# Title

## 9.66 Final Project

Leon Lin

abstract

### A grammar for concepts

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 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.

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

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

*Figure 1.* A slightly simplified 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.

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 con-

[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.

cept, 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 '$\text{Conj} \to P \wedge \text{Conj}$' and '$\text{Conj} \to \text{True}$' respectively.) If $\tau$ 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; $\beta$ 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. We use the 10 most likely formulae of the posterior.

## 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 $\tau$ 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 $P(F \mid D)$.

## Model performance on the categories of Medin and Schaffer

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 results are shown in figure 2. The fit between our model and the human data 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.

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.

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

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

*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 ($b = 1.3$). 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.