8803 Machine Learning Theory

Maria Florina Balcan Lecture 21: March 30th, 2010

The PAC model assumes access to a noise-free oracle for examples of the target concept. In reality, we need learning algorithms with at least some tolerance for mislabeled examples. In this lecture we study a classic model for learning in the presence of noise, the Random Classification noise model. This is a simple noise model in which one can get positive algorithmic results.

The Random Classification Noise Model

In this model, a learning algorithm will have access to a modified and noisy oracle for examples, denoted by $EX^{\eta}_{CN}(c^*, D)$. Here c^* and D are the target concept and distribution, and $0 \leq \eta < 1/2$ is a new parameter called the classification error rate.

$$EX^{\eta}_{CN}(c^*, D) = \begin{cases} \text{with probability } 1-\eta, \text{ returns } \langle x, c^*(x)\rangle \text{ from } EX(c^*, D) \\ \text{with probability } \eta, \text{ returns } \langle x, \neg c^*(x)\rangle \text{ from } EX(c^*, D) \end{cases}$$

This model was first introduced by Angluin and Laird (1988).

As the noise rate approaches 0.5, the labels provided by the noisy oracle are providing less and less information about the target concept. The learning algorithm thus needs more oracle calls and more computation time as the noise rate approaches 0.5. When the noise rate is equal to 0.5, PAC learning becomes impossible, because every label seen by the algorithm is the outcome of an unbiased coin flip, and gives no information about the target concept.

Definition 1 *C is PAC-learnable by $\mathcal{H}$ in the presence of noise if there exists a learning algorithm $\mathcal{L}$ with the property that $(\forall c \in C)(\forall D$ on $X)(\forall \varepsilon,\ 0 < \varepsilon < 1)$ $(\forall \delta,\ 0 < \delta < 1)$, $(\forall \eta,\ 0 \leq \eta < 0.5)$, if $\mathcal{L}$ is given inputs ε, δ, η_b $(\eta \leq \eta_b < 0.5)$ and access to $EX^{\eta}(c^*, D)$, then it will with probability $\geq 1-\delta$ produce an ouptut hypothesis $h \in \mathcal{H}$ s.t. $error(h) \leq \varepsilon$. C is efficiently PAC-learnable if the running time of $\mathcal{L}$ is polynomial in n, $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, $\frac{1}{1-2\eta_b}$.*

Note:

- Despite the classification noise in the examples received, the goal of the learner remains that of finding a good approximation to the target concept with respect to the distribution D. The error rate is measured with respect to the target concept and distribution.

$$error(h) = \sum_{x: c^*(x) \neq h(x)} \Pr_D(x)$$

- The learner is allowed more time as $\eta_b \to 0.5$.

Empirical Risk Minimization

The typical noise-free PAC algorithm draws a large number of samples and outputs a consistent hypothesis. With classification noise, however, there may not be a consistent hypothesis. Angluin

and Laird show that ERM (empirical risk minimization) can be used to learn in the random classification noise model, though not necessarily efficiently. To remind you, the ERM algorithm is specified as follows:

- Draw a "large enough" sample.
- Output hypothesis $c \in C$ which minimizes disagreements with the sample.

To analyze this, suppose concept c_i has true error rate d_i. What is the probability p_i that c_i disagrees with a labelled example drawn from $EX^{\eta}_{CN}(c^*, D)$? We have two cases.

1. $EX^{\eta}_{CN}(c^*, D)$ reports correctly, but c_i is incorrect: $d_i(1-\eta)$
2. $EX^{\eta}_{CN}(c^*, D)$ reports incorrectly, but c_i is correct: $(1-d_i)\eta$

$$p_i = d_i(1-\eta) + (1-d_i)\eta$$
$$p_i = \eta + d_i(1-2\eta)$$

An ε-good hypothesis has expected disagreement rate $\leq \eta + \varepsilon(1-2\eta)$.

We need m large enough such that an ε-bad hypothesis will not minimize disagreements with the hypothesis. Consider the point $\eta + \frac{\varepsilon(1-2\eta)}{2}$. If an ε-bad hypothesis minimizes disagreements, then the target function must have at least as large of a disagreement rate. Thus, at least one of the following events must hold:

1. Some ε-bad hypothesis c_i has empirical disagreement rate $\leq \eta + \frac{\varepsilon(1-2\eta)}{2}$.
2. Target concept c^* has empirical disagreement rate $\geq \eta + \frac{\varepsilon(1-2\eta)}{2}$.

If C is finite, by Hoeffding bound, if we choose

$$m = O\left(\frac{1}{\varepsilon^2(1-2\eta_b)^2} \ln\left(\frac{|C|}{\delta}\right)\right),$$

then the probability that either of the events occurs is at most δ. That is, the probability that an ϵ-bad hypothesis minimizes disagreements is at most δ. Thus, if we draw a sample of size m as specified above, and then find a hypothesis which minimizes disagreements with the sample, then we have an algorithm which PAC learns in the presence of classification noise. This gives the following theorem.

Theorem 1 *For all finite concept classes C, we can PAC learn in the presence of classification noise.*

We can get similar results for classes of finite VCdimension.

Minimizing disagreements (ERM) can be NP-hard

Theorem 2 *Finding monotone conjunctions which minimize disagreements with a given sample is NP-hard.*

Proof: The proof is a polynomial-time reduction from the decision version of the vertex cover problem to the decision version of our problem. The decision version of the vertex cover problem is specified by an undirected graph $G = (V, E)$ of n vertices and a positive integer $c < n$, and the question is whether there exists a set C of at most c vertices of G such that every edge of G is incident to at least one vertex in C. (Such a set C is called a vertex cover.) This problem is NP-complete.

Our reduction is as follows. If there are n vertices in the graph, then our instance space is $\{0,1\}^n$. For each vertex v_i in the graph, we introduce a positive example $(a_i, +)$, where $a_i = (11\cdots101\cdots11)$, with 0 in i-th position, and for each edge e in the graph we introduce $n+1$ negative examples $(b_e, -)$, where $b_e = (11\cdots101\cdots101\cdots11)$, with 0's in i-th and j-th positions. Thus we get a set S of labeled examples of size $n + |E|(n+1)$. Clearly, the computation of S from (G, c) can clearly be carried out in polynomial time.

It is easy to show the following.

Claim 1 *G has a vertex cover of size at most c if and only if there is a monotone conjunction with at most c disagreements.*

Suppose G has a vertex cover C of at most c vertices. Let f denote the product of those x_i such that v_i is in C. How many examples from S disagree with c? By construction, for each vertex v_i, $f(a_i) = -$ iff $v_i \in C$. Thus, f disagrees with at most c positive examples from S. For each edge $e = (v_i, v_j)$, the set C contains at least one of v_i or v_j, so f contains at least one of x_i or x_j. So, $f(b_e) = -$. Thus, f agrees with all the negative examples in S. Hence the number of disagreements is at most c, as claimed.

Now suppose that there exists some conjunction f such that f disagrees with at most c examples in S. Since $c < n$, this means that f must agree with all the negative examples in S, since each one is repeated $n+1$ times. Hence f can only disagree with positive examples in S, and at most c of them. Thus f must contain at most c literals x_i. Define the set C to be all those vertices v_i such that x_i appears in the conjunction f. Then C contains at most c vertices; it remains to see that it is a vertex cover. If $e = (v_i, v_j)$ is any edge in G then $f(b_e) = -$, since f agrees with all the negative examples. But $f(b_e) = -$, if and only if f contains at least one of x_i or x_j. Thus C contains at least one of v_i or v_j, so C is a vertex cover of G.

■

In the following lecture, we will show how to learn conjunctions *efficiently* in the random classification noise model by using statistical queries.