CSE 513T Theory of Artificial Intelligence & Machine Learning Sep 22, 2016

Lecture 8

Lecturer: Brendan Juba *Scribe: Wenjia Zhang, Editor: Tyler Ponder*

Topics

1. Agnostic vs "realizable" learning
2. Empirical Risk Minimization and Uniform Convergence

1 Agnostic Learning

1.1 Informally

What is a bird?

- Animal that can fly?
- Consider penguins, injured birds, etc... Are they birds?
- Consider Feathers. What if I pluck the feathers? Do you still have a bird? (Dinosaurs had feathers, too. And they are not birds.)
- Biologists have a formal notion of what makes a bird. The "natural" ("naive") notion of a "bird" seems to be
  1. Distinct from the formal notion.
  2. Based on "rules of thumb" that may be wrong.

More troubling, what is a game? (How do you recognize when someone is playing a game?)

Wittgenstein said there is no formal defining criteria you can use to recognize when people are playing game. This is a problem for AI: we use some kind of formal rule to classify things (birds, people playing games, etc...).

Today, we'll study a notion of learning these (falliable) rules of thumb. This is called *agnostic learning*. We will operate without the assumption that there is a perfect formal characterization of the target property. Alternatively, we can think of this as a model of learning with arbitrary noise.

1.2 Formally

Definition 1 *The agnostic learning problem for a class of representation* $\mathcal{C} = \bigcup_{n\in\mathbb{N}} \mathcal{C}_n$ *(where* $\mathcal{C}_n$ *is a class on* n *attributes* X_n*) is as follows:*
Given as input n, $1/\epsilon$ *and* $1/\delta$ *and given access to labeled examples drawn from an arbitrary distribution* $\mathcal{D}$ *on* X_n *labeled by an arbitrary* $f : X_n \to \{0,1\}$*, produce as output a representation of some efficiently evaluatable* $h \in \mathcal{H}$*, such that with probability* $1-\delta$*,*

$$\Pr_{x\in\mathcal{D}}[h(x) \neq f(x)] \leq \min_{c\in\mathcal{C}_n} \Pr_{x\in\mathcal{D}}[c(x) \neq h(x)] + \epsilon$$

More generally, we may consider $\alpha(n,\epsilon,\epsilon^*)$ *approximate agnostic learning in which the output* h *need only satisfy* $\Pr_{x\in\mathcal{D}}[h(x) \neq f(x)] \leq \alpha(n,\epsilon,\epsilon^*)$*, where* $\epsilon^* = \min_{c\in\mathcal{C}_n} \Pr_{x\in\mathcal{D}}[c(x) \neq h(x)]$.

Example: State of the art for agnostic learning of conjunctions achieves $\alpha(n, \epsilon, \epsilon^*) \approx O(n^{\frac{1}{3}}\epsilon^*)$.

Note: The usual PAC-learning corresponds to the special case where $\epsilon^* = \Pr_{x\in\mathcal{D}}[c^*(x) \neq h(x)] = 0$. We may refer to this as "realizable" learning, to distinguish it from agnostic (non-realizable) learning.

So far, our learning algorithms always solved $Consis^B_{\mathcal{H}}$. Now, as long as f is "non-realizable" by $\mathcal{H}$, we'll always see a counterexample to any $h \in \mathcal{H}$, so long as we have enough data.

2 Empirical Risk Minimization

As a generalization of solving $Consis^B_{\mathcal{H}}$, let's choose the $h \in \mathcal{H}$ that makes the fewest mistakes.

Definition 2 *The empirical error of a representation h on a list of labeled examples $(x^{(1)}, b^{(1)}), \ldots, (x^{(m)}, b^{(m)})$ is the proportion of $x^{(i)} \neq b^{(i)}$ of all examples, or*

$$\widehat{err}(h) = \frac{1}{m}\#\left\{i : h(x^{(i)} \neq b^{(i)})\right\}$$

Definition 3 *An empirical risk minimizer of a class of representations $\mathcal{H}$ and a list of labeled examples is any $h^* \in \mathcal{H}$ such that*

$$\widehat{err}(h^*) \;=\; \min_{h\in\mathcal{H}} \widehat{err}(h)$$

We argue that an empirical risk minimizer (ERM) is a good predictor for many classes $\mathcal{H}$ by generalizing our Occam's Razor Theorem to allow errors.

3 Uniform Convergence

The idea is that $\widehat{err}(h) \approx \Pr_{x\in D}[h(x) \neq f(x)]$ as long as the data set is large enough.

Definition 4 *For a class of representation $\mathcal{H}$, distribution D and target property f, then a list of labeled examples is ϵ-representative, if, for $err(h) := \Pr_{x\in D}[h(x) \neq f(x)]$ and for all $h \in \mathcal{H}$*

$$|err(h) \;-\; \widehat{err}(h)| \;\leq\; \epsilon$$

Lemma 5 *For $\epsilon/2$-representative list of examples any ERM h^* satisfies*

$$err(h^*) \;\leq\; \min_{h\in\mathcal{H}} err(h) \;+\; \epsilon$$

So it solves agnostic learning for $\mathcal{H}$.

Proof of **Lemma 5**: Since h^* is an ERM,

$$\widehat{err}(h^*) \;\leq\; \min_{h\in\mathcal{H}} \widehat{err}(h)$$

Since the data is $\epsilon/2$-representative,

$$\widehat{err}(h) \;\leq\; err(h) \;+\; \epsilon/2 \;\;\forall\, h \in \mathcal{H}$$

So

$$\begin{array}{rcl} \min_{h\in\mathcal{H}} \widehat{err}(h) & \leq & \min_{h\in\mathcal{H}} err(h) \;+\; \epsilon/2 \\ \widehat{err}(h^*) & \geq & err(h^*) \;-\; \epsilon/2 \end{array}$$

By joining above inequalities,

$$err(h^*) - \epsilon/2 \leq \min_{h \in \mathcal{H}} err(h) + \epsilon/2$$

Hence,

$$err(h^*) \leq \min_{h \in \mathcal{H}} err(h) + \epsilon$$

■

Definition 6 *We say that the class of representations $\mathcal{H}$ uniformly converges with sample complexity $m(n, \epsilon, \delta)$ if for every n, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1/2)$, target property f and distribution D, a sample of $m \geq m(n, \epsilon, \delta)$ examples drawn independently from D and labeled by f is ϵ-representative with probability at least $1 - \delta$.*

Theorem 7 *Any class of representations $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$ such that $\mathcal{H}_n$ is a class on n attributes and $|\mathcal{H}_n|$ is finite for all n uniformly converges with $m(n, \epsilon, \delta) = O(\frac{1}{\epsilon^2}(\log |\mathcal{H}_n| + \log 1/\delta))$. So an ERM $h^* \in \mathcal{H}$ solves the agnostic learning task.*

We'll use the following theorem:

Theorem 8 *(Hoeffding's Inequality) Let $Z_1, \ldots, Z_m$ be a sequence of $\{0, 1\}$ valued random variables such that $E[\frac{1}{m} \sum_{i=1}^{m} Z_i] = \mu$, then*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

Proof of **Theorem 6**: Let any $h \in \mathcal{H}$ be given and put

$$Z_i = \begin{cases} 1 & if\ h(x^{(i)}) \neq f(x^{(i)}) \\ 0 & otherwise \end{cases}$$

Then

$$E[Z_i] = \Pr_{x^{(i)} \in D}[h(x^{(i)}) \neq f(x^{(i)})] = err(h)$$

Hence, also

$$E\left[\frac{1}{m}\sum_{i=1}^{m} Z_i\right] = err(h)$$

While

$$\frac{1}{m}\sum_{i=1}^{m} Z_i = \frac{1}{m}\#\left\{i : h(x^{(i)}) \neq f(x^{(i)})\right\}$$

So Hoeffding's Inequality says

$$\Pr[|\widehat{err}(h) - err(h)| > \epsilon] \leq 2e^{-2m\epsilon^2}$$

We want to take a union bound over $\mathcal{H}_n$. So to get ϵ-representative data to have a probability of failing of at most δ, we wish to bound this quantity by $\delta/|\mathcal{H}_n|$.
Now

$$\begin{aligned} 2e^{-2m\epsilon^2} &\leq \frac{\delta}{|\mathcal{H}_n|} \\ m &\geq \frac{1}{2\epsilon^2}\log_2 \frac{2|\mathcal{H}_n|}{\delta} \end{aligned}$$

Thus, for $m \geq \frac{1}{2\epsilon^2}\left(\ln|\mathcal{H}_n| - \ln\frac{2}{\delta}\right)$, (since h is arbitrary) a union bound gives uniform convergence with probability at least $1-\delta$.
■

Difficulty: How do I find an ERM? If $|\mathcal{H}_n| \leq poly(n)$, for example conjunctions of $size \leq k(= O(1))$, then we have:

Corollary 9 *The agnostic learning task for conjunctions of size* $\leq k$ $(= O(1))$ *literals can be solved in poly-time.*

For most natural classes finding an ERM is NP-hard. Note that this doesn't rule out improper learning, etc. The best algorithm we have for these tasks are poly approximations.