6.858/18.428 Machine Learning	Lecture 19 : November 21, 1994
Lecturer: Javed Aslam	*Scribe: Gillian Lee*

19.1 Outline

- Standard PAC Model vs. Noisy PAC Model
- Noise Models
- Minimizing disagreements to learn with classification noise
- Minimizing disagreements for conjunctions is NP-hard

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 show that all finite concept classes can be learned with classification noise if we minimize disagreements. However, then we show that for the simple concept class of monotone conjunctions, minimizing disagreements is NP-hard. In the following lecture, we will show how to get *efficient* learning algorithms via the statistical query model.

19.2 Standard PAC Model vs. Noisy PAC Model

The standard PAC model assumes that the learning algorithms have access to a noise-free examples oracle for the target concept (see Figure 19.1). In the noisy PAC model, however, we may have noise which corrupts examples and/or the classification of examples (see Figure 19.2). We consider two types of noise models: the *malicious noise* model and the *classification noise* model.

19.2.1 Malicious Noise

$$EX_{MN}^{\beta}(c_*, \mathcal{D}) = \begin{cases} \text{with probability } 1-\eta, \text{ returns } \langle x, c_*(x)\rangle \text{ from } EX(c_*, \mathcal{D}) \\ \text{with probability } \eta, \text{ returns } \langle x, l\rangle \text{ for maliciously chosen} \\ \qquad x \in X \text{ and } l \in \{0,1\} \end{cases}$$

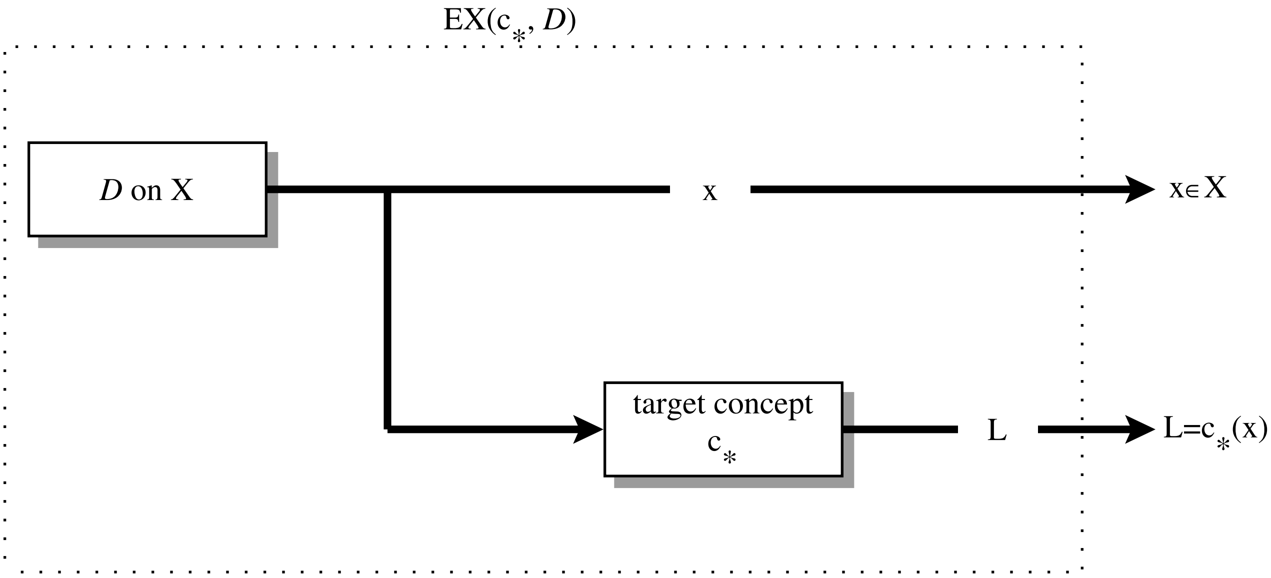

Figure 19.1: Standard PAC Model

The malicious error noise model was first introduced by Valiant (1985). His results for CNF(n,k) hold only for a small rate of noise. Kearns and Li (1987) show that the malicious noise model can tolerate only a small rate of noise for any domain.

19.2.2 Random Classification Noise

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

- label flipped with probability η
- example unchanged

This model was first introduced by Angluin and Laird (1988). We shall discuss their results below.

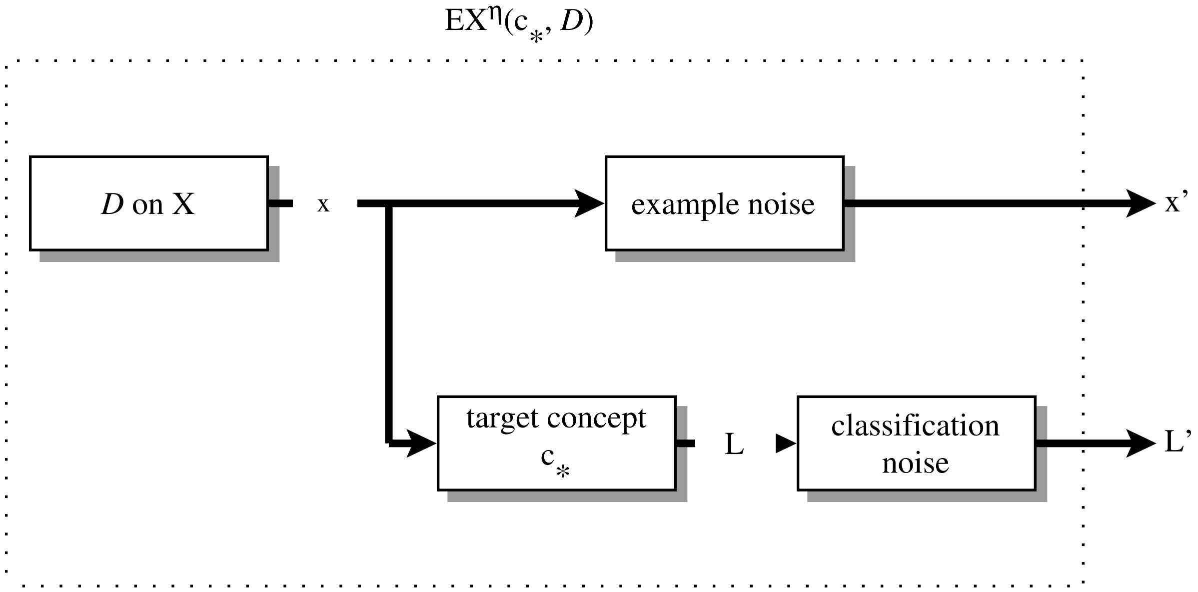

Figure 19.2: Noisy PAC Model

19.3 Learning With Classification Noise

We assume:

- The learner knows which type of noise exists.
- The learner has an upper bound η_b on noise rate $(0 \leq \eta \leq \eta_b < 0.5)$

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

Definition 2 *$\mathcal{C}$ 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:

- 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 $\eta_b \rightarrow 0.5$.

19.3.1 Angluin and Laird Method

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 propose the following method.

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

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 $EX^{\eta}_{CN}(c_*, \mathcal{D})$? We have two cases.

1. $EX^{\eta}_{CN}(c_*, \mathcal{D})$ reports correctly, but c_i is incorrect: $d_i(1-\eta)$
2. $EX^{\eta}_{CN}(c_*, \mathcal{D})$ 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 $\leq \eta + \varepsilon(1-2\eta)$.
An ε-bad hypothesis has expected disagreement rate $> \eta + \varepsilon(1-2\eta)$.

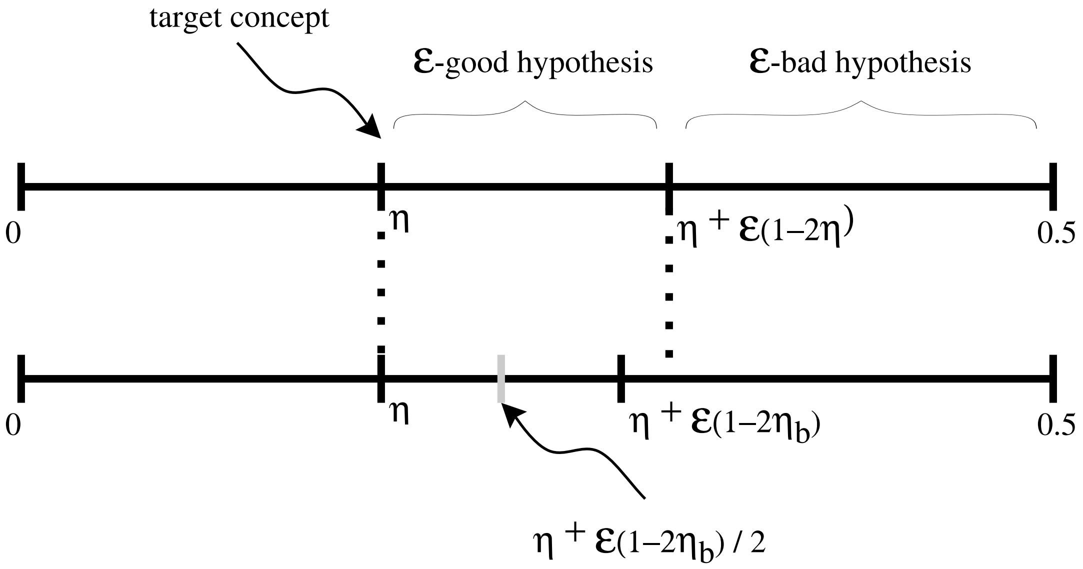

Figure 19.3: Disagreement Rates

We need m large enough such that an ε-bad hypothesis will not minimize disagreements with the hypothesis. Consider the point $\eta + \frac{\varepsilon(1-2\eta_b)}{2}$. 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 $\leq \eta + \frac{\varepsilon(1-2\eta_b)}{2}$.
2. Target concept c_* has empirical disagreement rate $\geq \eta + \frac{\varepsilon(1-2\eta_b)}{2}$.

We can make these probabilities small (i.e., $\leq \frac{\delta}{2}$) using Hoeffding Bounds:

$$GE(p, m, p + \alpha) \leq e^{-2m\alpha^2}$$

$$LE(p, m, p - \alpha) \leq e^{-2m\alpha^2}$$

We have:

$$\begin{aligned}
\Pr(\text{event } 2) &= GE(\eta, m, \eta + \frac{\varepsilon(1-2\eta_b)}{2}) \\
&\leq e^{-2m\left(\frac{\varepsilon(1-2\eta_b)}{2}\right)^2}
\end{aligned}$$

We want $\Pr(\text{event } 2) \leq \frac{\delta}{2}$, so choose

$$m \geq \frac{2}{\varepsilon^2(1-2\eta_b)^2} \ln\left(\frac{2}{\delta}\right).$$

We have:

$$\begin{aligned}
\Pr(\text{event } 1) &\leq LE(\eta + \varepsilon(1-2\eta), m, \eta + \frac{\varepsilon(1-2\eta_b)}{2})|\mathcal{C}| \\
&\leq LE(\eta + \varepsilon(1-2\eta_b), m, \eta + \frac{\varepsilon(1-2\eta_b)}{2})|\mathcal{C}| \\
&\leq |\mathcal{C}| e^{-2m(\frac{\varepsilon(1-2\eta_b)}{2})^2}
\end{aligned}$$

We want Pr(event 1)$\leq \frac{\delta}{2}$, so choose

$$m \geq \frac{2}{\varepsilon^2(1-2\eta_b)^2} \ln\left(\frac{2|\mathcal{C}|}{\delta}\right).$$

If we choose m larger than both values, 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 $\mathcal{C}$, we can PAC learn in the presence of classification noise.* ■

19.3.2 Minimizing disagreements can be NP-hard

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

Sketch of Proof: Reduce the vertex cover problem to that of finding a monotone conjunction which minimizes disagreements with a given sample. Given a graph $G = (V, E)$, and a constant c, the vertex cover problem is to determine whether there is a vertex cover of size at most c (i.e., whether there is a subset of the nodes of size at most c such that every edge in the graph is adjacent to some vertex in this subset). Our reduction is given by the following table.

Graph	Monotone Conjunction
$V = \{v_1, v_2, ...v_n\}$	$\{x_1, x_2, ...x_n\}$ define the instance space.
$\forall v_i \in V$	example $(\langle 11 \cdots 101 \cdots 11 \rangle, +)$, with 0 in i-th position.
$\forall$ edges $(v_i, v_j) \in E$	$c + 1$ examples $(\langle 11 \cdots 101 \cdots 101 \cdots 11 \rangle, -)$, with 0's in i-th and j-th positions.

If there are n vertices in the graph, then our instance space is $\{0, 1\}^n$. For each vertex in the graph, we introduce a positive example, and for each edge in the graph we introduce $c + 1$ negative examples. It is straightforward to show the following.

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

■

References

[1] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343-370, 1988.

[2] M. Kearns and M. Li. *Learning in the presence of malicious errors.* (Technical Report TR-03-87). Cambridge, MA: Harvard University, Center for Research in Computing Technology.

[3] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

[4] L. G. Valiant. Learning disjunctions of conjunctions. *Proceedings of the Ninth International Joint Conference of Artificial Intelligence* (pp. 560-566). Los Angeles, CA: Morgan Kaufmann.