Learning biases for teaching boolean concepts

Nick Searcy (nick.searcy@louisville.edu)

Patrick Shafto (p.shafto@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292 USA

Abstract

According to previous accounts, teaching is the helpful sampling of examples according to a learner's known biases. Using the domain of Boolean concepts, we show that biases are necessary, there is no single rational bias, and teaching is not possible when the teacher does not know the learner's bias. Taken together, these results suggest that teaching via sampling would be either ineffective or impossible for Boolean concepts. We offer an alternative account of teaching based on cooperation and the teacher's omission of irrelevant features. The result is a model of teaching that is computationally efficient, effective in concept spaces with infinitely many features, and suggestive of a natural concept representation based on cooperation.

Keywords: Concept learning; representation; teaching dimension; relevant features

Introduction

Learning from a cooperative source offers significant advantages over learning from other sources (e.g. disinterested, random, or adversarial) (Shafto & Goodman, 2008; Csibra & Gergely, 2009). Previously, cognitive science and machine learning researchers have used a sampling account to explain the advantage of teaching (Xu & Tenenbaum, 2007; Shafto & Goodman, 2008; Balbach, 2008; Zilles, Lange, Holte, & Zinkevich, 2009): the teacher selects examples to give to the learner in order to maximize the probability of learning the correct concept, given limitations on the number of examples. These approaches assume that learners have a known bias—that, a priori, learners are more inclined toward some concepts than others (e.g for 'red' over 'red or square' when both are consistent with the evidence).

In this paper, we analyze the role of a prior bias in the domain of Boolean concepts (Shepard, Hovland, & Jenkins, 1961) and show three results. First, without a bias, sampling-based teaching requires the observation of each and every example. Second, there are many optimal biases, and hence no single rational choice. Third, as the number of features grows, teaching is not possible if the teacher and learner have a different bias. These results indicate that sampling alone is an incomplete account of the effectiveness of teaching.

We offer a novel account of teaching Boolean concepts called cooperative inference that uses a notion of cooperation (such as the idea that a teacher may omit unnecessary features from examples) rather than biases to explain the effectiveness of teaching. We show that this method allows teaching without prior communication of the learner's bias. For the cooperative inference model, the difficulty of teaching depends on the complexity of the target concept irrespective of the concept space, and thus permits teaching in concept spaces with infinitely many features. The model also indicates a natural representation for teaching—Disjunctive Normal Form—and the resulting learning bias is consistent with the best models of human complexity of learning (Feldman, 2000; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Goodwin & P. Johnson-Laird, 2011).

The limitations of teaching by sampling

To analyze sampling-based teaching, we use an extensively studied model of algorithmic teaching: the *teaching set* model (Shinohara & Miyano, 1991; Anthony, Brightwell, Cohen, & Shawe-Taylor, 1992; Goldman & Mathias, 1993). The paper will use the following notation for Boolean features, instances, and concepts.

Definition 1 (Preliminaries) Let $\mathcal{F} = \{f_0, f_1, \dots\}$ be the *feature space*. Let the *instance space* be the function space from $\mathcal{F}$ to Boolean labels $\mathcal{X} = \{0,1\}^{\mathcal{F}}$. And let the *concept space* be the function space from $\mathcal{X}$ to the Boolean labels $\mathcal{C} = \{0,1\}^{\mathcal{X}}$. A concept class is some subset of the concept space $C \subseteq \mathcal{C}$.

Ordered pairs formed of features and Boolean labels such as those found in an instance $(f,b) \in x$ are referred to as *specifications*. Ordered pairs between instances and Boolean labels such as those found in concepts $(x,b) \in c$ are referred to as *examples*. Finally, let the *sample space* be any set of examples that can be found in a concept, $\mathcal{S} = \{s \subseteq c \mid c \in \mathcal{C}\}$. For abbreviation, samples and concepts may be represented as strings over $\{0,1,*\}$ such that $(x_i, A[i]) \in s$ for all $A[i] \neq *$.

A concept is *consistent* with a sample if each example in the sample is also in the concept.

$$\text{Cons}(s,C) = \{c \in C \mid s \subseteq c\} \tag{1}$$

For an intuitive illustration of these definitions, consider the two features 'red' and 'square'. These two features can be combined to form four instances: 'red and square', 'not red and square', 'red and not square', 'not red and not square'. A concept can be thought of as a definition for an unknown category such as 'fep'. The four instances allow for 16 concepts from 'feps are red' to 'feps are either not red and not square or red and square' to 'feps are nothing' (i.e. all four instances are false). A teacher samples evidence in the form of labeled examples to teach a concept; in order to teach the concept 'feps are red' a teacher might use the following examples 'red and square is a fep, not red and square is not a fep'.

Within this framework, a learner is both *consistent* and *class-preserving*. Consistent means that learners will only

learn a concept that is consistent with the teaching sample. Class-preserving means that learners will only learn a concept that is in the concept class. Equivalently, the learner can be thought of as beginning with all concepts in the concept class and ruling out concepts that are inconsistent with the sample.

Assuming that the teacher knows the concept class of the learner, the teaching set is the sample with the fewest examples that will teach the target concept for any consistent, class-preserving learner—i.e. that will rule out all other concepts in the concept class.

Definition 2 (Teaching Set) The teaching set of a target concept is the minimal sample that teaches the target concept to all consistent and class-preserving learners.

$$\mathrm{TS}(c,C) = \arg\min_{s\in\mathcal{S}}\{|s| \mid \mathrm{Cons}(C,s) = \{c\}\} \qquad (2)$$

The teaching dimension for a target concept is the size of the teaching set.

One learner's concept class might include only three concepts, 'feps are red', 'feps are not square' and 'feps are not red and not square'. Then, a teacher could teach 'feps are red' with a single labeled example, 'red and square is a fep', because that example rules out the other two concepts. However, the teacher would require more than one example to teach 'feps are not square' because none of the examples in that concept rule out both other concepts.

Ineffective without a bias

To this point, models of teaching have assumed various learning biases. Consider learning from an unknown source where, without a bias, all consistent concepts are equally likely and thus there can be no learning (Watanabe, 1969). We show that without a bias, there can also be no sampling-based teaching.

Theorem 1 (Teaching without bias) *Given a set of n features $\mathcal{F}_n$ and a concept class, $\mathcal{X}_n = \{0,1\}^{\mathcal{F}_n}$, $C_n = \{0,1\}^{\mathcal{X}_n}$. The teaching set for any $c \in C_n$ must include every instance in $\mathcal{X}_n$ labeled according to the concept such that* $\mathrm{TS}(c,C_n) = c$.

PROOF Let s, the teaching set for c, label all but m instances in c, such that $|c \setminus s| = m$.[1]

The set of unlabeled instances can be used to form a set of concepts $X^* = \{x \mid (x,b) \in c \setminus s\}$, $C^* = \{0,1\}^{X^*}$. These concepts may be used to construct $\mathrm{Cons}(s,C_n)$, the set of all concepts in C_n consistent with the s.

$$\mathrm{Cons}(s,C_n) = \left\{c = s \cup c' \text{ for all } c' \in \{0,1\}^{X^*}\right\} \qquad (3)$$

It follows that $\mathrm{Cons}(s,C_n) = \{c\}$ iff $X_n = \emptyset$—i.e. $m = 0$ and no instances are left unlabeled—and thus $\mathrm{TS}(c,\mathcal{C}) = c$. ■

Without a bias, a teacher must label *every* instance in order to teach the target concept—teaching is no better than other sampling methods, all of which trivially teach a concept when allowed to sample the entire example space.

[1] $A \setminus B$ is the set difference such that $A \setminus B = \{x \in A \mid x \notin B\}$.

Many optimal biases

If a learner needs a bias in order for teaching to be effective, the most sensible choice for the bias is the one that would, on average, lead to the most efficient teaching.

We will consider two types of biases: The first, *ordered bias*, adopts a total pre-order[2] on the concepts in $\mathcal{C}$ such that lower ordered concepts are given priority. One example of this is the Occam's Razor bias (Balbach, 2008; see also Goodman et al., 2008) where concepts are judged by the number of terms (separated by 'or's) in the description, e.g. 'feps are red and square' which has one term would be given priority over 'feps are red or square' which has two. The second, *functional bias*, permits a redefinition of the set of concepts that are consistent with any sample. This bias is more complex but also opens up many more possibilities. An example of this is the Subset Teaching Set (Zilles et al., 2009) where a learner rules out concepts based on a prediction of the teaching set rather than based on whether or not an example is consistent (see Shafto & Goodman, 2008, for a probabilistic example of such a bias). For both kinds of bias, we show that there are many optimal biases and that the selection of a bias through prior communication would allow arbitrarily efficient teaching—effectively equivalent to telepathy.

For an ordered bias, the learner learns the lowest-order consistent concept. Thus, an ordered bias amounts to a modification of the Cons() function. Given c and $\preceq$, we use $C_{\preceq c} = \{c' \in \mathcal{C} \mid c' \preceq c\}$ to refer to the set of concepts of the same or lower order as c and

$$\mathrm{Cons}(s,\mathcal{C},\preceq) = \{c' \in C_{\preceq c} \mid \text{ for all } c \supseteq s\} \text{ and} \qquad (4)$$

$$\mathrm{TS}(c,\mathcal{C},\preceq) = \arg\min_{s\in\mathcal{S}}\{|s| \mid \mathrm{Cons}(s,\mathcal{C},\preceq) = \{c\}\}\,. \qquad (5)$$

Next, we develop a novel ordered bias called the *Hamming distance bias* and show that it minimizes the average teaching dimension and is thus optimal. Informally, the proof is as follows. Each ordered bias includes a least-element concept that is of equal or lower order to all other concepts in the concept class. Any ordered bias would, at minimum, require the teaching set of a target concept to include examples sufficient to rule out the least-element concept (because it is necessarily in the set of concepts of lower order than the target concept). For the Hamming distance bias, we show that the teaching set of each concept is the minimal set of examples sufficient to rule out the least-element concept and therefore it is a minimal ordered bias.

Theorem 2 (Hamming distance bias) *Let $h(c_1,c_2)$ be the Hamming distance between c_1 and c_2 such that $h(c_1,c_2) = |c_1 \setminus c_2|$. Given an origin concept, c^*, the Hamming distance bias is $\preceq_{h(c^*)} = \{(c_1,c_2) \in \mathcal{C}\times\mathcal{C} \mid h(c^*,c_1) \le h(c^*,c_2)\}$ and is an optimal ordered bias.*

[2] A total-order $\preccurlyeq_t$ on a set X is a partial order such that any two elements in X are comparable (i.e. for all $a,b \in X$ either $a \preccurlyeq_t b$ or $b \preccurlyeq_t a$). A *pre-order* $\preccurlyeq_p$ on a set X is a total order that is both reflexive ($a \preccurlyeq_p a$) and transitive ($a \preccurlyeq_p b$ and $b \preccurlyeq_p c$ implies $a \preccurlyeq_p c$) for all $a,b,c \in X$.

PROOF Let $\preceq_{c_0}$ be an ordered bias with c_0 as a least element, i.e. $c_0 \preceq_{c_0} c$ for all $c \in \mathcal{C}$. Then c_0 is in $C_{\preceq_{c_0} c}$ for all $c \in \mathcal{C}$ and in order to rule out c_0, the teaching set for any c must include $c \setminus c_0$,

$$\mathrm{TS}(c, \mathcal{C}, \preceq_{c_0}) \supseteq c \setminus c_0 \ . \tag{6}$$

For $\preceq_{h(c^*)}$ and any $c_1, c_2 \in \mathcal{C}$, $c_1 \neq c_2$ if $c_1 \preceq_{h(c^*)} c_2$ (i.e. $|c_1 \setminus c^*| \leq |c_2 \setminus c^*|$) then $(c_2 \setminus c^*) \not\subseteq c_1$. So $c_2 \setminus c^*$ is sufficient to rule out any concept of a lesser order, i.e.

$$\mathrm{Cons}(c_2 \setminus c^*, \{c_1, c_2\}, \preceq_{h(c^*)}) = \{c_2\} \tag{7}$$

$$\mathrm{TS}(c, \mathcal{C}, \preceq_{h(c^*)}) = c \setminus c^* \ . \tag{8}$$

From eq. (6), the Hamming distance bias results in the minimal size teaching set for each concept and is thus optimal. ■

For example, if the origin concept is the concept with all negative examples, $c^* = 000\ldots$, the teaching set for each target concept c will include only the examples that differ from c to c^*, or, in other words, the set of positive examples in c. Thus the average teaching dimension for the Hamming distance bias is the average number of positive examples in each concept, or $\frac{2^n}{2} = 2^{n-1}$ for a concept space with n features.

The Hamming distance bias is optimal without regard to the origin concept, so the average teaching dimension would not change if the origin concept were changed. What does change, however, is the number of examples required to teach particular concepts—especially the origin concept which can be taught with an empty teaching sample.

Rational selection of a functional bias is similar to that of an ordered bias, with the exception that, because a functional bias may modify which concepts are consistent with which samples in any way, there is no analogous constraint to eq. (6). Thus, a functional bias may use any set of examples to teach a target concept so long as each concept is taught with a different set of examples.

The minimal functional bias would assign a concept to each unique $s \in \mathcal{S}$, beginning with the smallest. The first concept would be taught with $s = \emptyset$, the following concepts would be taught with samples $|s| = 1$, and so on. For $|s| = i$, and the number of features, n, the number of unique samples in $\mathcal{S}$ is $\binom{2^n}{i} 2^i$. It is outside of the scope of this paper to determine the average teaching dimension for the optimal functional bias, though it is clearly smaller than that of the optimal ordered bias.

Both the optimal ordered and functional biases have a free parameter in the choice of an origin concept. So there are at least as many distinct optimal choices as there are concepts and for each bias there is at least one concept that can be taught with an empty teaching sample.

Impossible for unknown bias

First, the different roles of the two biases should be clarified. For the following analysis, we will use classes $C_t, C_l \subseteq \mathcal{C}$ to stand in for more complex biases without a loss of generality (see eq. (4)). The teacher's class C_t is used to determine a teaching set that is *consistent* only with the target concept, $s_t = \mathrm{TS}(c, C_t)$, while the learner's class C_l is the class that is *preserved* when the learner uses the sample to rule out other concepts, $C' = \mathrm{Cons}(\mathrm{TS}(c, C_t), C_l)$. Given C_t and C_l, we say that a concept c is *teachable* iff $\{c\} = \mathrm{Cons}(\mathrm{TS}(c, C_t), C_l)$ and we use the following indicator function such that $\mathrm{Teach}(c) =$ True if the concept c is teachable and False otherwise.

To begin with, concepts not in either C_t or C_l are trivially unteachable. Of the concepts in the intersection of the teacher and learner's classes $c \in C_t \cap C_l$, a concept is teachable iff each example in the teaching set from the learner's perspective $\mathrm{TS}(c, C_l)$ is included in the teaching set from the teacher's perspective $\mathrm{TS}(c, C_t)$. That is, each example in $\mathrm{TS}(c, C_l)$ represents a necessary condition for the teachability of c.

For example, if $C_t = \{00, 01, 10\}$ and $C_l = \{00, 01, 11\}$, then the teaching set from the teacher's perspective includes both examples $\mathrm{TS}(00, C_t) = 00$. From the learner's perspective, 00 includes the teaching set $\mathrm{TS}(00, C_l) = *0$, so 00 is teachable. On the other hand, 01 is not teachable because the teaching set from the learner's perspective requires both examples while the teacher's includes only one, $\mathrm{TS}(01, C_t) = *1$.

The teaching set from the teacher's perspective includes an example only when it is necessary to rule out a concept that hasn't already been ruled out by other examples in the teaching set. We refer to such concepts as *adjacent* concepts. Given a sample, we say that a concept is adjacent along an example when the concept would be ruled out if the example is included in the teaching set but not otherwise. To illustrate, imagine that the teaching set for a target concept from the learner's perspective is $\mathrm{TS}(c, C_l) = 001*$. The teacher's class must include at least one concept adjacent to the third example (either 0000 or 0001) otherwise $\mathrm{TS}(c, C_t)$ will not include that example (e.g. $\mathrm{TS}(c, C_t) = 00**$).

To determine the probability that a concept is teachable we model a process where the teacher's class is determined by randomly drawing without replacement from the set of concepts. Then, for each example, the hypergeometric distribution gives the probability that all adjacent concepts are removed. Let $U = |\mathcal{C}|$ be the size of the universe of concepts and R be the number of concepts removed from $\mathcal{C}$ to get the teacher's class of size $T = |C_t|$ such that $U = R + T$. Given an example in the teaching set from the learner's perspective, let A_e be the number of adjacent concepts

$$P(\mathrm{Teach}(c) = \mathrm{False} \mid R, C_l, e) = \frac{\binom{A_e}{A_e}\binom{U-A_e}{R-A_e}}{\binom{U}{R}} \ . \tag{9}$$

As the number of features becomes very large, $n \to \infty$, the number of concepts U does as well. If R remains constant, then $\lim_{n\to\infty} P(\mathrm{Teach}(c) = \mathrm{False}) = 0$ meaning that a target concept will be teachable in the limit. But, because $U = R + T$, a constant R would mean that the teacher's concept class increases superexponentially, i.e. $O(T) = 2^{2^n}$, and this implies an implausible lack of constraints on the size of a concept space. If T increases less than superexponentially

such that R approaches U in the limit $R = U - T \to U$, the target concept will not be teachable.

Stirling's approximation of the factorial—appropriate because both U and R are very large—provides a simplification of the hypergeometric distribution in eq. (9) for $n \to \infty$.

$$\frac{\binom{A_e}{A_e}\binom{U-A_e}{R-A_e}}{\binom{U}{R}} \sim \frac{(U-A_e)^U R^U}{U^U (R-A_e)^U} \tag{10}$$

Then, because $R \to U$, $\lim_{n\to\infty} P(\text{Teach}(c) = \text{False}) = 1$.

This result depends on a set of reasonable assumptions: that the set of features is large, that the set of concepts considered by the teacher and learner is much smaller than the set of all concepts, and that the teacher cannot predict which concepts are in the learner's concept class. Given these assumptions, the likelihood that a concept could successfully be taught via sampling approaches zero. The fact that more complex biases, such as ordered, reduce to a concept class (see eq. (4)) means that this result applies to all such biases.

Cooperative Inference

Previous accounts of teaching leverage the idea that teachers purposefully choose which examples to provide to the learner but the above resuls show that this is an incomplete account of the effectiveness of teaching. Such an account cannot explain the effectiveness of teaching in realistic situations, such as when the bias of the learner is unknown or the set of features is large.

In what follows, we propose a solution. Building on previous approaches that analyze the consequences of the teacher's ability to select examples, we analyze the consequences of the teacher's ability to select features. When a teacher communicates an example, the teacher may omit unnecessary features, resulting in what we call a partial example. We show that this extension leads to a number of important consequences: teachers may successfully teach when ignoring all irrelevant features, the complexity of teaching depends on the complexity of the target concept irrespective of the concept space, and there is a natural representation for teaching.

Let a partial instance be any subset of a typical instance $\mathcal{X}' = \{x' \subseteq x \mid x \in \mathcal{X}\}$. A partial *example* is a partial instance paired with a Boolean label.

To understand how we will use partial examples, recognize that the relationship between partial instances and typical instances is analogous to the relationship between samples and concepts; partial instances are a subset of a typical instance and teaching samples are a subset of concepts. In both cases, inference is needed to relate the incomplete version to the complete version. The following application of partial examples takes advantage of the fact that the teacher and learner are mutually cooperative.

First consider the set of instances that may be consistent with a particular partial example

$$\text{Cons}(x', \mathcal{X}) = \{x \mid x' \subseteq x\} \; . \tag{11}$$

A partial example (x', b), may match some subset of consistent instances $\text{Cons}(x', X)$. Matching instances are given the label b, allowing a partial example to effectively stand in for one or more typical examples. Based on cooperation, a learner infers that all consistent examples should match a partial example[3]

$$\text{Match}(s, C) = \bigcup_{(x',b)\in s} \text{Cons}(x', \mathcal{X}) \times b \; . \tag{12}$$

Imagine the addition of a third feature to our intuitive example, so that we have 'red', 'square', and 'small'. For concepts such as 'feps are red or small and square', a teacher would likely wish to use all three features, but for others, such as 'feps are red and square', a teacher might find it helpful not to include the feature 'small'. Equation (12) represents our interpretation of cooperation for this case; when a cooperative source omits features from examples, the learner assumes that the partial example matches all consistent instances. So, 'red and square is a fep' would imply that 'red and square and small is a fep' as well as 'red and square and not small is a fep'.

This use of partial examples allows for a powerful improvement in the efficiency of teaching. First, we show that with cooperative inference a teacher only needs to use features known to be relevant in order to teach a concept. A feature is *relevant* if, for at least one pair of differently-labeled instances in the concept, the feature is the only feature to change[4]

$$\text{Relevant}(c) = \{f \mid (x_0, 0), (x_1, 1) \in c\} \tag{13}$$

$$\text{where } \; x_0 \triangle x_1 = \{(f, 0), (f, 1)\} \; . \tag{14}$$

Theorem 3 (Relevant instance space) *Given a concept c let F_R be a subset of $\mathcal{F}$ that contains all features that are relevant with respect to c and X_R be the set of partial instances formed of F_R, $X_R = \{0,1\}^{\mathcal{F}_R}$. Using cooperative inference, a teacher may successfully teach c by labeling each partial instance $x' \in X_R$ according to any consistent full instance, $x \supseteq x'$, $x \in \mathcal{X}$.*

PROOF The theorem follows immediately from the definitions.

Given an example in the teaching set, $(x', b) \in s$, $x' \in X_R$, consider the set of matching examples $s_{x'} = \{(x, b) \mid x \in \text{Cons}(x', \mathcal{X})\}$. Note that each example in $s_{x'}$ is a superset of x' and so must have the same label for each relevant feature. Assume for the sake of contradiction that two examples in $s_{x'}$ are differently labeled, $(x, 0), (x^*, 1) \in s_{x'}$. We may build a series of examples $x_0, x_1, \ldots x_i$ beginning with $x_0 = x$, and for each step, changing the label for one feature in x_0 to match x^* such that $i = \frac{1}{2}|x \triangle x_i|$.

[3] Others (e.g. Csibra & Gergely, 2009) have informally proposed a similar idea, that helpful teachers offer generic or semantically generalizeable examples.

[4] Here, $\triangle$ refers to the symmetric difference such that $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

Because, x_0 and x_n are differently labeled, there must exist some i such that $(x_i, b), (x_{i+1}, \overline{b}) \in c$. This implies that the feature f such that $x_i \triangle x_{i+1} = \{(f,0)(f,1)\}$ is relevant and is a contradiction. ■

So, a teacher who knows that only 'red' and 'square' are relevant to the concept 'fep', may omit the entire universe of irrelevant features.

Until this point, our formal discussion of concepts used an extensional sense, where a concept is defined by the set of outputs for all inputs. The alternative, intensional sense, is to represent concepts as a rule that generates the appropriate output label based on the content of input and is more similar to our intuitive discussion. For example, the concept with intension $c = f_0 \vee f_1$ (i.e. 'feps are red or square') has the following extension over two features

$$f_0 \vee f_1 = \left\{ \begin{array}{l} (\{(f_0,0),(f_1,0)\},0) \\ (\{(f_0,0),(f_1,1)\},1) \\ (\{(f_0,1),(f_1,0)\},1) \\ (\{(f_0,1),(f_1,1)\},1) \end{array} \right\} . \tag{15}$$

The intensional concept $f_0 \vee f_1$ will label instances outside of eq. (15), whereas the extensional concept is only defined for the included examples. For $f_0 \vee f_1$, the instance $\{(f_0,0),(f_1,1),(f_2,0)\}$ would be labeled positive and the instance $\{(f_0,0),(f_1,0),(f_2,0)\}$ would be labeled negative while both would be undefined for the extensional definition in eq. (15).

Given a set of examples, an intensional representation can be derived and we briefly describe this process.

Definition 3 (Intensional form) A *literal* is a negated or unnegated variable, e.g. $l_1 = f$ is true for $(f,1)$ and $l_2 = \overline{f}$ is true for $(f,0)$. A *term* is a conjunction (i.e. 'and') of literals such as $t = l_0 \wedge l_1 \wedge \ldots$ and is false unless *all* of the literals are true. A *clause* is a disjunction (i.e. 'or') of literals such as $cl = l_0 \vee l_1 \vee \ldots$ and is true unless all of the literals are false. Each term and clause is associated with a set of literals such that $t = \bigwedge_{l \in L_t} l$ and $cl = \bigvee_{l \in L_{cl}} l$. A concept is in *Disjunctive Normal Form (DNF)* if it is a disjunction of terms such as $(l \wedge l \wedge \ldots) \vee (l \wedge l \wedge \ldots) \vee \ldots$ and a concept is in *Conjunctive Normal Form (CNF)* if it is a conjunction of clauses $(l \vee l \vee \ldots) \wedge (l \vee l \vee \ldots) \wedge \ldots$

This allows us to derive an intensional concept from an extensional definition. The first method is to collect all of the positive examples and from each, form a term that is true when the specifications for that example are true and false otherwise. At this point, each positive example has an analogous term that evaluates to true for all instances in positive examples. To form an intensional concept, these terms then need to be 'or-ed' together. The resulting concept would be one that is true for all instances in positive examples and false otherwise.

$$\text{ext2int}^{+}(c) = \bigvee_{(x,1) \in c} \left(\bigwedge_{(f,1) \in x} f \quad \wedge \bigwedge_{(f,0) \in x} \overline{f} \right) \tag{16}$$

Through a similar process, the negative examples can be 'and-ed' together to form a concept that is negative for all of the instances in negative examples and positive otherwise.

$$\text{ext2int}^{-}(c) = \bigwedge_{(x,0) \in c} \left(\bigvee_{(f,0) \in x} f \quad \vee \bigvee_{(f,1) \in x} \overline{f} \right) \tag{17}$$

If we use the set of examples in equation eq. (15), we can form an intensional concept both ways. From the positive examples we have $c^{+} = (\overline{f_0} \wedge f_1) \vee (f_0 \wedge \overline{f_1}) \vee (f_0 \wedge f_1)$ and from the negative examples we have $c^{-} = (f_0 \vee f_1)$. In the case where the provided examples are *partial* examples, the learner can infer the concept label for instances not covered by the examples. If the intensional form of a concept is known, the output label can be predicted for any instance defined over the same features and thus intensional concepts conveniently provide a set of potentially relevant features as in theorem 3. Thus, a teacher can communicate an intensional concept through the use of partial examples.

Just as intensional concepts can be derived from the extensional definition, a sample compatible with cooperative inference can be derived from an intensional concept. Let c_d and c_c be concepts in DNF and CNF form, respectively.

$$\text{int2ext}^{+}(c_d) = \bigcup_{t \in c_d} \left\{ \left(\bigcup_{f \in L_t} \{(f,1)\} \cup \bigcup_{\overline{f} \in L_t} \{(f,0)\}, 1 \right) \right\} \tag{18}$$

$$\text{int2ext}^{-}(c_c) = \bigcup_{cl \in c_c} \left\{ \left(\bigcup_{\overline{f} \in L_c} \{(f,1)\} \cup \bigcup_{f \in L_c} \{(f,0)\}, 0 \right) \right\} \tag{19}$$

These two equations imply a logical correspondence between a teaching set for a concept constructed of partial examples and intensional forms of that concept. If a teacher has a concept stored intensionally, such as $c = f_1 \vee f_0$ 'feps are red or square', the teacher easily convert this definition to the teaching sample $\{((f_0,1),1),((f_1,1),1)\}$ 'red is a fep, square is a fep'.

This means that logical operations on one form can be leveraged for the other, e.g. simplifying the intensional definition results in an equivalent simplification of the extensional teaching set. For example the rule used to combine terms and clauses in the classic Quine-McCluskey algorithm (Roth, 2013)—e.g. 'feps are either red and small or red and not small' $\rightarrow$ 'feps are red'—has a logical equivalent with cooperative inference.

If a teacher has a concept in both DNF and CNF form, a concept can be inferred from the results of eqs. (18) and (19) using partial examples, $c = \text{Match}(s^{+}, \mathcal{X}) \cup \text{Match}(s^{-}, \mathcal{X})$ where $s^{+} = \text{int2ext}^{+}(c_d)$ and $s^{-} = \text{int2ext}^{-}(c_c)$ and vice versa for eqs. (16) and (17). A teacher may not need both DNF and CNF forms; a very large feature space combined with sparse positive instances would support a principle of truth (see e.g.

P. N. Johnson-Laird, 2001). In such a case, if a teacher were to only present positive examples a cooperative learner would infer that the teacher was helpfully omitting the vast set of negative examples.

This correspondence between the intensional form of a concept and its teaching set suggests that DNF is a natural representation for teaching. When a concept is stored in this way, teaching no longer involves searching a space of teaching samples for the solution; rather, the solution is the representation itself. In this way, teaching via cooperative inference would inform sampling biases through representation. A DNF-based representational bias is consistent with some of the most successful models and experiments in concept learning (Feldman, 2000; Goodwin & P. Johnson-Laird, 2011; Goodman et al., 2008).

Discussion

In cognitive science and machine learning, previous models have sought to explain teaching through the helpful sampling of examples with respect to a known bias. We have argued that there is no single rational bias, that if the teacher does not know the learner's bias, the probability of teaching goes to zero as the number of features increases, and that a sampling account is ineffective without a bias. Thus, sampling alone is an incomplete explanation of teaching.

We have proposed a solution that begins with a notion of cooperation instead of biases. In the cooperative inference model, a teacher not only samples examples but is able to omit unnecessary feature specifications. The most immediate effect of cooperative inference is a novel, powerful, and realistic account of teaching.

In the sampling-based accounts of teaching, the teacher searches the space of teaching samples for one that rules out all other concepts in the concept space—so the computational complexity of the model increases with the size and complexity of the concept space. This represents a significant limitation, as the size of the concept space increases superexponentially with the number of features and a realistic world includes many—if not infinitely many—features. For cooperative inference, the complexity of teaching increases with the complexity of the target concept rather than the complexity of the concept space. Additionally, the model indicates a natural representation for teaching such that once a concept has been learned through teaching or represented for teaching, minimal work is needed in order to teach the concept again.

Our work is related to a variety of trends in cognition and cognitive development (Csibra & Gergely, 2009; Tomasello, Carpenter, Call, Behne, Moll, et al., 2005; Clark & Wilkes-Gibbs, 1986). Most notably, the assumption of common knowledge and common ground is key to theories of language and communication (Clark & Wilkes-Gibbs, 1986). Our problem differs from the standard formulation in that theories of language have focused on how to link utterances to referents. In contrast, we have focused on the case where referents are clear, and the challenge is linking examples to concepts. Nevertheless, both problems require a priori assumptions about the other party, and a similar approach may yield insights in the other domain as well.

A critical component of any model of human behavior is the choice of bias. Before now, these choices have been made by a combination of intuition and informal argument (see especially Anderson, 1990; Feldman, 2000; Goodman et al., 2008; Goodwin & P. Johnson-Laird, 2011). We have presented a formal analysis that provides an a priori justification for choice of bias in the case of teaching. However, the promise of this work is in the generality of the approach, and we are optimistic that similar methods can be applied to more general learning problems.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press. (Cit. on p. 6).

Anthony, M., Brightwell, G., Cohen, D., & Shawe-Taylor, J. (1992). On exact specification by examples. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 311–318). New York: ACM. (Cit. on p. 1).

Balbach, F. J. (2008). Measuring teachability using variants of the teaching dimension. *Theoretical Computer Science*, *397*(1), 94–113. (Cit. on pp. 1, 2).

Clark, H. H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. (Cit. on p. 6).

Csibra, G. & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153. (Cit. on pp. 1, 4, 6).

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633. (Cit. on pp. 1, 6).

Goldman, S. A. & Mathias, H. D. (1993). Teaching a smarter learner. In *Proceedings of the sixth annual conference on computational learning theory* (pp. 67–76). ACM. (Cit. on p. 1).

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. (Cit. on pp. 1, 2, 6).

Goodwin, G. P. & Johnson-Laird, P. (2011). Mental models of boolean concepts. *Cognitive psychology*, *63*(1), 34–59. (Cit. on pp. 1, 6).

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, *5*(10), 434–442. (Cit. on p. 5).

Roth, J. C. H. (2013). *Fundamentals of logic design*. Cengage Learning. (Cit. on p. 5).

Shafto, P. & Goodman, N. D. (2008). Teaching games: statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the thirtieth annual conference of the cognitive science society* (pp. 1632–1637). (Cit. on pp. 1, 2).

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42. (Cit. on p. 1).

Shinohara, A. & Miyano, S. (1991). Teachability in computational learning. *New Generation Computing*, *8*, 337–347. (Cit. on p. 1).

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., et al. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and brain sciences*, *28*(5), 675–690. (Cit. on p. 6).

Watanabe, S. (1969). *Knowing and guessing: a quantitative study of inference and information*. Wiley New York. (Cit. on p. 2).

Xu, F. & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245. (Cit. on p. 1).

Zilles, S., Lange, S., Holte, R., & Zinkevich, M. (2009). Teaching dimensions based on cooperative learning. (Cit. on pp. 1, 2).