Active inference in concept learning

Jonathan D. Nelson (jnelson@cogsci.ucsd.edu)*
Joshua B. Tenenbaum (jbt@psych.stanford.edu)^
Javier R. Movellan (movellan@cogsci.ucsd.edu)*

*Cognitive Science Department, UCSD
La Jolla, CA 92093-0515

^Psychology Department, Stanford University
Stanford, CA 94305

Abstract

People are active experimenters, constantly seeking new information relevant to their goals. A reasonable approach to active information gathering is to ask questions and conduct experiments that minimize the expected state of uncertainty, or maximize the expected information gain, given current beliefs (Fedorov, 1972; MacKay, 1992; Oaksford & Chater, 1994). In this paper we present results on an exploratory experiment designed to study people's active information gathering behavior on a concept learning task. The results of the experiment suggest subjects' behavior may be explained well from the point of view of Bayesian information maximization.

Introduction

In scientific inquiry and in everyday life, people seek out information relevant to perceptual and cognitive tasks. Whether performing experiments to uncover causal relationships, saccading to informative areas of visual scenes, or turning towards a surprising sound, people actively seek out information relative to their goals.

Consider a person learning a foreign language, who notices a particular word, "tikos," used to refer to a baby moose, a baby penguin, and a baby cheetah. Based on those examples, she may attempt to discover what tikos really means. Logically, there are an infinite number of possibilities. For instance, tikos could mean *baby animals*, or simply *animals*, or even *baby animals and antique telephones*. Yet a few examples are often enough for human learners to form strong intuitions about what meanings are most likely.

Suppose the learner could point to a baby duck, an adult duck, or an antique telephone, to inquire whether that object is "tikos." What question would she ask? Why do we think that pointing to the telephone is not a good idea, even though from a logical point of view, a phone could very well be tikos? In this paper we present a normative theoretical framework, to try to predict the questions people ask in concept learning tasks (Fedorov, 1972; MacKay, 1992; Oaksford & Chater, 1994).

A Bayesian concept learning model

In the approach presented here, we evaluate questions in terms of their information value. Formally, information is defined with respect to a probability model. Here we use a Bayesian framework in the sense that we model internal beliefs as probability distributions. In order to quantify the information value (in bits) of a person's questions, we first need a model of her beliefs, and the way those beliefs are updated as new information is obtained. Tenenbaum (1999, 2000) provides such a model of people's beliefs, for a number concept learning task. While Tenenbaum (1999, 2000); and the first and last authors of the present paper, in a pilot study, found that his model described subjects' beliefs well, there were some deviations between model predictions and subjects' beliefs. The concept learning model used in the present study, which we describe below, is based on Tenenbaum's original model, but extended in ways that reduce previously observed deviations between model predictions and study participants' beliefs.

We formalize the concept learning situation described by the number concept model using standard probabilistic notation: random variables are represented with capital letters, and specific values taken by those variables are represented with small letters. The random variable C represents the correct hidden concept on a given trial. This concept is not directly observable by study participants; rather, they infer it on the basis of example numbers that are consistent with the true concept. Notation of the form *"C=c"* is shorthand for the event that the random variable C takes the specific value c, e.g. that the correct concept (or "hypothesis") is *prime numbers*. We represent the examples given to the subjects by the random vector X. The subject's beliefs about which concepts are probable prior to the presentation of any examples is represented by the probability function $P(C=c)$. The subject's updated belief about a concept's probability, after she sees the

examples X=x, is represented by *P(C=c|X=x)*. For example, if *c* is the concept *even numbers* and *x* the numbers "2, 6, 4", then *P(C=c|X=x)* represents the subject's posterior probability that the correct concept is *even numbers*, given that 2, 6, and 4 are positive examples of that concept. Study participants are not explicitly given the true hidden concept; rather, they infer it from examples of numbers that are consistent with the true concept.

The number concept model includes both *arithmetic* and *interval* concepts. Interval concepts are sets of consecutive integers between *n* and *m*, where $1 \le n \le 100$, and $n \le m \le 100$, such as *numbers between 5 and 8*, and *numbers between 10 and 35*. Thus, there are 5050 interval concepts. Arithmetic concepts include *odd numbers*, *even numbers*, *square numbers*, *cube numbers*, *prime numbers*, *multiples of n* ($3 \le n \le 12$), *powers of n* ($2 \le n \le 10$), and *numbers ending in n* ($1 \le n \le 9$). There are 33 arithmetic concepts.

Inferences are made with respect to the following model of how examples are generated: A concept is first chosen at random according to a prior probability distribution. The prior probability distribution of the model is designed to reflect the human intuition that a concept like *multiples of 10* is more plausible than a concept like *multiples of 10 except 30*. A portion of total prior probability is divided evenly into the arithmetic concepts, with the exception of *even numbers* and *odd numbers*. To reflect the higher salience of the concepts *even numbers* and *odd numbers*, each of those concepts is given five times the prior probability of the other arithmetic concepts. Among the interval concepts, prior probability is apportioned according to the Erlang distribution

$$P(H = h) \propto \frac{|h|}{\sigma^2} e^{\frac{-|h|}{\sigma}}$$

according to the concept's size $|h|$. (The concept *numbers between 15 and 30* is size 16.) Sigma gives the optimal interval length. In the simulations described in this paper we set σ to 15, although in principle, σ is a free parameter to fit to the data. Interval concepts of a given length, such as *numbers between 25 and 35*, and *numbers between 89 and 99*, receive the same prior probability, irrespective of their endpoints.

Once a concept is chosen, examples are randomly and independently generated, with equal probability, from the set of numbers in that concept. Thus, the likelihood of a particular vector of m examples X=x, given the concept h,

$$P(X = x \mid H = h) = \frac{1}{|h|^m},$$

if all m examples are in the concept *h*, and zero otherwise.

This generating assumption reflects the human intuition that although a given set of example numbers is typically compatible with more than one concept, it may be more representative of some concepts than others. For instance, although the example numbers 60, 80, 10, and 30 are compatible with both *multiples of 10* and *multiples of 5*, that set of numbers is a better example of the concept *multiples of 10* than it is of the concept *multiples of 5*, because it is much more likely to be observed as a random sample from the more specific hypothesis *multiples of 10*.

The generative model described above can be used to compute the probability that a new element *y* belongs to the hidden concept *C* given the examples in *x*:

$$P(y \in C \mid X = x) = \frac{\sum_{h: y \in h} P(X = x \mid H = h) P(H = h)}{\sum_{h} P(X = x \mid H = h) P(H = h)}$$

An ideal concept learning model would assign some prior probability to every possible concept, according to each concept's plausibility to human learners. The main difference between the concept learning model used in the current paper, and the model introduced in Tenenbaum (1999, 2000), is our inclusion of a large number of random "exception" concepts, which are formed by replicating and slightly changing, or "mutating," concepts from the basic model. Here, we include 50,830 exception hypotheses -- on average, 10 exception concepts for each concept in the basic model. To form an exception concept (or "hypothesis"), a concept is first picked from the basic model, according to the prior probability of concepts in the basic model. We include a parameter μ for the average number of changes to the original concept, and divide these changes equally, on average, into additions of new numbers and exclusions of existing numbers. The probability of each existing number being excluded from a concept is $\frac{\mu}{2|h|}$, and the probability of each currently excluded number (between 1 and 100) being added to the concept is $\frac{\mu}{2(100 - |h|)}$.

Each exception hypothesis receives a constant share of the total proportion of prior probability assigned to the exception hypotheses. In the simulation of the model reported in this paper, 60% of prior probability

was assigned to the exception hypotheses, and μ was set to 6. It takes approximately 30 minutes to simulate the set of trials in the study, for any setting of model parameters, and we are just beginning to explore the parameter space. Early exploration suggests that a wide range of parameters in the extended number concept model can improve on the basic model's correspondence to human beliefs.

Information-maximizing sampling

In the experiment reported in this paper, we allowed subjects to actively ask questions about number concepts, instead of making inferences solely on the basis of the examples given to them. For example, on one trial the subject was given the number 16 as an example of the hidden underlying concept, and then was allowed to test another number, to find out whether it was also consistent with the true, hidden concept.

In our formalism, the binary random variable Y_n represents whether the number n is a member of the correct concept. For example, $Y_8=1$ represents the event that 8 is an element of the correct, hidden concept, and $Y_8=0$ the event that 8 is not in that concept. Asking "is the number n an element of the concept?" is equivalent to finding the value taken by the random variable Y_n, in our formalism.

We evaluate how good a question is in terms of the information about the correct concept expected for that question, given the example vector $X=x$. The expected information gain for the question "Is the number n an element of the concept?" is calculated with respect to the learner's beliefs, as approximated with the extended number concept model described above. Formally, expected information gain is given by the following formula:

$$I(C, Y_n \mid X = x) = H(C \mid X = x) - H(C \mid Y_n, X = x),$$

where the uncertainty (entropy) about the hidden concept C given the example numbers in x,

$$H(C \mid X = x) = -\sum_c P(C = c \mid X = x) \log_2 P(C = c \mid X = x),$$

and the expected remaining uncertainty about the hidden concept C, given the example numbers in x and the answer to the question Y_n:

$$H(C \mid Y_n, X = x) = -\sum_{v=0}^{1} P(Y_n = v \mid X = x) \sum_c P(C = c \mid Y_n = v, X = x) \log_2 P(C = c \mid Y_n = v, X = x)$$

We consider only binary questions, of the form "is n consistent with the concept?" so the maximum information value of any question in our experiment is one bit. Note how information value of questions is relative to subjects' internal beliefs, which we approximate here by using the expanded number concept learning model. An information-maximizing strategy prescribes asking the question with the highest expected information gain, e.g., the question that minimizes the expected entropy, over all concepts.

Another strategy of interest is confirmatory sampling, which consists of asking questions whose answers are most likely to confirm current beliefs. In other domains it has been proposed that people have a bias to use confirmatory strategies, regardless of their information value (Klayman & Ha, 1987; Popper, 1959; Wason, 1960).

The active sampling concept game

Twenty-nine undergraduate students, recruited from Cognitive Science Department classes at the University of California, San Diego, participated in the experiment. Subjects gave informed consent, and received either partial course credit for required study participation, or extra course credit, for their participation. The experiment began with the following instructions:

> Often it is possible to have a good idea about the state of the world, without completely knowing it. People often learn from examples, and this study explores how people do so. In this experiment, you will be given examples of a hidden number rule. These examples will be randomly chosen from the numbers between 1 and 100 that follow the rule. The true rule will remain hidden, however. Then you will be able to test an additional number, to see if it follows that same hidden rule. Finally, you will be asked to give your best estimation of what the true hidden rule is, and the chances that you are right. For instance, if the true hidden rule were "*multiples of 11*," you might see the examples 22 and 66. If you thought the rule were "*multiples of 11*," but also possibly "*even numbers*," you could test a number of your choice, between 1-100, to see if it also follows the rule.

On each trial subjects first saw a set of examples from the correct concept. For instance, if the concept were even numbers, subjects might see the numbers "2, 6, 4" as examples. Subjects were then given the opportunity to test a number of their choice. Subjects were given feedback on whether the number they tested was an element of the correct concept.

We wrote a computer program to simulate the expanded number concept model, and to compute the information value of each possible question, given each set of examples. By considering beliefs and questions together, we may evaluate the information value of participants' questions, as well as that of information-maximizing and confirmatory sampling strategies. We define the confirmatory strategy as testing the number (excluding the examples) that has the highest posterior probability, as given by the extended number concept

model, of being consistent with the correct hidden concept.

Results

We discuss two types of trials, grouped according to the posterior beliefs of the extended number concept model, after all the example numbers have been seen. These results should be considered preliminary, as 29 data points on each trial are not sufficient for estimation of statistically reliable sampling distributions over the range of possible queries from 1 to 100.

Arithmetic trials

On some trials, the model is dominated by arithmetic concepts, and exception hypotheses based on arithmetic concepts. On each of these trials, good agreement between a number's information value and subjects' propensity to sample that number was observed. The information value of the confirmatory strategy was near to that of the information-maximizing strategy on these trials.

Consider the trial with the examples 81, 25, 4, and 36, in which the concept with the highest posterior probability is *square numbers*. Generalization behavior of the model, and beliefs of subjects, are shown in Figure 1. Note that the model and subjects alike assign certain, or near certain, probability to each of the example numbers, but less than certain probability to the other square numbers. Relative to the model's beliefs, the most informative numbers to test are non-example square numbers, such as 9, 16, 49, 64, or 100 (Figure 2). In fact, 20 of 29 subjects tested one of these numbers. Other subjects' samples do not show a clear pattern, except for testing the number 10 (5 of 29 subjects), which is unpredicted.

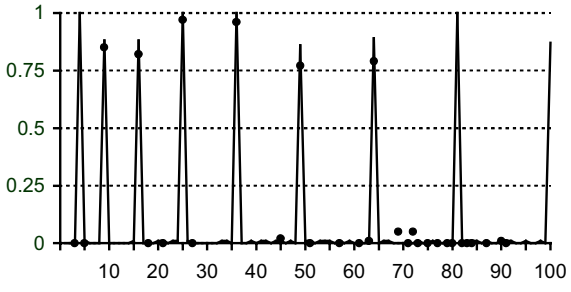

Figure 1. Generalization probabilities, given the examples 81, 25, 4, and 36. Model probabilities are given by the line. Subjects' probabilities, for the 30 probe numbers subjects rated, are given with circles.

Good agreement between subjects' samples and rated information value is also observed on the trial with the examples 16, 8, 2, and 64. The most informative numbers to test are non-example powers of two, 4 or 32. Most (16/29) subjects tested these numbers.

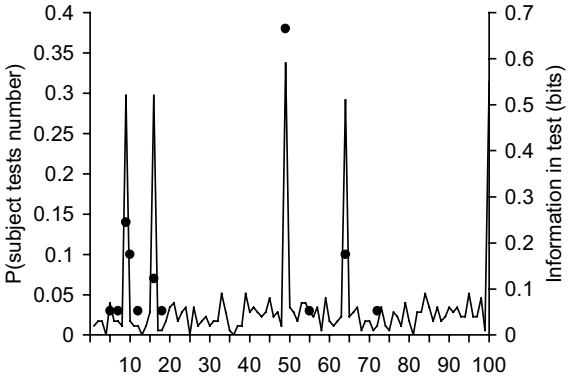

Figure 2. Information value of questions (line), and subjects' questions (circles), given the examples 81, 25, 4, and 36.

Finally, we may consider the trial with the examples 60, 80, 10, and 30, in which the hypothesis *multiples of 10* receives the highest posterior probability; multiples of 5 also receive moderate probability (Figure 4). On this trial, non-example multiples of 10, such as 20, and odd multiples of five, have the highest information value. Multiples of 10 were tested by 21 of 29 subjects; an additional 5 subjects tested odd multiples of five

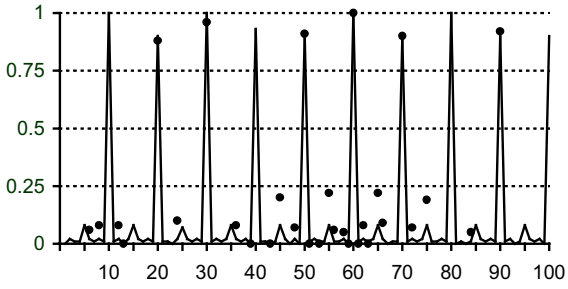

Figure 3. Generalization probabilities given the examples 60, 80, 10, and 30.

The difference between the first two arithmetic trials, and the trial with the examples 60, 80, 10, and 30 appears to be that a clear alternate hypothesis -- multiples of five -- receives moderate posterior probability in the multiples of 10 trial, but not on the other trials.

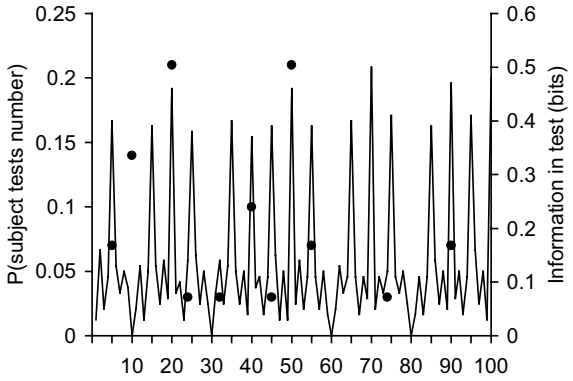

Figure 4. Information value of questions, and subjects' questions, given the examples 60, 80, 10, and 30.

Interval trials

On these trials, several examples of numbers of similar magnitude, such as 16, 23, 19, and 20, are given (these numbers are points where model probabilities are 1.00, Figure 5, and Figure 7). The model is certain that the example numbers themselves are consistent with the true concept. The model is fairly sure that non-example numbers within the range spanned by the examples, like 17, 18, 21, and 22, are consistent with the true concept. Finally, the model assigns decreasing probability to numbers as they move away from the range of observed examples (Figure 5).

It should be noted that there is some variability from one run of the model to the next. The general pattern of results, however, holds from run to run. In particular, (1) numbers slightly outside of the range of the observed examples are most informative, (2) information value of numbers decreases with increasing distance from the observed examples, and (3) there is moderate information value in non-example numbers within the range of observed examples.

Most subjects tested numbers outside of, but near the observed examples (Figure 6). About one-third of subjects tested (non-example) numbers within the range spanned by the examples. On the other interval trials -- with example numbers 60, 51, 57, and 55; and 81, 98, 96, 93 (illustrated in Figure 7 and Figure 8)-- similar patterns emerged.

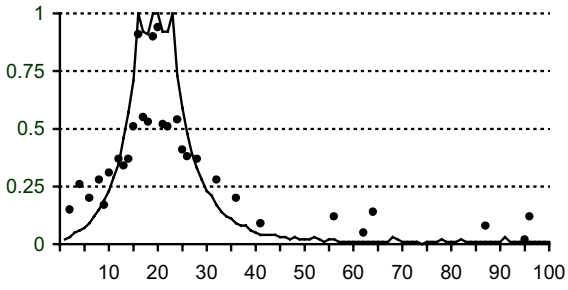

Figure 5. Generalization probabilities, given the examples 16, 23, 19, and 20.

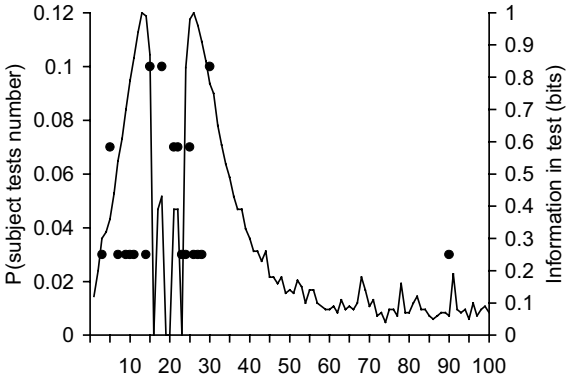

Figure 6. Information value of questions, and subjects' questions, given the examples 16, 23, 19, and 20.

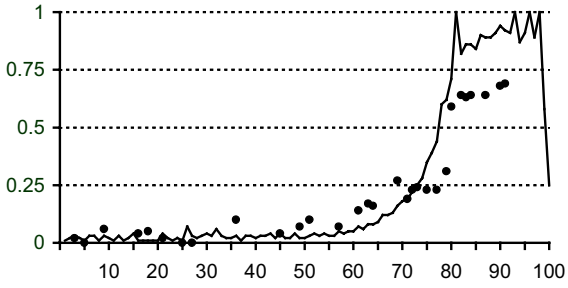

Figure 7. Generalization probabilities, given the examples 81, 98, 96, and 93.

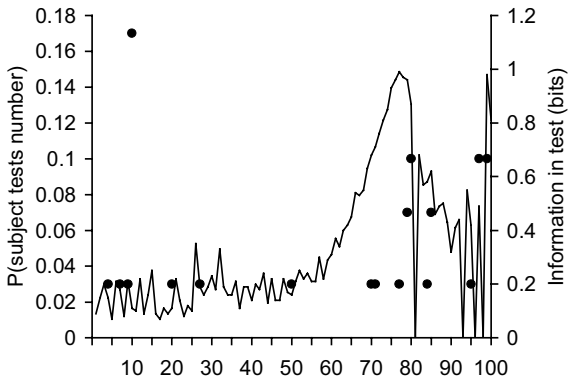

Figure 8. Information value of questions, and subjects' questions, given the examples 81, 98, 96, and 93.

Discussion

This paper presents work in progress to analyze active inference in concept learning from the point of view of the rational, probabilistic approach to cognition (Anderson, 1990). In the rational study of information-gathering behavior, the current research adds to existing analyses of Wason's (1966, 1968) selection task (Oaksford & Chater, 1994, 1998), and Wason's (1960) 2-4-6 task (Ginzburg & Sejnowski, 1996).

We found that a normatively inspired criterion of optimal sampling -- maximizing average information gain -- predicts human behavior well on a relatively unconstrained task. This result is strengthened by the fact that the extended number concept model we employed, as a proxy for subjects beliefs, was not originally developed with the goal of serving as a model for sampling. Nor were our extensions to it ad hoc. To the contrary, our extended model now has a better fit to data from earlier studies.

If rational theories of cognition are to explain thought and behavior in natural environments, then optimal sampling agents should also exhibit the systematic "biases" traditionally associated with human behavior. Indeed, we found that on many trials, a confirmatory sampling strategy approximates the information-maximizing strategy.

A final point is that whereas information gain, calculated with respect to the extended number concept model, predicts study participants' questions fairly well, information gain with respect to the original number concept model does not do so. This illustrates that particular queries are not informative or uninformative on their own, but only in relation to a particular probability model. To understand people's questions, or build artificial sampling systems that come closer to meeting human competence, developing appropriate probability models is critical.

Acknowledgments

Thanks to Gedeon Deák, Jeff Elman, Iris Ginzburg, Craig McKenzie, Terry Sejnowski, and three anonymous reviewers for their ideas; and Kent Wu, Dan Bauer and Jonathan Weh for their help in this research. J. Nelson was partially supported by a Pew graduate fellowship during this research.

References

Anderson, J. R. (1990). *The adaptive character of thought*. New Jersey: Erlbaum.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Ginzburg, I.; Sejnowski, T. J. (1996). Dynamics of rule induction by making queries: transition between strategies. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 121-125.

Klayman, J.; Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211-228.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 590-604.

Oaksford, M.; Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.

Oaksford, M.; Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. UK: Erlbaum

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Tenenbaum, J. B. (1999). *A Bayesian Framework for Concept Learning.* Ph.D. Thesis, MIT

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In *Advances in Neural Information Processing Systems*, 12, Solla, S. A., Leen, T. K., Mueller, K.-R. (eds.), 59-65.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*. 12, 129-140.

Wason, PC (1966). Reasoning. In Foss, B (ed.), *New Horizons in Psychology*, pp. 135-151.

Wason, PC (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.