COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire | Lecture #4
Scribe: Akshay Mittal | February 13, 2013

1 Proof of learning bounds

For intuition of the following theorem, suppose there exists a hypothesis h which is ϵ-bad and makes at least one mistake with a few 100 examples. Therefore, since h is ϵ-bad, then with high probability, it is going to be eliminated and will not be picked up by the algorithm. By the union bound, we will then show that all of the ϵ-bad hypotheses are inconsistent with that training set.

Theorem. *Say algorithm A always finds hypothesis $h_A \in \mathcal{H}$ consistent with m examples where*

$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

then

$$Pr[err_D(h_A) > \epsilon] \leq \delta$$

The underlying assumption is that the hypothesis space is finite, *i.e.* $|\mathcal{H}| < \infty$ and that the m examples are i.i.d. with respect to the distribution $\mathcal{D}$. The theorem provides a upper bound on the amount of training data m needed to achieve a low error ϵ with a confidence of at least $1 - \delta$.

Proof. We aim to bound the probability that h_A is both consistent and ϵ-bad, *i.e.* the generalization error of h_A is greater than ϵ. Let $\mathcal{B} = \{h \in \mathcal{H} : h\ \epsilon\text{-bad}\}$ be the set of all ϵ-bad hypotheses in $\mathcal{H}$. (Here $\mathcal{B}$ is a fixed set and not a random variable. The concept c and distribution $\mathcal{D}$ are fixed, thus, the hypotheses h having error on $\mathcal{D}$ are fixed. The only random variable is h_A which depends on the sample. The consistency of h_A is also random.)

$$\begin{aligned}
&Pr[h_A \text{ is } \textit{consistent} \text{ and } \epsilon\text{-bad}] \\
&\qquad \leq Pr[\exists h \in \mathcal{H} : h\ \textit{cons.}\ \&\ \epsilon\text{-bad}] && (\because \text{if } A \Rightarrow B, \text{ then } Pr[A] \leq Pr[B]) \\
&\qquad = Pr[\exists h \in \mathcal{B} : h\ \textit{cons.}] \\
&\qquad = Pr\left[\bigvee_{h\in\mathcal{B}} (h\ \textit{cons.})\right] \\
&\qquad \leq \sum_{h\in\mathcal{B}} Pr[h\ \textit{cons.}] && \text{(by union bound)} \\
&\qquad = \sum_{h\in\mathcal{B}} Pr[h(x_1) = c(x_1) \wedge \ldots \wedge h(x_m) = c(x_m)] && \text{(by defn. of consistency)} \\
&\qquad = \sum_{h\in\mathcal{B}} \prod_{i=1}^{m} Pr[h(x_i) = c(x_i)] && \text{(by independence)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{h\in\mathcal{B}} (1-\epsilon)^m && (\because h \in \mathcal{B} \Rightarrow Pr[h(x_i) \neq c(x_i)] \geq \epsilon)\\
&= |\mathcal{B}|(1-\epsilon)^m &&\\
&\leq |\mathcal{H}|(1-\epsilon)^m && (\because \mathcal{B} \subseteq \mathcal{H})\\
&\leq |\mathcal{H}|e^{-\epsilon m} && (\because \forall x, (1+x) \leq e^x)\\
&\leq \delta && \text{(follows by choice of } m)
\end{aligned}
$$

□

The negation of $Pr[\exists h \in \mathcal{H} : h\ cons.\ \&\ \epsilon\text{-bad}]$ leads us to conclude that with probability $\geq (1-\delta)$ and $\forall h \in \mathcal{H}$, if h is consistent, then

$$
err_{\mathcal{D}}(h) \leq \epsilon = \frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m} \tag{1}
$$

Equation 1 captures the bound on the generalization error in terms of the learning performance δ, the size of the hypothesis space $|\mathcal{H}|$ and the number of training sample m.

An Alternate Proof

We can attempt to get rid of the dependence of the generalization error on the number of hypotheses $|\mathcal{H}|$ as follows

$$
\begin{aligned}
&Pr[err_{\mathcal{D}}(h_A) > \epsilon \mid h_A\ cons.]\\
&\quad = \frac{Pr[h_A\ cons. \mid err(h_A) > \epsilon]\ Pr[err(h_A) > \epsilon]}{Pr[h_A\ cons.]} && \text{(by Bayes rule)}\\
&\quad = Pr[h_A\ cons. \mid err(h_A) > \epsilon]\ Pr[err(h_A) > \epsilon] && (\because Pr[h_A\ cons.] = 1)\\
&\quad \leq Pr[h_A\ cons. \mid err(h_A) > \epsilon] && (\because Pr[err(h_A) > \epsilon] \leq 1)\\
&\quad = Pr[h_A(x_1) = c(x_1) \wedge \ldots \wedge h_A(x_m) = c(x_m) \mid err(h_A) > \epsilon] && \text{(by defn. of consistency)}\\
&\quad = \prod_{i=1}^{m} Pr[h_A(x_i) = c(x_i) \mid err(h_A) > \epsilon] && \text{(by conditional independence)}\\
&\quad \leq (1-\epsilon)^m && (\because Pr[h(x_i) \neq c(x_i)] \geq \epsilon)\\
&\quad \leq e^{-\epsilon m} && (\because \forall x, (1+x) \leq e^x)\\
&\quad \leq \delta && (\text{if } m \geq \tfrac{ln\frac{1}{\delta}}{\epsilon})
\end{aligned}
$$

The argument above seems plausible, but it is actually incorrect. In the first proof, h is not a random variable, since we had picked it before the sample $\mathcal{S}$ was picked, hence use of *independence* is valid in that case. However in this alternate proof, the hypothesis h_A is generated from the sample $\mathcal{S}$, and therefore is a random variable that depends on the sample $\mathcal{S}$. Since h_A depends on the sample $\mathcal{S}$, given h_A is ϵ-bad, the samples from $\mathcal{S}$ are no longer i.i.d. Thus, use of conditional independence in the above proof is incorrect, *i.e.*

$$
Pr[h_A\ cons. \mid err(h_A) > \epsilon] \neq \prod_{i=1}^{m} Pr[h_A(x_i) = c(x_i) \mid err(h_A) > \epsilon]
$$

Moreover, $Pr[h_A \; cons. \mid err(h_A) > \epsilon]$ should be 1, because we assume that h_A is always consistent. Therefore, care must be taken to pick the hypothesis (for which the generalization error is being analysed) before the sample space $\mathcal{S}$ is selected.

2 Consistency via PAC

In the previous section, we have seen that if we can learn in the consistency model, then we can learn in the PAC-model, provided $|\mathcal{H}|$ is not too huge. A concept class $\mathcal{C}$ is said to be *properly* PAC learnable by $\mathcal{H}$ if the hypotheses space $\mathcal{H}$ is the same as the concept class $\mathcal{C}$. We will now take the situation considering the case vice-versa.

Proposition. *Given*

- *algorithm A that properly PAC-learns $\mathcal{C}$,* i.e. *given a set of random examples, A finds a hypothesis $h \in \mathcal{C}$, such that with high probability, the hypothesis has generalization error at most ϵ.*

- *a sample $\mathcal{S} = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$*

we can use A as a subroutine to find $c \in \mathcal{C}$ consistent with $\mathcal{S}$ (if one exists).

Intuitively, since a PAC learning algorithm must have examples from a random distribution and $\mathcal{S}$ is *not* a random set, we construct a distribution for it and sample the examples (for feeding to algorithm A) from it. We then use algorithm A to get a hypothesis h such that $err_{\mathcal{D}}(h) \leq \epsilon$ and use it to show that h is consistent.

Proof. Given m examples $\mathcal{S}$, we construct a distribution $\mathcal{D}$ that is *uniform* over the m examples in $\mathcal{S}$. We choose $\epsilon = \frac{1}{2m}$ and any desired value of $\delta > 0$. We then run algorithm A on $m' = poly(\frac{1}{\epsilon}, \frac{1}{\delta})$ examples chosen from the distribution $\mathcal{D}$. Here m' is the number of examples required by A to attain the desired accuracy $(1 - \epsilon)$ with high probability $1 - \delta$. If A outputs the algorithm h, we check whether h is consistent with $\mathcal{S}$. If h is consistent with $\mathcal{S}$, then we output the hypothesis h (thus proving the proposition), else (or if A failed to generate a hypothesis), then we say "nothing consistent". Mathematically, if there exists $c \in \mathcal{C}$ consistent with $\mathcal{S}$, then with probability at least $(1 - \delta)$ (since A is PAC-learning algorithm), we have

$$\begin{aligned} err_{\mathcal{D}}(h) &\leq \epsilon \\ &= \frac{1}{2m} \\ &< \frac{1}{m} \end{aligned} \tag{2}$$

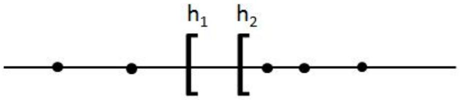
Since $\mathcal{D}$ is uniform, the probability assigned to each example is $\frac{1}{m}$ and therefore the generalization error is an integer multiple of $\frac{1}{m}$. By Equation 2, this leads to the conclusion that $err_{\mathcal{D}}(h) = 0$ and h is consistent. If, however, there does not exist a $c \in \mathcal{C}$ which is consistent, then algorithm A would fail somehow *i.e.* either give a hypothesis h which is inconsistent or terminate by saying that "nothing consistent". □

Consistency and PAC-learnability are closely related concepts.

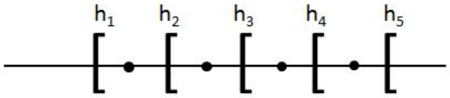
3 Learnability of Infinite Hypothesis Space

The result shown above holds only for the finite hypothesis spaces. There are still various examples, such as positive half-lines, rectangles, etc. that allow us to learn even though they have infinite hypothesis spaces. We will now discuss the characteristics of these hypothesis spaces to determine what makes a concept class $\mathcal{C}$ PAC-learnable.

Example 3.1. Positive Half-lines



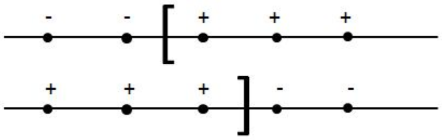
Given any unlabeled dataset of points on the x-axis, h_1 and h_2 behave exactly the same on the dataset. Although there are infinitely many hypotheses possible for this example, using the similarity of multiple hypotheses, we can divide the hypothesis into finite distinct equivalence classes. For instance, below are the 5 possible labelings/dichotomies/behaviors for a set of 4 unlabeled examples.



h1	+	+	+	+
h2	-	+	+	+
h3	-	-	+	+
h4	-	-	-	+
h5	-	-	-	-

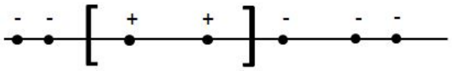
Therefore, in general, for m unlabeled examples we have $(m+1)$ possible equivalence classes compared to the 2^m possible labelings (there are infinitely many hypotheses but only finitely many labelings) for the unlabeled dataset $[m+1 \ll 2^m]$. The fact that the number of equivalence classes/labellings/dichotomies is so small, makes this concept class of positive half-lines PAC-learnable. Even though the hypothesis space is infinitely large, the *effective* hypothesis space is small $\mathcal{O}(m)$.

Example 3.2. Positive/Negative Half-lines



This case is similar to Example 3.1 except that the half-line can label the examples positive or negative on either side of the marker point. To compute the effective hypothesis space, we double the result of Example 3.1 and subtract 2 (to account for the double counting of all positive labels and all negative labels). Thus, the effective hypothesis space $= 2(m+1)-2 = 2m \equiv \mathcal{O}(m)$.

Example 3.3. Intervals
The concept class $\mathcal{C}$ consists of concepts which classify points, on the real axis, inside an interval (specific to every concept) as positive and those outside the interval as negative.



With m points, there are $(m+1)$ ways to place a marker for an interval boundary, thus the number of ways to select an interval, of this format, is $\binom{m+1}{2}$. To account for the empty interval (which can be placed between any two points), one extra value needs to be added, thus totaling the effective hypothesis cardinality to be $\binom{m+1}{2} + 1 \equiv \mathcal{O}(m^2)$. Alternatively, one could first pick pair of points $\binom{m}{2}$, then singleton points m and lastly the empty case 1, once again totaling to $\binom{m}{2} + m + 1$ which computes to be the same result.

Generalizing the characteristics of the aforementioned examples, we have a hypothesis space $\mathcal{H}$ and a set of unlabeled examples $\mathcal{S} = \langle x_1, x_2, \ldots, x_m \rangle$. We can thus define the set of all possible behaviors/dichotomies/labelings of $\mathcal{H}$ on $\mathcal{S}$ as

$$\Pi_{\mathcal{H}}(\mathcal{S}) = \{\langle h(x_1), \ldots, h(x_m) \rangle : h \in \mathcal{H}\}$$

Thus, for Example 3.1 we get $|\Pi_{\mathcal{H}}(\mathcal{S})| = 5$. We can now define a *growth function* over m samples to capture how complex the hypothesis space grows as we see more samples

$$\Pi_{\mathcal{H}}(m) = \max_{|\mathcal{S}|=m} |\Pi_{\mathcal{H}}(\mathcal{S})|$$

Intuitively, we would like to use $\Pi_{\mathcal{H}}(\mathcal{S})$ as an effective hypothesis space and thus replace $\ln |\mathcal{H}|$ in the error bound with $\ln |\Pi_{\mathcal{H}}(m)|$. The error bound depends on the growth function, which means that if the growth function grows as 2^m, *i.e.* $\forall m$, $\Pi_{\mathcal{H}}(m) = 2^m$, then we reduce the bound as $m \geq \frac{m + \ln \frac{1}{\delta}}{\epsilon}$, which is not useful. In this case, learning is impossible because we are working with something like all possible functions. However, we will see that the only other possible case for all hypothesis spaces is that the growth function grows as a polynomial in m, *i.e.* $\Pi_{\mathcal{H}}(m) = \mathcal{O}(m^d)$. Here d is called the *VC-dimension* and the error bound does not blow up

$$\lim_{m \to \infty} \frac{d \ln m + \ln \frac{1}{\delta}}{m} \longrightarrow 0$$

Thus we observe that for any $\mathcal{H}$, either $\Pi_{\mathcal{H}}(m) = 2^m$ for all m, or $\Pi_{\mathcal{H}}(m) = \mathcal{O}(m^d)$ for some constant d.