Exploiting Ontology Structures and Unlabeled Data for Learning

Maria-Florina Balcan NINAMF@CC.GATECH.EDU
School of Computer Science, Georgia Institute of Technology, Atlanta, GA

Avrim Blum AVRIM@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Yishay Mansour MANSOUR@TAU.AC.IL
Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

Abstract

We present and analyze a theoretical model designed to understand and explain the effectiveness of ontologies for learning multiple related tasks from primarily unlabeled data. We present both information-theoretic results as well as efficient algorithms.

We show in this model that an ontology, which specifies the relationships between multiple outputs, in some cases is sufficient to completely learn a classification using a large unlabeled data source.

1. Introduction

Overview In a number of modern applications of machine learning (including text understanding, medical diagnosis, search, and vision), one would like to learn multiple related tasks from primarily, or even solely, unlabeled data. One approach to this task (Carlson et al., 2010b) is to take advantage of known relationships among the classes being learned, which are often captured by an *ontology*; such an ontology can either be provided by domain experts or learned from past data. In this work, we develop and analyze a theoretical model aimed at providing a mathematical understanding of learning methods of this type.

More specifically, we propose and analyze a model for the following setting. Suppose we would like to solve L different classification problems, which potentially have very different sets of features, given only a large supply of unlabeled data. In order to help us with this difficult task, we are supplied with an ontology specifying relations among the L categories. For instance, this ontology might say that an example to be classified cannot be both a `company` and a `product`, or that if someone is an `athlete`, then they are also a `person`. The question our model aims to address is how learning these *multiple* concepts (`company, product, person, athlete,` etc.) *in parallel*, aided by constraints provided by the given ontology, can be done using unlabeled data.

Motivation A prime example motivating our work is the NELL Never-Ending-Language-Learning system (Carlson et al., 2010a;b) (`rtw.ml.cmu.edu`) which has had substantial success in learning a wide-range classification of objects in the world from primarily unlabeled data. These include facts such as that AMGEN is a BIOTECHCOMPANY, that a PORSCHE BOXSTER is a VEHICLE and more specifically an AUTOMOBILE-MODEL, etc. It learns propositions such as these starting from only a small number of seed instances of each category, by identifying patterns observed on the web and bootstrapping from a large amount of unlabeled data. This bootstrapping is aided by constraints provided from a given ontology: for example that something cannot be both a person and a company, that an automobile is a special kind of vehicle, and so on. These constraints are used to prevent overgeneralization by classifiers for each of the categories.

Other examples of this type abound. In medical diagnosis, we often have readily available known relationships concerning the different concepts (corresponding to different diseases) we are trying to learn to predict. For example if a patient presents with chest pain, this might be due to a heart attack or gastroesophageal reflux but very unlikely both; a patient with diabetic neuropathy must also have diabetes. Note that in this setting, different diseases often have different sets of

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

relevant tests (which would lead to different features for the concepts we are trying to learn).

Another motivating example is *machine reliability*: given a machine with multiple parts, we want to detect which part failed. The ontology specifies which parts (say, gear wheel) belong to a larger part (say, engine). Again, different tests are often aimed at different parts.

Our Results We model a classifier h as an L-tuple $(h_1, \ldots, h_L)$ over the L categories, with each h_i providing an up-or-down prediction on category i and each having its own feature space. An ontology $R \subseteq \{0,1\}^L$ specifies which L-tuples of labelings are legal and which are not. We can analyze the incompatibility between a classifier and a given ontology by examining the probability the classifier assigns a labeling that violates the ontology; we call this its *unlabeled error rate*. Clearly, this is also a lower bound on its true error rate. Our first result shows that for the ontology in which each example belongs to exactly one category, if $L \geq 3$ and no category is dominant, then for a product distribution (given the labels), unlabeled error is also (up to a factor of 2) an *upper bound* on true error, for hypotheses that also produce no overly dominant category. This implies that by optimizing on unlabeled data we can find a near optimal hypothesis.

Our second result involves the ontology in which each example belongs to at *most* one category. In this case we find unlabeled error is closely related to *false-positive* error. As a result, we can achieve a near optimal hypothesis by optimizing on unlabeled data subject to both upper and lower bounds on prediction rates for each category.

Next, we extend these results to the case of a general graph-based ontology described by binary NAND and subset relationships between categories (a NAND relation between categories i, i' indicating that an object cannot be both type i and i', and a subset relation indicating that an object of type i must also be of type i'), and also relax the independence condition. We show that the classifier of minimum unlabeled error that maximizes a measure we call *tightness* (that can be estimated from unlabeled data alone) will be a good approximation to the target function. We further give efficient algorithms for the case of discrete domains, corresponding to settings where data within each view falls into a small number of high-quality clusters. Finally, we provide finite sample guarantees, extending VC-dimension to apply to unlabeled error.

Related Work This type of constraint-aided learning is reminiscent of co-training (Blum & Mitchell, 1998) and multi-view learning more generally (Sridharan & Kakade, 2008; Chapelle et al., 2006; Zhu & Goldberg, 2009); however in those settings the multiple views correspond to a single task, whereas in our framework the views correspond to different tasks related by an ontology. The use of unlabeled error is also related to the semi-supervised learning framework of (Balcan & Blum, 2005). (Cavallanti et al., 2010) study multi-task learning where the tasks have different feature spaces in the context of online *supervised* learning, and (He & Lawrence, 2011) study multi-task semi-supervised learning where the *individual* tasks involve multiple views; however, these do not use an ontology among tasks. Empirical work on using ontologies for semi-supervised and unsupervised learning in the NELL system appears in (Carlson et al., 2010a;b; Verma & Hruschka Jr., 2012), and empirical work on learning ontologies from data appears in (Mohamed et al., 2011).

2. The model

Learning model: We assume there are L *categories*. Let $X = X_1 \times \cdots X_L$ be the instance space where X_i denotes the instance space for category i. An example is an L-tuple $\vec{x} = (x_1, \ldots, x_L)$ where $x_i \in X_i$.

We have a hypothesis class C such that each classifier is $\vec{c} = (c_1, \ldots, c_L) \in C^L$ where $c_i : X_i \to \{0,1\}$. The classification of $\vec{c}$ of $\vec{x}$ is $\vec{c}(\vec{x}) = (c_1(x_1), \ldots, c_L(x_L))$. We assume that the *target function* $c^* \in C^L$. Given the target function c^*, let $X_i^+ = \{x_i \in X_i : c_i^*(x_i) = 1\}$ and $X_i^- = \{x_i \in X_i : c_i^*(x_i) = 0\}$.

We assume a joint distribution D over X. For category i, the distribution D induces on X_i is D_i. D_i^+ and D_i^- are the induced distributions on X_i^+ and X_i^-, respectively. Let $p_i = \mathbf{Pr}_D(c_i^*(x_i) = 1)$. We will often make the assumption that each category is meaningful (for all i, $p_i \geq \alpha$ for some $\alpha > 0$) and no category is dominant (e.g., $p_i \leq 1/3$ for all i).

Ontology: An ontology $R \subseteq \{0,1\}^L$ is the set of legal labelings. We assume that $c^*(\vec{x}) \in R$ for any $\vec{x} \in X$. A broad class of ontologies we will focus on are *graph based ontologies*, where the ontology is given by a graph over the L categories, where each edge corresponds to a binary relation between pairs of categories indicating which relations are disallowed. We focus specifically on two relationships: A "NAND" relationship (i, i') implies that for any $z \in R$ we never have $z_i = z_{i'} = 1$. A "$\subseteq$" relationship (i, i') implies that for any $z \in R$ if $z_i = 1$ then $z_{i'} = 1$.

2.1. Intuition and high-level idea

The high-level idea for how the ontology R enables the use of unlabeled data in our model is that it induces a

natural notion of the *unlabeled error rate* of a proposed classifer. Specifically, fixing R, given a classifier $h = (h_1, \ldots, h_L)$, we define

$$\text{err}_{unl}(h) \quad = \quad \mathbf{Pr}_{x\sim D}[(h_1(x_1), \ldots, h_L(x_L)) \notin R].$$

In other words, the unlabeled error rate of a classifier h is the probability mass of data for which it assigns an illegal labeling. Since any labeling disallowed by R must also be an incorrect vector of predictions, unlabeled error rate is a lower bound on labeled error rate; that is, $\text{err}_{unl}(h) \leq \text{err}(h)$, where

$$\text{err}(h) \quad = \quad \mathbf{Pr}_{x\sim D}[\exists i, h_i(x_i) \neq c_i^*(x)].$$

What we will show is that under appropriate assumptions we can get bounds in the other direction as well. This will imply that unlabeled error rate can be used as a proxy for labeled error rate in a global optimization procedure. It actually will be a bit more complicated than this: for example, a classifier h such that $h_1(x_1) = 1$ for every example x, and $h_i(x_i) = 0$ for all x and all $i > 1$, would be perfectly consistent with an ontology that is the complete graph of NAND edges; yet such an h would presumably have high labeled error (so in this case, $\text{err}_{unl}(h)$ is not at all upper-bounding $\text{err}(h)$). However, notice that in this case we could discard such an h due to violating an assumption that no category is overly dominant. What we will show is that this, and a few other cases that can also be addressed from unlabeled data only, are the only types of obstacles that arise, at least for the portion of $\text{err}(h)$ caused by overgeneralizing. This motivates algorithms that aim to generalize as much as possible subject to the ontology and category constraints.

Of course, the ease or difficulty of optimizing for low unlabeled error rate depends on the form of the classifiers. After discussing general results under the assumption of an optimization oracle, we then consider a more specific cluster-based setting where we will be able to solve this problem efficiently.

3. Optimization-based learning

In this section we analyze the relationship between unlabeled and labeled error rates under natural conditions. We begin by assuming data satisfies independence given the labeling (IGL). In Section 3.3 we will relax that assumption to a notion of weak dependence.

We begin by considering the ontology R_1 where every example is positive for *exactly* one of the L categories. For this ontology we have $\sum_i p_i = 1$ and under IGL we can view the generation of examples as follows. Examples are drawn by first selecting a label $i \sim (p_1, \ldots, p_L)$, and then selecting $x_{i'} \sim D_{i'}^-$ for each $i' \neq i$ and selecting $x_i \sim D_i^+$. We will then relax this to a complete graph of NAND edges (every example positive for *at most* one category), i.e., $R_{01} = \{z \in \{0,1\}^L, \sum_i z_i \leq 1\}$. Finally we will relax it to an *arbitrary* ontology of NAND and "$\subseteq$" edges, subject only to very mild conditions such as each category having at least one incident edge and each relation being nontrivial (see Section 3.3).

For a hypothesis $h = (h_1, \ldots, h_L)$, for $a, b \in \{0,1\}$, we define $p_{i,a|b}$ as the probability that $h_i(x_i) = a$ given that $c_i^*(x_i) = b$.

3.1. Each example positive for exactly one category

For ontology R_1, where each example is positive for exactly one category, we have: $\text{err}_{unl}(h) = \mathbf{Pr}_{x\sim D}[h_i(x) = 0 \text{ for all } i \text{ OR } \exists\, i \neq i' \text{ s.t. } h_i(x) = 1, h_{i'}(x) = 1]$.

As a warmup we begin by analyzing the case that the different views all look like identical copies of the same learning problem. In particular, $D_1 = \ldots = D_L$ and $c_1^* = \ldots = c_L^*$, and D satisfies independence given the labeling. We call this the *independent-identical* model. In this case we can shorten the notation $p_{i,a|b}$ to simply $p_{a|b}$. We will show that for ontology R_1, if $\text{err}(h) \geq \epsilon$ then $\text{err}_{unl}(h) = \Omega(\epsilon)$. That is, all the hypotheses of large labeled error rate have large unlabeled error rate as well, and thus these hypotheses can be discarded by using unlabeled examples.

To get an intuition of why this might be true note that

$$\begin{aligned} \text{err}(h) \quad &\leq \quad \sum_{i\in\{1,\ldots,L\}} p_i(p_{i,0|1} + \sum_{i'\neq i} p_{i',1|0}) \\ &= \quad p_{0|1} + (L-1)p_{1|0} \equiv \alpha. \end{aligned}$$

Note that either $p_{0|1} \geq \alpha/2$ or $p_{1|0} \geq \frac{\alpha}{2(L-1)}$. For simplicity assume below that $\alpha \leq 1$.[1]

We will consider a few cases based on $p_{1|0}$ and $p_{0|1}$. The first case is $p_{1|0} \geq \frac{\alpha}{2(L-1)}$ and $p_{0|1} \leq \alpha/2$. In this case we will consider the probability of having at least two positive predictions, which is,

$$(1-p_{0|1})(1-(1-p_{1|0})^{L-1}) \geq \tfrac{1}{2}(1-(1-\tfrac{\alpha}{4})) = \tfrac{\alpha}{8}.$$

Now we consider the case that $p_{0|1} \geq \alpha/2$. Consider the unlabeled accuracy, namely the probability of having a single positive label,

$$\begin{aligned} 1 - \text{err}_{unl} \quad &\leq \quad (L-1)p_{0|1}p_{1|0} + (1-p_{0|1}) \\ &= \quad (\alpha - p_{0|1})p_{0|1} + (1-p_{0|1}). \end{aligned}$$

[1] If $\alpha > 1$ then a different argument using the fact that $L \geq 3$ can be applied.

Since this is decreasing with $p_{0|1}$, the unlabeled accuracy is at most $1-(\alpha/2)+(\alpha/2)(\alpha-\alpha/2)$. So, the unlabeled error rate is at least $\alpha/2-\alpha^2/4 \geq \alpha/4 \geq \mathrm{err}(h)/4$. We have derived the following theorem.

Theorem 1 *In the independent-identical model with ontology R_1 and $L \geq 3$, if $\mathrm{err}(h) \geq \epsilon$ then $\mathrm{err}_{unl}(h) = \Omega(\epsilon)$.*

We now remove the (artificial) restriction that all the D_i^+ and D_i^- are equal. The requirement in Theorem 2 that $p_i \leq 1/3$ can be viewed as requiring that we need at least three non-trivial categories.

Theorem 2 *For ontology R_1, if D satisfies IGL and we have $p_i \leq 1/3$ and $\mathbf{Pr}[h_i = 1] \leq 1/3$ for all i, then $\mathrm{err}_{unl}(h) \geq \mathrm{err}(h)/2$.*

Proof: Let us write the labeled error rate as:

$$\mathrm{err}(h) = 1 - \sum_i p_i p_{i,1|1} \prod_{i' \neq i} p_{i',0|0}$$

The unlabeled error rate is,

$$\mathrm{err}_{unl}(h) = \mathrm{err}(h) - \sum_i p_i p_{i,0|1} \sum_{i' \neq i} p_{i',1|0} \prod_{j \notin \{i',i\}} p_{j,0|0}$$

Intuitively, the difference between the unlabeled error rate and the true error rate are the cases when we have an outcome with a single one, but it is the incorrect one. We would like to lower bound the ratio between the unlabeled error rate and the true error rate, i.e.,

$$\frac{\mathrm{err}_{unl}(h)}{\mathrm{err}(h)} = 1 - \frac{\sum_i p_i p_{i,0|1} \sum_{i' \neq i} p_{i',1|0} \prod_{j \notin \{i',i\}} p_{j,0|0}}{1 - \sum_i p_i p_{i,1|1} \prod_{i' \neq i} p_{i',0|0}}$$

Let $x_i = p_{i,0|1}$ and $y_{i'} p_{i',1|0}$, then we like to maximize

$$F(x_1, y_1, \ldots, x_L, y_L) = \frac{\sum_i p_i x_i \sum_{i' \neq i} y_{i'} \prod_{j \notin \{i',i\}} (1-y_j)}{1 - \sum_i p_i (1-x_i) \prod_{i' \neq i} (1-y_{i'})}$$

For each i, the function F is increasing in x_i, so it will get its maximum value at 1 (the maximum value of x_i). This implies that we need to maximize

$$G(y_1, \ldots, y_L) = \sum_i p_i \sum_{i' \neq i} y_{i'} \prod_{j \notin \{i',i\}} (1-y_j).$$

Since each $p_i \leq 1/3$ and h_i is 1 with probability at most $1/3$ then $p_{i,1|0} \leq 1/2$.

We know that the maximizing solution to G out of all $y_i \in [0, 1/2]$ will have a certain number, say k, of y's equal $1/2$ and the rest are zero. Therefore,

$$G \leq \sum_{i \in S} p_i k \frac{1}{2} (1 - \frac{1}{2})^{k-1} = \sum_{i \in S} p_i k \frac{1}{2^k} \leq \frac{1}{2}.$$

This implies that $\mathrm{err}_{unl}(h)/\mathrm{err}(h) \geq \frac{1}{2}$. ■

3.2. Each example positive for at most one category

We now consider the more general ontology R_{01}, which allows for the all-negative labeling. We start with the following important lemma, that relates the error of switching 0 to 1, with some 'observable event'.

Lemma 3 *Given ontology R_{01}, assume D satisfies IGL and that $\sum_i p_i = 1 - \beta \geq 3/4$, and $p_i \leq 1/4$ for all i. Consider a hypothesis h such that $\sum_i p_{i,1|0} = \gamma$. Then, for any $\delta \leq 1/4$ one of the following holds: (1) the probability of two or more positive labels is at least $\gamma^2(1-\beta-\delta)/6000$ (2) The probability the hypothesis h predicts all categories negative is more than $\beta + \delta$, or (3) for some category i, the probability that $h_i(x_i) = 1$ is at least $0.982(1-\beta-\delta)$*

Proof: Assume that the probability that h predicts all categories negative is at most $\beta + \delta$ (otherwise we are done). Let $s = \arg\max_i p_{i,1|0}$. We consider two cases depending on the magnitude of $p_{s,1|0}$. In the first case, where $p_{s,1|0} < \gamma/6$, we show that the unlabeled error rate is significant. In the second case, where $p_{s,1|0} \geq \gamma/6$, we show that either the unlabeled error rate is significant or that h predicts positive in category s with a very high probability.

Case 1: Assume $p_{s,1|0} < \gamma/6$. Then we can partition the categories into three subsets S_1, S_2 and S_3 such that $\sum_{i \in S_j} p_{i,1|0} \geq \gamma/6$. Let $B_j = \sum_{i \in S_j} p_{i,1|0}$. For one of the three sets we have $\sum_{i \in S_j} p_i \geq (1-\beta-\delta)/3$; assume that this is S_1. Then the probability that the true label is in S_1 and in each other set we have at least one positive label (and hence at least two positive labels overall) is at least,

$$\begin{aligned}
&(\sum_{i \in S_1} p_i)(1 - \prod_{i \in S_2} (1 - p_{i,1|0}))(1 - \prod_{i \in S_3} (1 - p_{i,1|0})) \\
&\geq (\textstyle\sum_{i \in S_1} p_i)(\sum_{i \in S_2} \frac{p_{i,1|0}}{e})(\sum_{i \in S_3} \frac{p_{i,1|0}}{e}) \\
&\geq \tfrac{1-\beta-\delta}{3} \tfrac{\gamma}{6e} \tfrac{\gamma}{6e} = \tfrac{(1-\beta-\delta)\gamma^2}{108e^2} > \tfrac{(1-\beta-\delta)\gamma^2}{6000},
\end{aligned}$$

where the first inequality uses the fact that having k independent events of probability q_j each implies that the probability some event occurs is at least $\sum_j q_j/e$. This implies that the probability of an unlabeled error in this case is $\Omega(\gamma^2(1-\beta-\delta))$.

Case 2: Assume that $p_{s,1|0} \geq \gamma/6$. Then $\gamma \leq 6$. Consider the probability of observing a 1 in s and also a label 1 in some $i \neq s$. Note that the probability that $c_s^*(x) = 0$ is at least $3/4$. The only way the unlabeled

error rate can be less than $\gamma^2(1-\beta-\delta)/6000$ is if

$$\mathbf{Pr}[\exists i \neq s : h_i(x) = 1 | c_s^*(x) = 0] < \frac{\gamma(1-\beta-\delta)}{1000(3/4)}.$$

Namely, given that $c_s^*(x) = 0$, the probability of predicting 1 in any $i \neq s$ is at most the above bound. Since $\gamma \leq 6$ this implies that the probability that $h_s = 1$ is at least $1 - \mathbf{Pr}[\forall i h_i(x) = 0] - \mathrm{err}_{unl}(h) - \frac{6(1-\beta-\delta)}{1000(3/4)} > 0.982(1-\beta-\delta)$. ■

Unlike the case of the ontology R_1, for R_{01} there can be a hypothesis that has zero unlabeled error and a very high labeled error: for example, a hypothesis that predicts negative always for every category. To overcome this, we will focus on a set of "plausible" hypotheses. Let $Good(\beta, \lambda)$ be the set of hypotheses $h \in C$ such that $\mathbf{Pr}[\forall i h_i(x_i) = 0] \leq \beta$ and for any category $\mathbf{Pr}[h_i(x_i) = 1] \leq \lambda$.

Theorem 4 *Suppose $\sum_i p_i = 1-\beta \geq 4/5$ and for any i we have $p_i \leq \lambda = 1/5$. If $h \in Good(\beta+\epsilon/20, \lambda+\epsilon/20)$ and $\mathrm{err}(h) \geq \epsilon$ then $\mathrm{err}_{unl}(h) = \Omega(\epsilon^2)$. In addition $Good(\beta+\epsilon/20, \lambda+\epsilon/20) \neq \emptyset$.*

Proof: First, note that $c^* \in Good(\beta, \lambda)$. Now, consider $h \in Good(\beta+\epsilon/20, \lambda+\epsilon/20)$ and let $\sum_i p_{i,1|0} = \gamma$. Assume that $\mathrm{err}(h) \geq \epsilon$. We consider two cases depending on the value of γ.

Case 1: Assume $\gamma \geq 0.25\epsilon$. By Lemma 5, with $\delta = \epsilon/20$, we have three possible outcomes: (1) the probability of two or more positive labels is $\gamma^2(1-\beta-\epsilon/20)/6000$, (2) The probability that h predicts all categories negative is more than $\beta + \epsilon/20$, or (3) for some category i, the probability that $h_i(x_i) = 1$ is at least $0.982(1-\beta-\epsilon/20)$.

Since $h \in Good(\beta+\epsilon/20, \lambda+\epsilon/20)$, outcome (2) is impossible. For outcome (3), since $\beta + \epsilon/20 \leq 1/5 + 1/20 = 1/4$, then $0.982(1-\beta-\epsilon/20) > 0.7$. Since the probability that $h \in Good(\beta+\epsilon/10, \lambda+\epsilon/10)$ labels any category positive is at most $\lambda + \epsilon/20 = 1/4 < 0.7$, outcome (3) is also impossible. Therefore, the unlabeled error rate of h is $\Omega(\gamma^2(1-\beta-\epsilon/20)) = \Omega(\epsilon^2)$.

Case 2: Assume $\gamma < 0.25\epsilon$. Since $\mathrm{err}(h) \geq \epsilon$, this implies that $\xi = \sum_i p_i p_{i,0|1} \geq 0.75\epsilon$. The probability that h predicts negative in all categories is at least $\beta - \gamma + \xi > \beta - 0.25\epsilon + 0.75\epsilon = \beta + 0.5\epsilon$, which is a contradiction that $h \in Good(\beta+\epsilon/20, \lambda+\epsilon/20)$. ■

3.3. General Graph-Based Ontologies

The previous analysis demonstrated an extremely tight connection between labeled and unlabeled error, allowing for learning from unlabeled examples only, and moreover (see Section 4) from a number of unlabeled examples not much larger than the sample complexity of *supervised* learning. However, these results required an ontology such that any example is positive for at most one category. They also assumed full independence given the labeling. In this section we extend these results to show we can learn from purely unlabeled data given an *arbitrary* graph based ontology of NAND and "$\subseteq$" edges, subject only to the following conditions on the ontology and the distribution:

(1) For some given $\alpha > 0$, we have $\alpha \leq p_i \leq 1-\alpha$ for all i. That is, each category is nontrivial.

(2) The ontology has no isolated vertices. That is, for every category i, there exists at least one edge in the ontology graph (either NAND or "$\subseteq$") with i as one of its endpoints.

(3) For some given $\alpha_2 > 0$, for any edge (i, i') in the ontology graph, each of the pairs of labels not explicitly *disallowed* by the edge has probability mass at least α_2 under D.

For example, if there is an edge "athlete $\subseteq$ person", then each of the three cases: "person-athlete", "person-nonathlete", and "nonperson-nonathlete" should have probability at least α_2. This condition is natural because if it is not satisfied for some edge, then it means the two categories are effectively either synonyms (nearly always equal) or antonyms (nearly always opposite).

Finally, we replace independence given the labeling with the following much weaker condition related to the "weak dependence" of (Abney, 2002):

(4) For some given $\lambda > 0$, for any edge (i, i'), any $h_i, h_{i'} \in C$, and any $a_i, \ell_i, a_{i'}, \ell_{i'} \in \{0, 1\}$, we have

$$\mathbf{Pr}[h_i = a_i | c_i^* = \ell_i \wedge c_{i'}^* = \ell_{i'} \wedge h_{i'} = a_{i'}] \geq \lambda \mathbf{Pr}[h_i = a_i | c_i^* = \ell_i]$$

so long as $\mathbf{Pr}[c_i^* = \ell_i \wedge c_{i'}^* = \ell_{i'} \wedge h_{i'} = a_{i'}] > 0$. (For compactness, we are using "$h_i = a_i$" to denote the event "$h_i(x_i) = a_i$" and similarly for $h_{i'}, c_i^*, c_{i'}^*$). Note that full (or even pairwise) independence would correspond to $\lambda = 1$.

For additional intuition, conditions (2) and (3) above would be satisfied if every category were contained in a triangle of NAND edges.[2]

[2]Specifically, if the ontology graph satisfies this property, then after deleting any edges that do not appear in

Under the above conditions we will still be able to learn from unlabeled data only. However, we will need a larger sample size than for a complete graph.

In order to describe the algorithm and present its analysis, we need a few more definitions. First, if category i has either an incident NAND edge or an incident "$\subseteq$" edge for which it is on the subset-side, say that it is *positive-constrained* (because the ontology is limiting when it can be positive). If category i has an incident "$\subseteq$" edge for which it is on the superset-side, we say that it is *negative-constrained* (because the ontology is limiting when it can be negative). Note that by condition (2) above, for every category at least one applies. If both cases apply to some category i, then for convenience we will just call it positive-constrained. Let cat^+ denote the set of positive-constrained categories and cat^- denote the set of negative-constrained categories. We now define how *tightly constrained* a hypothesis $h = (h_1, \ldots, h_L)$ is as:

$$tightness(h) = \sum_{i \in cat^+} \mathbf{Pr}_{x \sim D}[h_i(x_i) = 1] + \sum_{i \in cat^-} \mathbf{Pr}_{x \sim D}[h_i(x_i) = 0].$$

Finally, for a NAND edge (i, i') define $\mathrm{err}_{unl}(h, i, i') = \mathbf{Pr}_D[h_i(x_i) = 1 \wedge h_{i'}(x_{i'}) = 1]$ and for a "$\subseteq$" edge (i, i') define $\mathrm{err}_{unl}(h, i, i') = \mathbf{Pr}_D[h_i(x) = 1 \wedge h_{i'}(x_{i'}) = 0]$.

The optimization procedure for learning from unlabeled data is now simply this: given an unlabeled sample S, find the most tightly constrained hypothesis h possible (maximizing $tightness(h)$) subject to satisfying $\alpha \leq \mathbf{Pr}[h_i(x_i) = 1] \leq 1 - \alpha$ for all i and h having zero empirical unlabeled error over S. To analyze this procedure, we begin with two key lemmas.

Lemma 5 *For a NAND edge (i, i'), for any h:*

$$\mathrm{err}_{unl}(h, i, i') \geq \alpha_2 \lambda^2 \cdot p_{i,1|0} \cdot \mathbf{Pr}[h_{i'}(x_{i'}) = 1].$$

Proof: We will expand $\mathrm{err}_{unl}(h, i, i')$ by looking at two specific cases for $c_i^*(x_i)$ and $c_{i'}^*(x_{i'})$ and then will apply conditions (3) and (4) of our assumptions to achieve our desired lower bound. For compactness of notation, let "$c_{ii'}^* = ab$" denote the event "$c_i^*(x_i) = a \wedge c_{i'}^*(x_{i'}) = b$", and similarly for $h_{ii'}$. Then we have:

$$\begin{aligned}
&\mathrm{err}_{unl}(h, i, i') \\
&\quad = \mathbf{Pr}[h_i = 1 \wedge h_{i'} = 1] \\
&\quad \geq \mathbf{Pr}[c_{ii'}^* = 01] \cdot \mathbf{Pr}[h_i = 1 | c_{ii'}^* = 01] \cdot \\
&\qquad \mathbf{Pr}[h_{i'} = 1 | c_{ii'}^* = 01 \wedge h_i = 1] + \\
&\qquad \mathbf{Pr}[c_{ii'}^* = 00] \cdot \mathbf{Pr}[h_i = 1 | c_{ii'}^* = 00] \cdot \\
&\qquad \mathbf{Pr}[h_{i'} = 1 | c_{ii'}^* = 00 \wedge h_i = 1] \\
&\quad \geq \alpha_2(\lambda \mathbf{Pr}[h_i = 1 | c_i^* = 0] \cdot \lambda \mathbf{Pr}[h_{i'} = 1 | c_{i'}^* = 1]) + \\
&\qquad \alpha_2(\lambda \mathbf{Pr}[h_i = 1 | c_i^* = 0] \cdot \lambda \mathbf{Pr}[h_{i'} = 1 | c_{i'}^* = 0]) \\
&\quad \geq \alpha_2 \lambda^2 p_{i,1|0} \mathbf{Pr}[h_{i'} = 1].
\end{aligned}$$

The first inequality above is from considering just two of the three possible settings of $c_i^*(x_i)$ and $c_{i'}^*(x_{i'})$, the second comes from applying condition (3) to the events involving c and (4) to the events involving h, and finally the last inequality is using $\mathbf{Pr}[h_{i'} = 1] \leq \mathbf{Pr}[h_{i'} = 1 | c_{i'}^* = 0] + \mathbf{Pr}[h_{i'} = 1 | c_{i'}^* = 1]$. ■

Lemma 6 *For a "$\subseteq$" edge (i, i'), for any h we have:*

$$\mathrm{err}_{unl}(h, i, i') \geq \alpha_2 \lambda^2 \cdot p_{i,1|0} \cdot \mathbf{Pr}[h_{i'}(x_{i'}) = 0]$$

$$\mathrm{err}_{unl}(h, i', i) \geq \alpha_2 \lambda^2 \cdot p_{i',0|1} \cdot \mathbf{Pr}[h_i(x_i) = 1].$$

Proof: Using conditions (3) and (4) as in the proof of Lemma 5, we can expand $\mathrm{err}_{unl}(h, i, i')$ as:

$$\begin{aligned}
&\mathrm{err}_{unl}(h, i, i') \\
&\quad = \mathbf{Pr}[h_i = 1 \wedge h_{i'} = 0] \\
&\quad \geq \mathbf{Pr}[c_{ii'}^* = 01] \cdot \mathbf{Pr}[h_i = 1 | c_{ii'}^* = 01] \cdot \\
&\qquad \mathbf{Pr}[h_{i'} = 0 | c_{ii'}^* = 01 \wedge h_i = 1] + \\
&\qquad \mathbf{Pr}[c_{ii'}^* = 00] \cdot \mathbf{Pr}[h_i = 1 | c_{ii'}^* = 00] \cdot \\
&\qquad \mathbf{Pr}[h_{i'} = 0 | c_{ii'}^* = 00 \wedge h_i = 1] \\
&\quad \geq \mathbf{Pr}[c_{ii'}^* = 01] \cdot \lambda \mathbf{Pr}[h_i = 1 | c_i^* = 0] \cdot \\
&\qquad \lambda \mathbf{Pr}[h_{i'} = 0 | c_{i'}^* = 1] + \\
&\qquad \mathbf{Pr}[c_{ii'}^* = 00] \cdot \lambda \mathbf{Pr}[h_i = 1 | c_i^* = 0] \cdot \\
&\qquad \lambda \mathbf{Pr}[h_{i'} = 0 | c_{i'}^* = 0] \\
&\quad \geq \alpha_2 \lambda^2 p_{i,1|0} \mathbf{Pr}[h_{i'} = 0].
\end{aligned}$$

We also similarly have:

$$\begin{aligned}
&\mathrm{err}_{unl}(h, i, i') \\
&\quad = \mathbf{Pr}[h_i = 1 \wedge h_{i'} = 0] \\
&\quad \geq \mathbf{Pr}[c_{ii'}^* = 01] \cdot \lambda \mathbf{Pr}[h_i = 1 | c_i^* = 0] \cdot \\
&\qquad \lambda \mathbf{Pr}[h_{i'} = 0 | c_{i'}^* = 1] + \\
&\qquad \mathbf{Pr}[c_{ii'}^* = 11] \cdot \lambda \mathbf{Pr}[h_i = 1 | c_i^* = 1] \cdot \\
&\qquad \lambda \mathbf{Pr}[h_{i'} = 0 | c_{i'}^* = 1] \\
&\quad \geq \alpha_2 \lambda^2 p_{i',0|1} \mathbf{Pr}[h_i(x_i) = 1].
\end{aligned}$$

■

We now use these to show that for any plausible hypothesis h (one such that $\alpha \leq \mathbf{Pr}[h_i(x_i) = 1] \leq 1 - \alpha$

triangles, the resulting ontology satisfies (2) and (3) with $\alpha_2 = \alpha$. That is because for any edge (i, i') there is at least α probability of labeling $(1, 0)$ (when $c_i^*(x_i) = 1$), at least α probability of labeling $(0, 1)$ (when $c_{i'}^*(x_{i'}) = 1$) and at least α probability of labeling $(0, 0)$ (when $c_{i''}^*(x_{i''}) = 1$ where i'' is the third category in the triangle).

for all i), if h has high error then either its unlabeled error will be high or else its tightness will be low compared to the target function. This then will imply that maximizing tightness subject to low unlabeled error must lead to a hypothesis of low labeled error.

Theorem 7 *Suppose* $\alpha \leq \mathbf{Pr}[h_i(x_i) = 1] \leq 1 - \alpha$ *for all* i. *If* $\mathrm{err}(h) \geq \epsilon$ *then either* $\mathrm{err}_{unl}(h) \geq \alpha\alpha_2\lambda^2\epsilon/(4L)$ *or tightness*$(h) \leq$ *tightness*$(c) - \epsilon/2$.

Proof: Suppose $\mathrm{err}(h) \geq \epsilon$. We have three cases.

Case 1: Suppose for some positive-constrained category i, we have $p_{i,1|0} \geq \frac{\epsilon}{4L}$. By definition of positive-constrained, there must be some NAND or "$\subseteq$" edge (i, i'). So, by Lemmas 5 and 6, we have $\mathrm{err}_{unl}(h) \geq \mathrm{err}_{unl}(h, i, i') \geq \alpha_2\lambda^2(\frac{\epsilon}{4L})\alpha = \alpha\alpha_2\lambda^2\epsilon/(4L)$.

Case 2: Suppose for some negative-constrained category i', we have $p_{i',0|1} \geq \frac{\epsilon}{4L}$. By definition of negative-constrained, there must be some "$\subseteq$" edge (i, i'). So, by Lemma 6, we again have $\mathrm{err}_{unl}(h) \geq \mathrm{err}_{unl}(h, i, i') \geq \alpha_2\lambda^2(\frac{\epsilon}{4L})\alpha = \alpha\alpha_2\lambda^2\epsilon/(4L)$.

Case 3: Suppose neither of the above two cases occurs. Define $p_{i,1,0} = \mathbf{Pr}[h_i = 1 \wedge c_i^* = 0] \leq p_{i,1|0}$ and similarly $p_{i,1,0} = \mathbf{Pr}[h_i = 0 \wedge c_i^* = 1] \leq p_{i,0|1}$. We then have:

$$\textstyle\sum_{i \in cat^+} p_{i,1,0} + \sum_{i \in cat^-} p_{i,0,1} \leq \epsilon/4.$$

Therefore, since $\mathrm{err}(h) \geq \epsilon$,

$$\textstyle\sum_{i \in cat^+} p_{i,0,1} + \sum_{i \in cat^-} p_{i,1,0} \geq 3\epsilon/4.$$

Putting these together we have:

$$\begin{aligned} &tightness(c) - tightness(h) \\ &= \sum_{i \in cat^+} (p_{i,0,1} - p_{i,1,0}) + \sum_{i \in cat^-} (p_{i,1,0} - p_{i,0,1}) \\ &\geq \epsilon/2, \end{aligned}$$

as desired. ■

3.4. Implications

We now show how the above results can be used to analyze an idealized form of a natural iterative learning algorithm motivated by algorithms used in systems such as NELL (Carlson et al., 2010a;b).

We consider here the same asumptions (1-4) used in Section 3.3 and additionally assume each category is incident to at least one NAND edge. Define a hypothesis h_i to be "safe" if $h_i \subseteq c_i^*$, i.e., $p_{i,1|0} = 0$. Say that g_i is an "α_1-extension" of h_i if $h_i \subseteq g_i$ and $\mathbf{Pr}[g_i = 1] \geq \mathbf{Pr}[h_i = 1] + \alpha_1$. We now consider the following idealized iterative learning procedure. Let $\alpha_0, \alpha_1 > 0$ and $k \in Z^+$. Suppose:

1. Given an initial set of positive *labeled* examples for each category, the algorithm is able to produce safe initial hypotheses h_i that are nontrivial in that $\mathbf{Pr}(h_i(x_i) = 1) \geq \alpha_0$ for each i.
2. For any category i such that $p_{i,0|1} \geq \alpha_1$ (i.e., we are not yet done), the algorithm is able to propose up to k α_1-extensions $g_i^1, \ldots, g_i^k$ of h_i such that at least one is safe and any non-safe g_i^j satisfies $p_{i,1,0} = \Omega(\alpha_1)$ (i.e., none are "just barely" safe).
3. The procedure then is just to greedily replace any h_i with any extension g_i^j found that does not incur unlabeled error over NAND edges until no such extension of any h_i exists.

Corollary 8 *The above procedure correctly maintains safe hypotheses and halts after at most* L/α_1 *steps with a hypothesis of total error at most* $L\alpha_1$.

Proof: Assume inductively that $h_1, \ldots, h_L$ are safe and suppose g_i^j is not a safe extension of h_i. We are given that category i is connected by a NAND edge to at least one other category i'. Since $\mathbf{Pr}[h_{i'} = 1] \geq \alpha_0$, by Lemma 5 we have $\mathrm{err}_{unl}(h, i, i') \geq \alpha_2\lambda^2\alpha_1\alpha_0$. Thus any unsafe extension will not be taken. On the other hand, if g_i^j *is* a safe extension, then replacing h_i with g_i^j maintains the inductive property that $h_1, \ldots, h_L$ are safe and thus no *NAND* edge will incur unlabeled error. So the algorithm will not halt until no safe extension exists, which happens only if total error $\leq L\alpha_1$. ■

4. Sample Complexity

In this section we derive bounds relating empirical unlabeled error to true unlabeled error, in order to apply the results of Section 3 to finite sample sizes. To do so we will analyze the VC dimension of the unlabeled error sets.

We start with the identical-independent model. Recall that in this setting we select only a single hypothesis $h \in C$. We need to compute the $\mathrm{VCdim}(C, L)$ which is the VC dimension when we have L points for each example. Recall that a point is an L-tuple $\vec{x} = (x_1, \ldots, x_L)$. We first need to define $\ell(h, \vec{x})$, the loss of h on $\vec{x}$. Let R be an ontology, i.e., the set of legal labelings. We define $\ell(h, \vec{x}) = 0$ if $(h(x_1), \ldots, h(x_L)) \in R$, and otherwise it is 1.

Theorem 9 *Let* $d = \mathrm{VCdim}(C)$. *For any ontology* R *we have* $\mathrm{VCdim}(C, L) = O(d \log L)$.

Proof: Consider a set $T = \{\vec{x_1}, \ldots, \vec{x}_m\}$ that C shatters. Let $S = \{x | \exists i, x \in \vec{x}_i\}$ be the set of inputs

which appear in some category. Clearly, $|S| \leq Lm$. By Sauer's lemma we have that the number of different labelings of the set S is at most $(mL)^d$. Any such labeling can induce at most one labeling on the points in T. Therefore, for T to be shattered we need that $2^m \leq (mL)^d$, which implies that $m = O(d \log L)$. ■

It is worth noting that there is indeed a difference between $\text{VCdim}(C, L)$ and $\text{VCdim}(C)$.

Claim 1 *Consider the class TH_1 the threshold functions on $[0, 10]$. Then $\text{VCdim}(TH_1) = \text{VCdim}(TH_1, 1) = 1$ and $\text{VCdim}(TH_1, 2) = 2$.*

Proof: We show that for $L = 2$, TH_1 shatters 2 points. Consider $x_1 = (0.5, 4)$ and $x_2 = (2, 6)$. For $\theta = 8$ the labeling is $(1, 1)$, for $\theta = 1$ the labeling is $(1, 0)$, for $\theta = 5$ the labeling is $(0, 1)$, and for $\theta = 0$ the labeling is $(0, 0)$. Therefore TH_1 shatters $\{x_1, x_2\}$. The fact that no three points is shattered can be shown by case analysis. ■

We can now use Theorem 9 to derive generalization bounds based on Theorem 1. The result follows since the observed unlabeled error rate is close to the true unlabeled error rate, given our bound on the VC dimension.

Corollary 10 *In the independent-identical model with ontology R_1 and $L \geq 3$, a sample of $m_u = O\left(\frac{1}{\epsilon}\left(\text{VCdim}(C) \log L \ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$ unlabeled examples (and no labeled examples) are sufficient to achieve error $\leq \epsilon$. In particular, with probability $1-\delta$, all h with $h_1 = \ldots = h_L \in C$ satisfying $\widehat{\text{err}}_{unl}(h) = 0$ have $\text{err}(h) \leq \epsilon$.*

We now need to derive a generalization bound for the general case, that will relate the observed and the true unlabeled error rates. The main difference is that we consider the hypothesis class C^L where each hypothesis is $\vec{h} = (h_1, \ldots, h_L)$ and $h_i \in C$. As before, the examples are L-tuples $\vec{x} = (x_1, \ldots, x_L)$ and the loss is ℓ defined using the set R of legal labelings. (The theorem holds for *any* ontology R.) We need to compute the $\text{VCdim}(C^L, L)$.

Theorem 11 *Let $d = \text{VCdim}(C)$. For any ontology R we have $\text{VCdim}(C^L, L) = O(dL \log dL)$.*

Proof: Consider a set $T = \{\vec{x_1}, \ldots, \vec{x}_m\}$ that C^L shatters, where $\vec{x}_i = (x_{i,1}, \ldots, x_{i,L})$. Let $T_j = \{x_{i,j} : 1 \leq j \leq L\}$ be the set of inputs which appear in category i. By Sauer's lemma we have that the number of different labelings by C for T_j is at most m^d, and hence the total number of labelings for T using C^L is at most m^{dL}. Therefore, for T to be shattered we need that $2^m \leq m^{dL}$, which implies that $m = O(dL \log dL)$. ■

We can therefore derive the following.

Corollary 12 *Consider ontology R_1, and assume $p_i \leq 1/3$ and $\mathbf{Pr}[h_i = 1] \leq 1/3$ for all i. Then, $m_u = O\left(\frac{1}{\epsilon}\left(dL \log(dL) \ln\left(\frac{1}{\epsilon}\right) + \ln(\frac{1}{\delta})\right)\right)$ unlabeled examples (and no labeled examples) are sufficient to achieve error $\leq \epsilon$, where $d = \text{VCdim}(C)$. In particular, with probability $1-\delta$, all $h \in C^L$ with $\widehat{\text{err}}_{unl}(h) = 0$ have $\text{err}(h) \leq \epsilon$.*

We now consider the difference between the observed $tightness(h)$, denoted by $\widehat{tightness}(h)$, and the true $tightness(h)$. The following theorem follows by considering a generalization bound for each category, independently, and using the union bound over categories.

Theorem 13 *For an unlabeled sample size*

$$m_u = O\left(\frac{L^2}{\epsilon^2}\left(\text{VCdim}(C) \ln\left(\frac{1}{\epsilon}\right) + \ln(\frac{L}{\delta})\right)\right),$$

we have that with probability $1 - \delta$, for any $h \in C$ we have $|tightness(h) - \widehat{tightness}(h)| \leq \epsilon$.

5. Discrete domains

In this section, we assume the different domains X_i are discrete and small: specifically, $X_i = \{1, \ldots, N\}$, where N is not too large, and let class C consist of all boolean functions over N elements. For example, this could model data satisfying the condition that points within each view can be easily grouped into at most N clusters, with each cluster having a single label. Note that even though we view N as small, the total number of examples possible is N^L and could be quite large. Our goal here will be to learn from unlabeled data using time and samples polynomial in N and L.

We will assume here the same conditions (1-4) used in Section 3.3, namely (1) each category has at least some nonnegligible probability mass α, (2) the ontology graph has no isolated vertices, (3) for any edge (i, i'), any pair of labels not *disallowed* by the ontology has probability mass at least α_2, and (4) weak dependence. We begin with the following lemma.

Lemma 14 *For any edge (i, j) of the ontology, any pair of labels ℓ_i, ℓ_j not disallowed by this edge, and any pair of values a_i, a_j s.t. $c_i^*(a_i) = \ell_i$ and $c_j^*(a_j) = \ell_j$:*

$$\mathbf{Pr}[x_i = a_i, x_j = a_j | c_i^* = \ell_i, c_j^* = \ell_j] \geq \lambda^2 \mathbf{Pr}[x_i = a_i | c_i^* = \ell_i] \cdot \mathbf{Pr}[x_j = a_j | c_j^* = \ell_j].$$

Proof: Since C contains all boolean functions over $\{1,\ldots,N\}$, using condition (4) we have:

$$\begin{aligned}
&\mathbf{Pr}[x_i = a_i, x_j = a_j | c_i^* = \ell_i, c_j^* = \ell_j] \\
&\quad = \mathbf{Pr}[x_i = a_i | c_i^* = \ell_i, c_j^* = \ell_j] \cdot \\
&\qquad \mathbf{Pr}[x_j = a_j | c_i^* = \ell_i, c_j^* = \ell_j, x_i = a_i] \\
&\quad \geq \lambda^2 \mathbf{Pr}[x_i = a_i | c_i^* = \ell_i] \cdot \mathbf{Pr}[x_j = a_j | c_j^* = \ell_j].
\end{aligned}$$

■

We now present the algorithm, using Lemma 14. The idea is simple. First, we sample enough unlabeled examples. Next, examine each edge (i,j) of the ontology graph. For each such edge, if two values that individually have reasonably high probability mass never co-occur, then this must be because they form the disallowed labeling for that edge (e.g., both are positive if it is a NAND edge or the first is positive and the second is negative for a subset edge).

Algorithm LearnDiscrete:
Given error bound ϵ, confidence bound δ, values α, α_2:

1. Let $\epsilon' = \min(\epsilon, \alpha L)$. Draw a sample S of $m = O(\lambda^{-2}\alpha_2^{-1}(NL/\epsilon')^2 \log(NL/\delta))$ examples.

2. Define $\hat{p}_i(a_i)$ to be the empirical frequency of $a_i \in X_i$, i.e., $\frac{1}{m}\sum_{x \in S} I(x_i = a_i)$. Similarly, $\hat{p}_{ij}(a_i, a_j)$ is the empirical frequency of the pair $(a_i, a_j) \in X_i \times X_j$, i.e., $\frac{1}{m}\sum_{x \in S} I(x_i = a_i, x_j = a_j)$.

3. Let T_i include all values if X_i that have empirical frequency at least $0.5\epsilon'/(NL)$, i.e., $T_i = \{a_i \in X_i | \hat{p}_i(a_i) \geq 0.5\epsilon'/(NL)\}$.

4. For each NAND edge (i,j), for each $a_i \in T_i$ and $a_j \in T_j$, mark them as `positive` if they never co-occur, i.e., $\hat{p}_{ij}(a_i, a_j) = 0$. For each "$\subseteq$" edge (i,j), for each $a_i \in T_i$ and $a_j \in T_j$, mark a_i as `positive` and a_j as `negative` if they never co-occur.

5. For each category i, for each value $a_i \in X_i$ not marked in step 4: if category i is positive-constrained (it is the endpoint of some NAND edge or on the subset side of some "$\subseteq$" edge) then mark a_i as `negative`; else mark a_i as `positive`.

The following theorem shows that the error rate of the algorithm is low.

Theorem 15 *Under assumptions (1-4) of Section 3.3, Algorithm* **LearnDiscrete** *produces a labeling of the domain with error at most ϵ with probability $1-\delta$.*

Proof: First, condition (1) implies that each category must have at least one positive value a_i of probability at least $\alpha/N \geq \epsilon'/(NL)$ and at least one negative value a_i of probability at least $\alpha/N \geq \epsilon'/(NL)$. By Chernoff bounds, the sample size m is sufficient so that with probability $1-\delta/2$, all values with probability at least $\epsilon'/(NL)$ have $\hat{p}_i(a_i) > \frac{1}{2}\epsilon'/(NL)$ and so will fall into the sets T_i, and furthermore no values with probability less than $\epsilon'/(4NL)$ will fall into the sets T_i. Assume for the remainder of the argument that this indeed is the case. This in turn implies that in step 4, in each category i incident to a NAND edge or on the subset side of a "$\subseteq$" edge, Algorithm **LearnDiscrete** correctly marks all positive values a_i of probability mass at least $\epsilon'/(NL)$ as `positive`, and in each category i on the superset side of a "$\subseteq$" edge, the algorithm correctly marks all negative values a_i of probability at least $\epsilon'/(NL)$ as `negative`.

Next, ignore all values not in sets T_i since together they have probability at most ϵ. For the remainder, we need to show that with probability at least $1-\delta/2$ we correctly mark the *negative* values a_i in positive-constrained categories and the *positive* values in negative-constrained categories. (We may ignore categories that are both positive- and negative-constrained since they are covered by the previous analysis.) For this, we use Lemma 14. Specificially, pick one such $a_i \in T_i$. Consider any edge (i,j) and any $a_j \in T_j$. By definition, (a_i, a_j) is not disallowed by the ontology, so, by condition (3), the probability of the associated labeling is at least α_2. Therefore, the probability of observing the pair (a_i, a_j) is at least $\alpha_2\lambda^2(\epsilon'/(4NL))^2$ by Lemma 14. The sample size m is therefore sufficient so that with probability at least $1-\delta/(2N^2L^2)$, we indeed have $\hat{p}_{ij}(a_i, a_j) > 0$. Applying the union bound, with high probability all such pairwise events occur. Therefore, such values a_i are marked only in step 5 of the algorithm and so are marked correctly. ■

6. Open Problems

It would be interesting to extend the results of Section 3.4 to remove the assumption that extensions will never be "barely safe", as well as to extend these results (using perhaps a different idealized algorithm) to the case where not all categories are incident to a NAND edge. It would also be of interest to further weaken the limited independence conditions used in Section 3.3.

Acknowledgments

This work was supported in part by NSF grants CCF-0953192, CCF-1116892, and IIS-1065251, AFOSR grant FA9550-09-1-0538, ONR grant N00014-09-1-0751, by a Microsoft Faculty Fellowship, and by grants from the Israel Science Foundation, the US-Israel Binational Science Foundation, the Israeli Ministry of Science and the Israeli Centers of Research Excellence (I-CORE) program (Center No. 4/11). We thank Tong Zhang and Lev Reyzin for early discussions on models of multiple-task learning, and Tom Mitchell for a number of exciting discussions of the NELL system.

References

Abney, S. Bootstrapping. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 360–367, 2002.

Balcan, M.-F. and Blum, A. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT)*, 2005.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *The 11th Annual Conference on Computational Learning Theory (COLT)*, 1998.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E. R., and Mitchell, T.M. Toward an architecture for never-ending language learning. In *AAAI*, 2010a.

Carlson, A., Betteridge, J., Wang, R. C., Hruschka Jr., E.R., and Mitchell, T.M. Coupled semi-supervised learning for information extraction. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 2010b.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934, 2010.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. URL http://www.kyb.tuebingen.mpg.de/ssl-book.

He, Jingrui and Lawrence, Rick. A graphbased framework for multi-task multi-view learning. In *ICML*, pp. 25–32, 2011.

Mohamed, T., Hruschka Jr., E.R., and Mitchell, T. Discovering relations between noun categories. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1447–1455, July 2011.

Sridharan, K. and Kakade, S.M. An information theoretic framework for multi-view learning. In *COLT*, pp. 403–414, 2008.

Verma, S. and Hruschka Jr., E.R. Coupled bayesian sets algorithm for semi-supervised learning and information extraction. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2012)*, Bristol, UK, September 2012. Association for Computing Machinery.

Zhu, Xiaojin and Goldberg, Andrew B. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.