1 PAC-learning

A concept class $\mathcal{C}$ is strongly learnable (with $\mathcal{H}$) provided there exists an algorithm A such that for all $\{c \in \mathcal{C}, \mathbb{P} \text{ on } \mathcal{X}, \epsilon > 0, \delta \leq 1\}$, there exists an $h \in \mathcal{H}$ where

$$\mathbb{P}^n(\text{err} \leq \epsilon) \geq \delta$$

An example of strongly learnable concept class: The instances be $\mathcal{X} = \mathbb{R}$, The concept class $\mathcal{C}$ be the set of positive half lines The consistent learner would find any $h \in \mathcal{H} = \mathcal{C}$ in the transition region from negative to positive

Fix an h and c. Let $h_*, c_* \in \mathbb{R}$ be the classification boundaries. Then, we only make a mistake if $x \in [h_*, c_*]$, when $h_* < c_*$ or $x \in [c_*, h_*]$, when $c_* < h_*$. Hence,

$$\text{err} = \mathbb{P}(h(x) \neq c(x)) = \mathbb{P}([h_*, c_*]) \text{ or } \mathbb{P}([c_*, h_*])$$

h is formed based on n data points $x_1, \ldots, x_n \overset{i.i.d}{\sim} \mathbb{P}$, and $\mathcal{D} = \{x_i\}_{i=1}^n$.

Fix an $\epsilon > 0$. Let g_+ be the minimal quantity such that $G_+ = [c_*, g_+]$ obeys $\mathbb{P}(G_+) \geq \epsilon$. Define G_- and g_- similarly for the other side of the interval.

Let's define two (bad) events $B_+ = (g_+, \infty)$, $B_- = (-\infty, g_-)$. By the previous argument, we can bound the probability of B_+

$$\begin{aligned}
\mathbb{P}^n(h_* \in B_+) &\leq \mathbb{P}^n(x_1 \notin G_+, \ldots, x_n \notin G_+) \\
&= \mathbb{P}(x_1 \notin G_+) \cdots \mathbb{P}(x_n \notin G_+) \\
&\leq (1 - \epsilon)^n \\
&\leq e^{-\epsilon n}
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{P}^n(\text{err} > \epsilon) &= \mathbb{P}^n(\mathbb{P}([h_*, c_*] > \epsilon) \text{ or } \mathbb{P}([c_*, h_*]) > \epsilon) \\
&= \mathbb{P}^n(h_* \in B_- \cup B_+) \\
&= \mathbb{P}^n(h_* \in B_-) + \mathbb{P}^n(h_* \in B_+) \\
&\leq 2e^{-\epsilon n}
\end{aligned}$$

Let $\delta > 0$ be given.

$$\mathbb{P}^n(\text{err} > \epsilon) \leq 2e^{-\epsilon n} \overset{set}{=} \delta$$

Result: $n > \epsilon^{-1} \log(2/\delta)$, this concept class is PAC-learnable.

Suppose $|\mathcal{H}| < \infty$ If an algorithm A finds a hypothesis $h \in \mathcal{H}$ that is consistent with n examples, where

$$n > \frac{1}{\epsilon}(log(|\mathcal{H}|) + \log(1/\delta)).$$

Then

$$\mathbb{P}^n(\text{err} > \epsilon) \leq \delta$$

A general result (Proof):

$$\begin{aligned}
\mathbb{P}^n(\text{err} > \epsilon) &= \mathbb{P}^n(\text{err} > \epsilon \cap h \text{ is consistent}) \\
&\leq \mathbb{P}^n(\exists h \text{ that is } \epsilon - \text{bad}, h \text{ is consistent}) \\
&\leq \sum_{h:h \text{ is } \epsilon-\text{bad}} \mathbb{P}^n(h \text{ is consistent}) \\
&= \sum_{h:h \text{ is } \epsilon-\text{bad}} \mathbb{P}^n(h(x_1) = c(x_1), \ldots, h(x_n) = c(x_n)) \\
&= \sum_{h:h \text{ is } \epsilon-\text{bad}} \prod_{i=1}^{n} \mathbb{P}(h(x_i) = c(x_i)) \\
&\leq \sum_{h:h \text{ is } \epsilon-\text{bad}} (1-\epsilon)^n \\
&= |\{h : h \text{ is } \epsilon - \text{bad}\}|(1-\epsilon)^n \\
&\leq |\mathcal{H}|(1-\epsilon)^n \\
&\leq |\mathcal{H}|e^{-\epsilon n} \\
&\stackrel{set}{=} \delta
\end{aligned}$$

Invert to get result.

What we need to refine this result for infinite $\mathcal{H}$ is to know more about its intrinsic complexity. This gives us a better idea of the complexity of the hypothesis space $\mathcal{H}$.

2 VC-dimension

Let $\mathcal{A}$ be a class of sets, $F = \{x_1, \ldots, x_n\}$ be a finite set, and $G \subseteq F$.

We say that $\mathcal{A}$ picks out G (relative to F) if $\exists A \in \mathcal{A}$ s.t.

$$A \cap F = G$$

Let $S(\mathcal{A}, F)$ be the number of subsets of F that can be picked out by $\mathcal{A}$. Of course, $S(\mathcal{A}, F) \leq 2^n$. We say that F is shattered by $\mathcal{A}$ if $S(\mathcal{A}, F) = 2^n$. Also, let $\mathcal{F}_n$ be all finite sets with n elements. Then we have the shattering coefficient of $\mathcal{A}$:

$$s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} S(\mathcal{A}, F)$$

Famous theorem: Let $\mathcal{A}$ be a class of sets. Then

$$\mathbb{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}(A) - \mathbb{P}(A)| > \epsilon) \leq 8 s_n(\mathcal{A}) e^{-n\epsilon^2/32}$$

This partly solves the problem. But, how big can $s_n(\mathcal{A})$ be?

Often times, $s_n(\mathcal{A}) = 2^n$ for all n up to some d, then $s_n(\mathcal{A}) < 2^n$ for all $n > d$. This d is the Vapnik-Chervonenkis (VC) dimension.

This should be compared with SVMs, where we wish to separate points with hyperplanes.

Imagine that $\mathcal{A}$ is the set of hyperplanes in $\mathbb{R}^2$. Let $\mathcal{F}_n$ be all sets of n points. We can shatter almost all $F \in \mathcal{F}_3$ (one is enough, though). But, we cannot shatter any $F \in \mathcal{F}_4 \Rightarrow d = 3$. It is tempting to think that the number of parameters determines the VC dimension. However, if we let $\mathcal{A} = \{\sin(tx) : t \in \mathbb{R}\}$, this is a one parameter family. $\mathcal{A}$ can shatter a F for any $\mathcal{F}_n \Rightarrow d = \infty$

Suppose that $\mathcal{A}$ has VC-dimension $d < \infty$. Then, for any $n \geq d$,

$$s_n(\mathcal{A}) \leq (n+1)^d$$

For small n, the shattering coefficient increases exponentially. For large n, the shattering coefficient increases polynomially.

If n is large enough and $d < \infty$ then

$$\begin{aligned}\mathbb{P}(\sup_{A \in \mathcal{A}} |\hat{\mathbb{P}}(A) - \mathbb{P}(A)| > \epsilon) &\leq 8 s_n(\mathcal{A}) e^{-n\epsilon^2/32} \\ &\leq 8(1+n)^d e^{-n\epsilon^2/32}\end{aligned}$$

3 Boosting

Suppose a learning algorithm cannot attain an error rate below a fixed amount (say 40%). Can we still drive the error rate arbitrary close to zero?Boosting considers this problem, augmenting classifiers that are only marginally better than random guessing.

A concept class $\mathcal{C}$ is weakly learnable (with $\mathcal{H}$) provided there exists an algorithm A and $\gamma > 0$ such that for all $\{c \in \mathcal{C}, \mathbb{P} \text{ on } \mathcal{X}, \delta \leq 1\}$, there exists an $h \in \mathcal{H}$ produced by A on n examples where:

$$\mathbb{P}^n(\text{err} \leq 1/2 - \gamma) \geq \delta$$

Again, only polynomial growth of n is allowed. This means that there is an algorithm that, with high probability can do slightly better than random guessing.

Example: Let $\mathcal{X} = \{0,1\}^n \cup \{Z\}$, $\mathcal{C}$ be all functions on $\mathcal{X}$, and $\mathbb{P}(\{Z\}) = 1/4$, uniform on all other elements. Given a set of examples, the algorithm will quickly learn $c(Z)$, as Z is very likely.

However, as we are only viewing polynomial number of examples from $|\mathcal{X}| \geq 2^n$, the algorithm will do not much better than random guessing. Then

$$\text{expected error} \approx \frac{1}{4}(0) + \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8}$$

So, there exists situations that are only weakly learnable.

Schapire (1990) paper proves a concept class $\mathcal{C}$ is weakly learnable if and only if it is strongly learnable. The proof is constructive and produces the first boosting algorithm.

Given:

1. n examples from some fixed, unknown $\mathbb{P}$
2. a weak learning algorithm A producing $h \in \mathcal{H}$

Produce: a new hypothesis $H \in \mathcal{H}_{new}$ with error $\leq \epsilon$.

Schapire (1990) gives the first boosting algorithm for a weak learning algorithm A.

1. A hypothesis h_1 is formed on n instances
2. A hypothesis h_2 is formed on n instances, half of which are misclassified by h_1

   More specifically, a fair coin is flipped. If the result is heads then A draws samples $x \sim \mathbb{P}$ until $h_1(x) = c(x)$. If the result is tails then we wait until $h_1(x) \neq c(x)$
3. A hypothesis h_3 is formed on n instances, for which h_1 and h_2 disagree
4. the boosted hypothesis h_b is the majority vote of h_1, h_2, h_3

Schapire's "strength of weak learnability" theorem shows that h_b has improved performance over h_1.