text{Conj} \\
\text{Conj} &\rightarrow P \wedge \text{Conj} \\
\text{Conj} &\rightarrow P \\
P &\rightarrow \text{Prim} \\
P &\rightarrow \neg\text{Prim} \\
\text{Prim} &\rightarrow \text{True} \\
\text{Prim} &\rightarrow f_1(x) \\
\text{Prim} &\rightarrow f_2(x) \\
&\vdots \\
\text{Prim} &\rightarrow f_N(x)
\end{aligned}
$$

Figure 3. A modified version of our grammar, designed to reduce the weight of formulas equivalent to True or False in the prior.

that the most likely concept given several positive examples is the smallest subtree containing them. Thus for poodles and labrador retrievers the category is dogs; for poodles, lions, and horses — mammals; and for poodles, fish, and roaches — all animals.

In many cases we find that the concept grammar model does show this smallest subtree effect. (With the original grammar the model runs into the same problem as it did with the number game, so again we use the grammar in figure 3.) Below are summaries of the model's generalizations (30000 samples each, $b = 3$) on various inputs.

Positive examples	Dog, Wolf
Likely formulae[4]	howls; is a canine ∧ eats rodents
Likely examples[5]	Dog, Wolf
Smallest subtree[6]	(same)
Positive examples	Cat, Dog, Wolf
Likely formulae	eats rodents
Likely examples	Cat, Dog, Wolf, Eagle
Smallest subtree	Cat, Dog, Wolf
Positive examples	Cat, Tiger, Wolf
Likely formulae	has paws; digs holes
Likely examples	Tiger, Lion, Cat, Dog, Wolf
Smallest subtree	(same)
Positive examples	Chimp, Wolf, Rhino
Likely formulae	has visible ears; has 4 legs; runs; has a nose
Likely examples	Rhino, Chimp, Gorilla, Lion, Dog, Wolf, Horse, Camel, Giraffe, Tiger, Cat, Elephant, Mouse, Squirrel, Deer, Cow, Seal
Smallest subtree	(same, except for Seal)
Intended category	mammals
Positive examples	Camel, Cow, Elephant
Likely formulae	has hooves
Likely examples	Elephant, Rhino, Horse, Cow, Camel, Giraffe, Deer
Smallest subtree	(same)
Positive examples	Ant, Bee
Likely formulae	is an insect; ¬ has bones; ¬ has red blood; flies
Likely examples	Bee, Butterfly, Ant, Cockroach
Smallest subtree	(same)
Positive examples	Chicken, Eagle
Likely formulae	is a bird, has a beak, has 2 legs
Likely examples	Robin, Eagle, Chicken, Ostrich, Finch, Penguin
Smallest subtree	(same, except for Penguin)
Intended category	birds
Positive examples	Finch, Robin
Likely formulae	sings; eats nuts
Likely examples	Robin, Finch
Smallest subtree	(same)
Intended category	birds that can fly

However, when the smallest subtree containing the positive examples is the entire of tree of animals, the model, ever faithful to the size principle, tends to come up with categories that do not match human intuitions:

Positive examples	Salmon, Wolf
Likely formulae	lives in cold climates; eats fish
Likely examples	Wolf, Salmon, Dog, Seal, Penguin, Whale, Cat, Eagle, Dolphin, Tiger, Trout, Ant
Smallest subtree	(all 33 animals)
Positive examples	Ant, Ostrich, Tiger
Likely formulae	¬ is brown; lives in hot climates; ¬ has a snout
Likely examples	(23 animals)
Smallest subtree	(all 33 animals)

[4]The smallest cut from the top of the list of the most likely formulae (in the posterior) that accounts for more than half the probability mass of the top $\phi = 10$ formulae. The formulae are listed in order of decreasing posterior probability.

[5]Those objects with generalization probability at least 0.5, in order of decreasing generalization probability.

[6]The animals in the smallest subtree (of the tree in Kemp & Tenenbaum 2008) that contains all of the positive examples.

It's not easy to come up with a set that the model generalizes to all 33 animals.

Positive examples	Bee, Giraffe, Penguin
Likely formulae	travels in groups; ¬ is smart; is black
Likely examples	(21 animals)
Smallest subtree	(all 33 animals)

Positive examples	Bee, Giraffe, Penguin, Eagle
Likely formulae	is soft; ¬ bites; ¬ digs holes
Likely examples	(31 animals)
Smallest subtree	(all 33 animals)

Positive examples	Bee, Giraffe, Penguin, Eagle, Ant
Likely formulae	¬ has a large brain; ¬ eats seeds; ¬ eats bugs; ¬ is a feline
Likely examples	(32 animals)
Smallest subtree	(all 33 animals)

Most of the formulae above only involve one feature, suggesting that our compositional concept space is overkill. Here is a failed attempt at getting the model to learn a compound rule:

Positive examples	Robin, Finch, Chicken, Ant, Bee, Butterfly, Cockroach
Likely formulae	¬ has teeth
Likely examples	Cockroach, Robin, Chicken, Bee, Butterfly, Ant, Finch, Ostrich, Eagle
Smallest subtree	the above, and Penguin, Salmon, Trout, Alligator, Iguana

And a successful attempt:

Positive examples	Salmon, Trout, Wolf, Dog
Likely formulae	is a fish ∨ is a canine; is a fish ∨ howls
Likely examples	Dog, Wolf, Salmon, Trout
Smallest subtree	(all 33 animals)

Still, the full combinatorial power of the model is hardly engaged.

If we allowed the input to be lists of observations rather than sets of objects, that is, if objects could be repeated in the input, there would be more possible experiments. For example, the repeated observation of, say, a cockroach, a salmon, and a dog as being part of a category should eventually cause the model to posit a concept containing those three animals *and nothing else*, no matter how convoluted the formula is.

References

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. *Advances in neural information processing systems*, *12*, 59–65.