Toward a Theory of Learning Coherent Concepts

Dan Roth Dmitry Zelenko

Department of Computer Science
University of Illinois at Urbana-Champaign
{danr,zelenko}@cs.uiuc.edu

Abstract

We develop a theory for learning scenarios where multiple learners co-exist but there are mutual compatibility constraints on their outcomes. This is natural in cognitive learning situations, where "natural" compatibility constraints are imposed on the outcomes of classifiers so that a valid sentence, image or any other domain representation is produced.

We suggest that work in this direction may help to resolve the contrast between the hardness of learning as predicted by the current theoretical models and the apparent ease at which cognitive systems seem to learn.

A model of concept learning is studied in which the target concept is required to cohere with other concepts of interest. The coherency is expressed via a (Boolean) constraint that the concepts have to satisfy. Under this model, learning a concept is shown to be easier (in terms of sample complexity and mistake bounds) and the concepts learned are shown to be more robust to noise in their input (attribute noise). These properties are established for half spaces and the connection to large margin theory is discussed.

Introduction

The emphasis of the research in learning theory is on the study of learning single concepts from examples. In this framework the learner attempts to learn a single hidden function from a collection of examples (or more expressive modes of interaction) and its performance is measured when classifying future examples. The theoretical research in this direction (Valiant 1984; Vapnik 1995) has already proved useful in that it has contributed to our understanding of some of the main characteristics of the learning phenomenon as well as to applied research on classification tasks (Druker, Schapire, & Simard 1993; Golding & Roth 1999). One puzzling problem from a theoretical and a practical point of view, is the contrast between the hardness of learning problems – even for fairly simple concepts – as predicted by the theoretical models, and the apparent ease at which cognitive systems seem to learn those concepts. Cognitive systems seem to use far less examples and learn more robustly than is predicted by the theoretical models developed so far.

In this paper we begin the study of a new model within which an explanation of this phenomenon may be developed. Key to this study is the observation that cognitive learning problems are do not usually occur in isolation. Rather, the input is observed by multiple learners that may learn different functions on the same input. In our model, the mere existence of the other functions along with the constraints Nature imposes on the relations between these functions – all unknown to the learner – contribute to the effective simplification of each of the learning tasks.

Assume for example that given a collection of sentences where each word is tagged with its part-of-speech (pos) as training instances, one wants to learn a function that, given a sentence as input, predicts the pos tag of the ith word in the sentence. E.g., we would like to predict the pos tag of the word `can` in the sentence `This can will rust`[1]. The function that predicts this pos may be a fairly complicated function of other tokens in the sentence; as a result, it may be hard to learn. Notice, however, that the same sentence is supplied as input to the function that predicts the pos of the word `will` and that, clearly, the predictions of these functions are not completely independent. Namely, the presence of the function for `will` may somewhat constrain the function for `can`. For example, the constraint may be that these functions never produce the same output when evaluated on a given sentence. This exemplifies our notion of coherency: given that these two functions need to produce coherent outputs, the input sentence may not take any possible value in the input space of the functions (that it could have taken when the function's learnability is studied in isolation) but rather may be restricted to a subset of the inputs on which the functions outcomes are coherent.

The learning scenario is that of concept learning from examples, where a learner is trying to identify

[1] This may not be the exact way one chooses to model the problem (Brill 1995; Roth & Zelenko 1998). However, this is a reasonable abstraction that helps deliver the intuition behind our point of view.

a concept $f \in \mathcal{F}$ when presented with examples labeled according to f. We study learning in the standard pac (Valiant 1984) and mistake bound (Littlestone 1988) learning models. It is well known that learnability in the pac model depends on the complexity of the hypothesis class. Specifically, it is equivalent to the finiteness of the VC-dimension (Vapnik & Chervonenkis 1971), a combinatorial parameter which measures the richness of the function class (see (Vapnik 1995; Kearns & Vazirani 1994) for details). Moreover, it is known (Blumer et al. 1989; Ehrenfeucht et al. 1989) that the number of examples required for learning is linear in the VC-dimension of the class. Mistake bound learning is studied in an online setting (Littlestone 1988); the learner receives an instance, makes a prediction on it, and is then told if the prediction is correct or not. The goal is to minimize the overall number of mistakes made throughout the learning process. The usual way to constrain the learning task is to explicitly restrict the concept class. Instead, here we are mostly concerned with the case in which the restriction is imposed implicitly via interaction between concepts. More precisely, we are interested in a learning scenario that involves several concepts $f_1, f_2, \ldots, f_k$ from the concept class $\mathcal{F}$. Let $g : \{0,1\}^k \to \{0,1\}$ be any Boolean function of k variables. The notion of coherency we study is formalized by assuming that the concepts $f_1, f_2, \ldots, f_k$ are subjected to a constraint g. In all cases, however, we are interested in learning a single function $f_1 \in \mathcal{F}$ under these conditions.

There exists several possible semantics for the coherency conditions and here we present only the one that we find most promising in that we can present results that indicate that the task of learning f becomes easier in these situations. We then study the effect of the coherence assumption on the robustness of the learned concepts. In particular, we study the robustness of learnable concepts to attribute noise. This type of robustness is important in cognitive systems, where multiple concepts are learned and "chained" (Valiant 1999; Khardon, Roth, & Valiant 1999). Namely, the output of one learned predictor may be used as input to another learned predictor. Errors in the output of one predictor therefore translate to attribute noise in the input to another, and predictors have to tolerate it to support chaining well; we show that learning coherent concepts is robust and briefly discuss relations to large margin classification and future work.

Class Coherency

Before getting to the main definition we introduce a condition that turns out to be too strong and leads to a restriction on the function class.

Let $\mathcal{F}$ be a concept class over X. The direct k-product $\mathcal{F}^k$ of the concept class $\mathcal{F}$ is the set $\mathcal{F}^k = \{f : f = (f_1, \ldots, f_k), f_i \in \mathcal{F}, i = 1, \ldots, k\}$. That is, if $f \in \mathcal{F}^k$, $f : X \to \{0,1\}^k$. Thus, learning k functions with a binary range can be reduced to learning a single function with range $\{0, \ldots, 2^k - 1\}$. The following theorem (Ben-David et al. 1995) states that this transformation (and its inverse) preserves PAC learnability[2].

Theorem 1 $\mathcal{F}^k$ is learnable iff $\mathcal{F}$ is learnable.

Definition 1 (Class Coherency) Let $\mathcal{F}$ be a concept class and $g : \{0,1\}^k \to \{0,1\}$ a Boolean constraint. $\mathcal{F}_g^k \subseteq \mathcal{F}^k$ is a coherent collection of functions if $\mathcal{F}_g^k = \{(f_1, \ldots, f_k) \in \mathcal{F}^k : \forall x \in X, (g(f_1(x), \ldots, f_k(x)) = 1)\}$.

Intuitively we can think of g as reducing the range of functions in $\mathcal{F}^k$. That is, if $Y = g^{-1}(1)$, then we do not care about elements $f \in \mathcal{F}^k$ for which $range(f) \not\subseteq Y$.

The observation that a constraint g reduces the range of the functions in $\mathcal{F}^k$ leads to the following sample size bound for pac-learning $\mathcal{F}^k$ (immediate from (Ben-David et al. 1995)).

Theorem 2 Let $m = |g^{-1}(1)|$. Then, the pac learning sample complexity of $\mathcal{F}_g^k$ is $O(\frac{1}{\epsilon}(d(\log m)\log\frac{1}{\epsilon} + \log\frac{1}{\delta}))$ where d is any appropriate capacity measure of $\mathcal{F}_g^k$.

Example 1 Let $\mathcal{F}$ be the class of axis-parallel rectangles inside $[0,1]^2$. Let $g(f_1, f_2) \equiv (f_1 \neq f_2)$. Then $\mathcal{F}_g^2$ is the class of the pairs (f_1, f_2) of axis-parallel rectangles, where f_1 is the complement of f_2 in $[0,1]^2$. Note that in this case $\mathcal{F}_g^2$ is a class of functions with the binary range $\{01, 10\}$. For binary-valued functions, the appropriate capacity measure of Theorem 2 is the VC-dimension of $\mathcal{F}_g^2$. It is not difficult to see that three points can be shattered by the concept class, but no four points can. Therefore, $VCD(\mathcal{F}_g^2) = 3$; however, $VCD(\mathcal{F}) = 4$ and, hence, Theorem 2 implies that the sample complexity of learning the concept class $\mathcal{F}$ alone is greater than the sample complexity of learning it in the presence of other functions when they are all constrained by g. Thus, adding more concepts may make learning easier.

While definition 1 captures the simultaneous nature of the learning scenario, it is too restrictive in that it imposes global constraints on all the k functions. We would like to relax this further and emphasize that the goal here is to study the learnability of a single function, say, f_1, and how it is affected by the presence of the other functions and the requirement that they behave coherently. The next section develops the main definition of this paper.

Distributional Coherency

In the previous section we removed from $\mathcal{F}^k$ any f, such that $g(f(x)) = 0$ for some $x \in X$. Now, we change the semantics of the constraint imposed on the direct product $\mathcal{F}^k$. For each $f \in \mathcal{F}^k$, we simply restrict the domain of f to X', where $\forall x \in X', g(f(x)) = 1$.

Definition 2 (Distributional Coherency) Given a Boolean constraint g and a class $\mathcal{F}$ of functions, we define the class of g-coherent functions $\mathcal{F}_g^\star$ to be the

[2] PAC learnability for multi-valued functions is shown to be characterized by the finiteness of a capacity measure of a function class, see (Ben-David et al. 1995) for details.

collection of all functions $f^\star : X \to \{0,1\}^k \cup \{\star\}$ in $\mathcal{F}^k$ defined by

$$f^\star(x) = \begin{cases} f(x) & \text{if } g(f(x)) = 1 \\ \star & \text{otherwise} \end{cases}$$

We interpret the value of "$\star$" as a forbidden value for the function f. In this way we restrict the domain of f to the subset X' of X satisfying the constraint g.

The constraint semantics in Def. 1 is stronger (more restricting) than the one in Def. 2. E.g., let $\mathcal{F}$ be the class of (non-identically false) monotone DNF, and g is $(f_1 \neq f_2)$. Then, $\mathcal{F}_g^2$ is empty, because $f_1(1) = 1 = f_2(1)$, for any $f_1, f_2 \in F$. But, in $\mathcal{F}_g^\star$, we simply restrict the domain of each f_1, f_2 to the non-overlapping areas of f_1, f_2.

In the pac learning model the above constraint can be interpreted as restricting the class of distributions when learning a function $f_1 \in \mathcal{F}$. Only distributions giving zero weight to the region $X \setminus X'$ are allowed.

Definition 3 Let $\mathcal{F}$ be a class of Boolean functions over X. Let $f_1, \ldots, f_k \in \mathcal{F}$ be subjected to a constraint g. Then, a distribution D over X is said to be f_1-compatible w.r.t to $f_2, \ldots, f_k \in \mathcal{F}$ and g, if $D\{x : f^\star(x) = \star\} = 0$, where $f = (f_1, \ldots, f_k)$. We denote by $\mathcal{D}_{f_1}$ the class of all f_1-compatible distributions (w.r.t to $f_2, \ldots, f_k \in \mathcal{F}$ and g).

To motivate investigation into the gain one might expect to have in this learning scenario, consider the following example.

Example 2 Let $\mathcal{F}$ be the class of disjunctions. Consider learning f_1 from examples, in the presence of f_2 and the constraint $g \equiv (f_1 \neq f_2)$. Suppose that both f_1 and f_2 include a literal l. The constraint implies that X' does not contain examples where l is 1 and effectively reduces the size of the target disjunction f_1 since the existence of literals common to f_1 and f_2 in the target disjunction is irrelevant to predictions on X'. Thus, if n_1, n_2 is the number of literals in f_1, f_2, respectively, n_c is the number of common literals, then using a feature efficient algorithm like Winnow (Littlestone 1988) to learn f_1 in the presence of f_2 and the constraint g gives an improved mistake bound of $2(n_1 - n_c)(\log n_1 + 1)$.

The only model we know of that is related to the model studied here is the one studied in (Blum & Mitchell 1998). Our model can be viewed as a generalization of the Blum and Mitchell model. They study learning two functions f_1, f_2 over different domains (X_1 and X_2, respectively), where the learner sees only pairs $(x_1, x_2) \in X = X_1 \times X_2$ that satisfy $f_1(x_1) = f_2(x_2)$. This is a special case of our model, when $x = (x_1, x_2)$ and the functions f_1, f_2 are defined over subdomains X_1, X_2 rather than the whole X. In example 2, if restricted to monotone disjunctions, we get the domain decomposition for free, because the constraint forces the literal sets of the disjunctions to be disjoint. Thus, by applying the results of (Blum & Mitchell 1998)[3], one can quantify the reduction in the number of examples needed for learning constrained monotone disjunctions. Next we analyze a more general case of learning in the coherency model.

Learning Linear Separators

Let $\mathcal{F}$ be the class of half-spaces in R^2 and let g be $(f_1 = f_2)$. f_1 and f_2 are depicted in Figure 1. The arrows point in the direction of the positive half-spaces with respect to the corresponding lines. The constraint g restricts the domains of both f_1 and f_2 to the shaded areas. Therefore, when learning f_1 (and similarly, f_2) we will see examples only from the shaded areas $X' \subseteq X$. For $x \in X'$, $f_1(x) = f_2(x)$. While, in principle, learning f_1 may be hard due to examples nearby the separator, now there are many linear separators consistent with $f_1(x)$. Therefore, at least intuitively, finding a good separator for $f_1(x)$ would be easier. For the case when the linear separator is learned via the Perceptron learning algorithm, we can show the following. (Proof omitted. The theorem is stated for the constraint $(f_1 = f_2)$ but can be phrased to any symmetric constraint.)

Theorem 3 Let f_1 and f_2 be two hyperplanes (w.l.o.g, passing through the origin) with unit normals $w_1, w_2 \in R^n$, respectively. Let $\alpha = \cos(w_1, w_2) = w_1 \cdot w_2$. Let $S = S^+ \cup S^-$ be the sequence of positive and negative examples so that $\forall x \in S, f_1(x) = f_2(x)$. Let S be linearly separable by both f_1 and f_2 with margins $2\delta_1$ and $2\delta_2$, respectively. If $\sup_{x \in S} |x| < R$ then the number of mistakes the Perceptron makes on S is bounded by $\beta \frac{R^2}{\delta^2}$, where $\beta = \frac{1+\alpha}{2}, \delta = \frac{\delta_1 + \delta_2}{2}$.

The general Perceptron mistake bound is $\frac{R^2}{\delta^2}$(Novikoff 1963), where δ is the margin of the target hyperplane. The presence of f_2 and the constraint g improves the mistake bound by a factor of β. As the shaded regions become smaller, α approaches -1, and, hence, β approaches 0.

While Theorem 3 shows the gain in mistake bound when learning w_1 (as a function of w_2 and the constraint) it is possible to quantify this gain in an algorithmic independent way by characterizing the set $E(w_1, w_2)$ of linear separators consistent with the imposed constraint[4]. Given w_2 and the constraint g, denote by $E(w_1, w_2)$ the set of all linear separators that can be learned without any loss in accuracy when the target concept is w_1. Formally (omitting the dependence on g from the notation), for any two vectors $w_1, w_2 \in R^n$, let $X^{'} = \{x : x \in R^n, sgn(w_1 \cdot x) = sgn(w_2 \cdot x)\}$. That is, $X^{'}$ corresponds to the shaded area in Figure 1. Then:

$$E(w_1, w_2) = \{w : w \in R^n, \forall x \in X', sgn(w \cdot x) = sgn(w_1 \cdot x)\}$$

Theorem 4 uses the well-known Farkas' Lemma (Mangarasian 1969) from linear programming.

[3] The results in (Blum & Mitchell 1998) require in addition certain conditional independence assumptions.

[4] The set $E(w_1, w_2)$ depends on the constraint g. The results in this section can be presented for any symmetric constraint on w_1, w_2, but will be presented, for clarity, only for equality.

Lemma 1 For any $m \times n$ matrix A and vector c in R^n, either

$$\{x : Ax \leq 0, c \cdot x > 0\} \; is \text{ non-empty} \tag{1}$$

or

$$\{y : A^T y = c, y \geq 0\} \; is \text{ non-empty} \tag{2}$$

but never both.

We now have:

Theorem 4 $E(w_1, w_2) = \{w : w = aw_1 + bw_2; a, b \in R; a, b > 0\}$.

Proof: Denote $W = \{w : w = aw_1 + bw_2; a, b \in R; a, b > 0\}$. Clearly, $W \subseteq E(w_1, w_2)$. In order to prove that $E(w_1, w_2) \subseteq W$, we partition X' in two sets,

$$X'_+ = \{x : x \in R^n, w_1 \cdot x \geq 0, w_2 \cdot x \geq 0\}$$

and

$$X'_- = \{x : x \in R^n, w_1 \cdot x \leq 0, w_2 \cdot x \leq 0\}.$$

Observe that $X'_- = \{-x : x \in X'_+\}$. Fix a $w \in R^n$, so that $w \cdot x \geq 0$ on X'_+. Hence, $w \cdot x \leq 0$ on X'_-, and $w \in E(w_1, w_2)$. Now apply Lemma 1 with $A = \binom{w_1}{w_2}$ (A is an $2 \times n$ matrix whose rows are w_1, w_2), and $c = w$. Since $w \cdot x \leq 0$ on X'_-, (1) is not satisfied; hence, (2) is satisfied, and $w = aw_1 + bw_2$, where a, b are some positive numbers. Thus, $E(w_1, w_2) \subseteq W$. $\Box$

Requiring the members of $E(w_1, w_2)$ to be unit vectors, unconstrained learning of f_1 can be viewed geometrically as searching for a point on the unit sphere that is close to the target w_1. In the presence of w_2 and the constraint, we have:

Corollary 1 The intersection of $E(w_1, w_2)$ with the unit sphere is a curve on the unit sphere in R^n connecting w_1 to w_2, with length $cos^{-1}(w_1 \cdot w_2)$.

Thus in the presence of w_2 and the constraint, it is sufficient now for the learning algorithm to find a point on the sphere that is close to any of the curve points; algorithmically, for the Perceptron, this translates to reducing the mistake bound proportionally to the length of this curve.

Robustness

In this section we show that the coherence assumption made in this paper has the effect of making the learned concepts more robust. We start by defining robustness and proving that concepts learned under this model can indeed be evaluated robustly (generalizing previous models of attribute noise); we then show that learning coherent concepts is robust and discuss the relation to large margin theory.

Definition 4 (Attribute Robustness) For $x, y \in \{0, 1\}^n$ let $H(x, y)$ be the Hamming distance between x and y (that is, the number of bits on which x and y differ). Let $S_k(f) = \{x : \forall y, \text{ if } H(x, y) \leq k \text{ then } f(x) = f(y)\}$. We say that the pair (D, f) is (ϵ, k)-robust, if $D(S_k(f)) > 1 - \epsilon$.

Intuitively, the condition means that w.h.p. all the points in a ball of radius k around any point x have the same label. This can be relaxed by requiring $f(y) = f(x)$ to hold only for a $(1 - \gamma)$ portion of the points in the ball $B_k = \{y : H(x, y) \leq k\}$, but we will not discuss this to simplify technical details.

Let f be a concept over $X = \{0, 1\}^n$ and let D be a distribution over X. We denote by D^k_{flip} the distribution that results from choosing k bits uniformly and flipping them. It is easy to see that if (D, f) is (ϵ, k)-robust, and $x \in S_k(f)$, then flipping k bits of x does not change the value of f. Hence, $error_{D^k_{flip}}(f) \leq D(x \notin S_k(f)) \leq \epsilon$ and the robustness condition guarantees a small error when evaluating f on the noisy distribution. It follows that if h is an ϵ-good hypothesis for f under D, and if (D, f) is (ϵ, k)-robust, then h is a 2ϵ-good hypothesis under D^k_{flip}.

We note that the distribution D^k_{flip} is an example of an attribute noise model (Goldman & Sloan 1995; Decatur & Gennaro 1995). These models usually assume the presence of noise in the learning stage and aim at learning a good approximation of the target concept over the original noiseless distribution. However, as can be readily seen (and has been pointed out in (Goldman & Sloan 1995)), in a more realistic setting in which the learned hypothesis is to be evaluated under noisy conditions, this hypothesis may be useless. The robustness condition defined above guarantees that a hypothesis learned in the presence of noise also performs well when being evaluated under these conditions. This holds for a more general attribute noise model, the product attribute noise, defined as follows. Let D be a distribution on the instance space $X = \{0, 1\}^n$. Assume that an attribute i of an example $x \in X$ sampled according to D is flipped independently with probability p_i, $i = 1, \ldots, n$. Denote by $p = \sum_{i=1}^n p_i$ the expected number of bits flipped in an example. We denote the distribution induced this way on X by D^p_{flip}.

Theorem 5 Let (D, f) be (ϵ, k)-robust. If $k \geq p + \sqrt{2n \ln \frac{1}{\epsilon}}$, then $error_{D^p_{flip}}(f) \leq 2\epsilon$.

Proof: Let $(x, f(x))$ be an example sampled according to D. Let x' be the result of flipping the bits of x according to the noise scheme described above. Denote by Pr the product distribution induced by the bit flipping. Then we have:

$$\begin{aligned} error_{D^p_{flip}}(f) &= D^p_{flip}\{x' : f(x') \neq f(x)\} \\ &= D^p_{flip}\{x' : x \in S_k, f(x') \neq f(x)\} + \\ &\quad D^p_{flip}\{x' : x \notin S_k, f(x') \neq f(x)\} \\ &\leq D^p_{flip}\{x' : x \in S_k, f(x') \neq f(x)\} + \\ &\quad \epsilon \leq Pr\{H(x, x') > k\} + \epsilon \end{aligned}$$

To bound $Pr\{H(x, x') > k\}$, we let Y be the random variable describing the number of bits flipped in an example. Note that $Y = H(x, x')$ and $E[Y] = p$. Also let

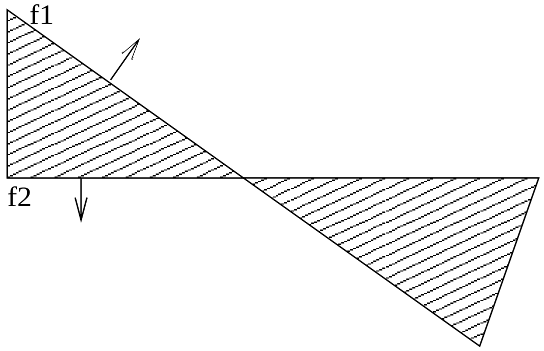

Figure 1: Constrained Half-spaces in R^2.

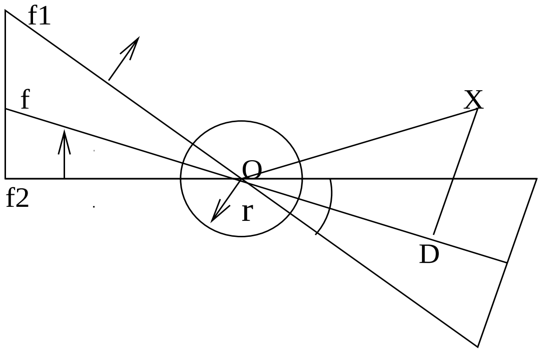

Figure 2: Constrained Robust Half-spaces

$m = \sqrt{2n \ln \frac{1}{\epsilon}}$. Then,

$$Pr\{Y > k\} = Pr\{Y > p+m\} = Pr\{Y - E[Y] > m\} \le e^{\frac{-m^2}{2n}},$$

where the last inequality follows directly from the Chernoff bounds (Hoeffding 1963). Hence, $Pr\{H(x,x') > k\} \le \epsilon$, and $error_{D^p_{flip}}(f) \le 2\epsilon$. □

Thus, if we have an ϵ-good hypothesis for noiseless distribution, and the target concept with the underlying distribution satisfy the above (ϵ, k)-robustness condition then the hypothesis will also be 3ϵ-good for the noisy distribution.

Coherency implies Robustness

We now establish the connection between coherency and robust learning. This is done again in the context of learning linear separators. As before, the target function is f_1, and we assume the presence of f_2 (w.l.o.g., both f_1, f_2 pass through the origin), and that they are subjected to the equality constraint g. However, here we restrict the domain of f_1 and f_2 to $X = \{0,1\}^n$. Let D be a distribution over X. We require the distribution to give small weight to points around the origin. Formally, let $B_r = \{x : x \in X, |x| \le r\}$ be the origin-centered ball of radius r. Then we require D to satisfy $D(B_r) < \epsilon$.

Notice that in general, when learning a single linear separator f, this property of D does not imply that (D, f) is robust. The following theorem shows that with the equality constraint imposed, the property is sufficient to make (D, f) robust.

Theorem 6 Let f_1 and f_2 be hyperplanes over $\{0,1\}^n$ (through the origin) with unit normals $w_1, w_2 \in R^n$, respectively. Let $\alpha = \cos(w_1, w_2) = w_1 \cdot w_2$. Let D be a f_1-compatible distribution (w.r.t f_2 and the equality constraint g) that satisfies $D(B_r) < \epsilon$, where $r > k\sqrt{\frac{2}{1-\alpha}}$. Then, there is a linear separator f, so that $D(x : f(x) \ne f_1(x)) = 0$ and f is (ϵ, k)-robust.

Proof: (Sketch) The idea of the proof is to exhibit a linear separator f that is consistent with f_1 on D and, for all points lying outside B_r, has a large "margin" separating positive examples from negative ones. Let f be the hyperplane bisecting the angle between f_1 and f_2. That is, $w = \frac{1}{2}(w_1 + w_2)$, where w is the normal vector of f. By theorem 4, $w \in E(w_1, w_2)$; hence, $D(x : f(x) \ne f_1(x)) = 0$. Now fix a point $x \in R^n$, so that $f_1(x) = f_2(x)$ and $x \notin B_r$. Figure 2. is the projection of f_1, f_2, f to the 2-dimensional plane determined by the origin, the point x and x's projection onto f. (Notice that in Figure 2 X' is the complement of the shaded region in Figure 1.) Then we have that $|XD| = |x| \sin(XOD) \ge r\sqrt{\frac{1-\alpha}{2}}$. If $r > k\sqrt{\frac{2}{1-\alpha}}$, then $|XD| > k$, hence flipping $\le k$ bits of x will not change the value of f ($|XD|$ is the distance from x to f). Therefore, f is $(0, k)$-robust for any point of the subdomain X' satisfying the constraint and lying outside of the ball B_r (and robustness grows as the size of X' decreases). This implies that (D, f) is (ϵ, k)-robust. □

We note that the assumption $D(B_r) < \epsilon$ in the Theorem 6 is satisfied if there is a margin r separating positive examples of f_1 from its negative examples (Vapnik 1995), so that the weight (with respect D) of examples lying inside the margin is less than ϵ. Also, existence of such a distributional margin implies that a sample of examples from the distribution will be linearly separable with margin at least r with high probability, thus guaranteeing that there is a large margin hyperplane consistent with the sample, that has small error with respect to D (Freund & Schapire 1998). In particular, we construct such a hyperplane f in the proof of Theorem 6.

Conclusions

This paper starts to develop a theory for learning scenarios where multiple learners co-exist but there are mutual compatibility constraints on their outcomes. We believe that these are important situations in cognitive learning, and therefore this study may help to resolve some of the important questions regarding the easiness and robustness of learning that are not addressed adequately by existing models. In addition, we view this model as a preliminary model within which to study learning in a multi-modal environment. Several classifiers feed from the same data, each makes predictions with respect to some facets of the target concept, depending on the modality (identifying images of dogs, identifying barks, etc.). However, these predictions need to satisfy constraints by virtue of representing different aspects of the same concept; this is exactly the situation studied in the coherency model introduced here.

We have shown that within this model the problem of learning a single concept – when it is part of an existing collection of coherent concepts – is easier than in the general situation. Moreover, this gain is due only to the existence of the coherency, even if the learner is unaware of it. Although the results of this paper are restricted mostly to linear separators we do not view this as a severe restriction given their universal nature in theory and applications (Roth 1999; Cortes & Vapnik 1995).

Some of the future directions of this work include the study of coherent concepts under more general families of constraints as well as algorithmic questions that arise from this point of view. These include, in particular, algorithmic questions that aim at exploiting the constraints in order to better learn coherently behaving concepts as in (Munoz et al. 1999).

Acknowledgments

This research is supported by NSF grants IIS-9801638 and IIS-9984168.

References

Ben-David, S.; Cesa-Bianchi, N.; Haussler, D.; and Long, P. M. 1995. Characterizations of learnability of {0,...,n}-valued functions. Journal of Computer and System Sciences 74–86.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In Proc. of the Annual ACM Workshop on Computational Learning Theory, 92–100.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1989. Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM 36(4):929–865.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21(4):543–565.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. Machine Learning 20:273–297.

Decatur, S. E., and Gennaro, R. 1995. On learning from noisy and incomplete examples. In Proc. of the Annual ACM Workshop on Computational Learning Theory, 353–360.

Druker, H.; Schapire, R.; and Simard, P. 1993. Improving performance in neural networks using a boosting algorithm. In Neural Information Processing Systems 5, 42–49. Morgan Kaufmann.

Ehrenfeucht, A.; Haussler, D.; Kearns, M.; and Valiant, L. 1989. A general lower bound on the number of examples needed for learning. Information and Computation 82(3):247–251.

Freund, Y., and Schapire, R. 1998. Large margin classification using the Perceptron algorithm. In Proc. of the Annual ACM Workshop on Computational Learning Theory, 209–217.

Golding, A. R., and Roth, D. 1999. A Winnow based approach to context-sensitive spelling correction. Machine Learning 34(1-3):107–130. Special Issue on Machine Learning and Natural Language.

Goldman, S. A., and Sloan, R. H. 1995. Can PAC learning algorithms tolerate random attribute noise? Algorithmica 14(1):70–84.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58(301):13–30.

Kearns, M., and Vazirani, U. 1994. Introduction to computational Learning Theory. MIT Press.

Khardon, R.; Roth, D.; and Valiant, L. G. 1999. Relational learning for NLP using linear threshold elements. In Proc. of the International Joint Conference of Artificial Intelligence.

Littlestone, N. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine Learning 2:285–318.

Mangarasian, O. L. 1969. Nonlinear Programming. McGraw-Hill.

Munoz, M.; Punyakanok, V.; Roth, D.; and Zimak, D. 1999. A learning approach to shallow parsing. In EMNLP-VLC'99, the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Novikoff, A. 1963. On convergence proofs for perceptrons. In Proceeding of the Symposium on the Mathematical Theory of Automata, volume 12, 615–622.

Roth, D., and Zelenko, D. 1998. Part of speech tagging using a network of linear separators. In COLING-ACL 98, The 17th International Conference on Computational Linguistics, 1136–1142.

Roth, D. 1999. Learning in natural language. In Proc. Int'l Joint Conference on Artificial Intelligence, 898–904.

Valiant, L. G. 1984. A theory of the learnable. Communications of the ACM 27(11):1134–1142.

Valiant, L. G. 1999. Robust logic. In Proceedings of the Annual ACM Symp. on the Theory of Computing.

Vapnik, V. N., and Chervonenkis, A. Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its applications XVI(2):264–280.

Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. New York: Springer-Verlag.