**COS 511: Theoretical Machine Learning**

Lecturer: Rob Schapire
Scribe: Allen(Zhelun) Wu

Lecture #5
February 19, 2018

---

# Review

**Theorem** (Occam's Razor)**.** *Say algorithm $A$ finds a hypothesis $h_A \in \mathcal{H}$ consistent with $m$ examples where $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$. Then:*

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

*In other words, the probability that the generalization error of $h_A$ is $\epsilon$-bad is bounded by $\delta$.*

However, what if the hypothesis set $\mathcal{H}$ is infinite? In this case, we wish we could obtain a more general result by replacing $|\mathcal{H}|$ by the number of labellings on a finite sample with growth function $\Pi_{\mathcal{H}}(m)$. And what is $\Pi_{\mathcal{H}}(m)$ again?

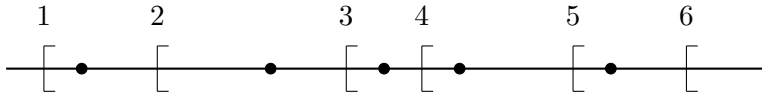
In the general case of unlabeled points $\mathcal{S} = \langle x_1, \ldots, x_m\rangle$, the set of labelings is:

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

**Definition** (Growth function)**.** *The maximum size of the set of labellings on a sample of size $m$ is the growth function:*

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

**Example** (Positive Half-Lines)**.** *Consider positive half-lines over a set of $m = 5$ data points. Although there are infinitely many hypotheses, there are only $m + 1 = 6$ different labelings (also known as behaviors or dichotomies).*




| | | | | | |
|---|---|---|---|---|---|
| 1 | + | + | + | + | + |
| 2 | − | + | + | + | + |
| 3 | − | − | + | + | + |
| 4 | − | − | − | + | + |
| 5 | − | − | − | − | + |
| 6 | − | − | − | − | − |

For half-lines with positive and negative cases,

$$\Pi_{\mathcal{H}}(m) = 2m.$$

For intervals,

$$\Pi_{\mathcal{H}}(m) = \binom{m}{2} + m + 1.$$

Actually we will later prove that there are only two possible cases:

- "nice" case: $\Pi_{\mathcal{H}}(m) = O(m^d)$, $d$ is a constant which turns out to be the VC-dimension.
- worst case: $\Pi_{\mathcal{H}}(m) = 2^m$.

These two cases exactly characterize when PAC learning is or is not possible — in the "nice" case, it is possible, but in the other case, it is not.

In this way, we can reduce the problem of probability and statistics to combinatorial properties of the hypothesis set such as the cardinality of $\mathcal{H}$, its growth function and later, VC-dimension.

# 1 Error bound for consistent hypotheses when $|\mathcal{H}| = \infty$

Occam's Razor provides a general technique to show that a learning algorithm generalizes with low error to new data with high probability — is probably approximately correct (PAC). And now we are generalizing the theorem to the case when $|\mathcal{H}| = \infty$.

**Theorem** (Generalized Occam's Razor)**.** *With probability at least $1 - \delta$, if $h_A \in \mathcal{H}$ is consistent, then its true generalization error is bounded:*

$$err_D(h_A) \leq \epsilon$$

*where*

$$\epsilon = O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right).$$

*(Constants will be filled in below.)*

# 2 Notation and Proof

We let $B$ denote the bad event that there exists a hypothesis $\mathcal{H}$ that is consistent with the sample $\mathcal{S}$ but for which $err_D(h) > \epsilon$. Instead of working with infinite samples, we prefer working with finite samples, so we'll imagine drawing a second "ghost" sample of size $m$. Thie is called the double-sample trick.

$$B = \exists h \in \mathcal{H} : h \textit{ consistent with } S \wedge err_D(h) > \epsilon$$

$$\mathcal{S} : x_1, x_2, \ldots, x_m, \quad \textit{real samples}$$

$$\mathcal{S}' : x'_1, x'_2, \ldots, x'_m (i.i.d \sim D), \ \textit{imagined/ghost samples}$$

We have these imagined samples because we want to use them to bound the true error rate by proving if $\mathcal{S}$ is good then $\mathcal{S}'$ being very bad is unlikely. We then use $h$ to test on $\mathcal{S}$ and count the number of errors $h$ makes on $\mathcal{S}$, which is $M(h, \mathcal{S})$:

$$M(h, \mathcal{S}) = \textit{ number of mistakes } h \textit{ makes on } \mathcal{S}$$

We also define $B'$ to be the event that there exists a hypothesis in $\mathcal{H}$ that makes no mistakes on $\mathcal{S}$, but at least $m\epsilon/2$ mistakes on $\mathcal{S}'$:

$$B' : \exists h \in \mathcal{H} : M(h, \mathcal{S}) = 0 \wedge M(h, \mathcal{S}') \geq \frac{m\epsilon}{2}$$

There are seven steps to finish the proof. Later, we are going to introduce more notation.

## Step 1: $Pr[B'|B] \geq 1/2$

If $B$ occurs, there must be a hypothesis that is consistent with $\mathcal{S}$ but has error at least $\epsilon$. Given such an $h$, the probability that $B'$ occurs is at least the chance that $h$ has more than $m\epsilon/2$ mistakes on $\mathcal{S}'$, in other words, at least half the expected number of mistakes, which is $m\epsilon$. This probability can be shown to be at least $1/2$, but to prove this formally will require Chernoff bounds, which will be discussed later in the course.

## Step 2: $Pr[B] \leq 2Pr[B']$

$$Pr[B'] \geq Pr[B \wedge B'] = Pr[B] \cdot Pr[B'|B] \geq \frac{1}{2} Pr[B].$$

## Step 3: $Pr[B'] = Pr[B'']$

$B'$ is the event that some $h$ has a lot of mistakes, and none of them occur on the real sample, but they instead all occur on the ghost sample. Because the data is i.i.d., this is very unlikely. To draw this out, we randomly permute the datasets $\mathcal{S}$ and $\mathcal{S}'$. We flip coins for $m$ times. For the $i$th time if we get a head, we do nothing, otherwise, we switch $x_i$ and $x_i'$. And after $m$ flips we get new datasets $\mathcal{T}$ and $\mathcal{T}'$.

First of all, let us define $B''$ here as the same as $B'$, but with $\mathcal{S}, \mathcal{S}'$ replaced by $\mathcal{T}, \mathcal{T}'$.

$$B'' : \exists h \in \mathcal{H} : M(h, T) = 0 \wedge M(h, T') \geq \frac{m\epsilon}{2}.$$

In this case, given the distribution of $\mathcal{T}$,$\mathcal{T}'$ and $\mathcal{S}$,$\mathcal{S}'$ are equal, all i.i.d according to $D$, $Pr[B'] = Pr[B'']$ is obviously true.

## Step 4: $Pr[b(h)|\mathcal{S}, \mathcal{S}'] \leq 2^{-m\epsilon/2}$

Let us introduce the event $b(h)$ that for fixed $h$, $h$ makes no mistakes on $\mathcal{T}$ but many mistakes on $\mathcal{T}'$.

$$b(h) : M(h, \mathcal{T}) = 0 \wedge M(h, \mathcal{T}') \geq \frac{m\epsilon}{2}$$

$Pr[b(h)|\mathcal{S}, \mathcal{S}']$ represents the probability of $b(h)$ happening when $\mathcal{S}, \mathcal{S}'$ are fixed.

In the following, we will use 1 for sample $x$ if $h(x) \neq c(x)$ and 0 for $h(x) = c(x)$.

**Case 1: $\exists x_i, x_i'$ with both of them labeled as 1**

In this case, no matter how the examples in $\mathcal{S}$ and $\mathcal{S}'$ are switched, $M(h, \mathcal{T}) = 0$ is not satisfied since $h$ makes a mistake on both $x_i$ and $x_i'$. Therefore $Pr[b(h)|\mathcal{S}, \mathcal{S}'] = 0$.

$$\mathcal{S} : 0\ 1\ 0\ 0\ 1\ 0$$

$$\mathcal{S}' : 1\ 0\ 0\ 0\ 1\ 1$$

$$\uparrow$$

One possible $\mathcal{T}, \mathcal{T}'$ could be

$$\mathcal{T} : 1\ 1\ 0\ 0\ 1\ 0$$

$$\mathcal{T}' : 0\ 0\ 0\ 0\ 1\ 1$$

As pointed out by the arrow, no matter how we switch it, still two 1s there, contradicting the fact that $\mathcal{T}$ is consistent. So there are only two cases, either both $x_i$ and $x_i'$ are 0 or one of which is 1. And let us define $r$ as the number of pairs with exactly one error on $x_i$ or $x_i'$.

**Case 2: $r < \frac{m\epsilon}{2}$**

In this case $Pr[b(h)|\mathcal{S}, \mathcal{S}'] = 0$ because with $r < \frac{m\epsilon}{2}$, there are only less than $\frac{m\epsilon}{2}$ 1s in total, not enough to satisfy $M(h, \mathcal{T}') \geq \frac{m\epsilon}{2}$.

**Case 3: $r \geq \frac{m\epsilon}{2}$**

In this case, we need to flip all errors in $\mathcal{T}$ so that all of the errors are in $\mathcal{T}'$. Since the permutations are all independent, the probability is $(1/2)^r = 2^{-r}$.

$$Pr[b(h)|\mathcal{S}, \mathcal{S}'] = 2^{-r} \leq 2^{-m\epsilon/2}.$$

## Step 5: $Pr[B''|\mathcal{S}, \mathcal{S}'] \leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}$

For every labeling on $\mathcal{S} \cup \mathcal{S}'$, we select one "representative" hypothesis from $\mathcal{H}$ realizing that particular labeling. And let $\mathcal{H}'$ denote this space of "representative" hypotheses.

$$\begin{aligned}
Pr[B''|\mathcal{S}, \mathcal{S}'] &= Pr[\exists h \in \mathcal{H} : b(h)|\mathcal{S}, \mathcal{S}'] \ \ (\textit{definition}) \\
&= Pr[\exists h \in \mathcal{H}' : b(h)|\mathcal{S}, \mathcal{S}'] \ \ (\mathcal{H}' \textit{ has all the representative hypotheses}) \\
&\leq \sum_{h \in \mathcal{H}'} Pr[b(h)|\mathcal{S}, \mathcal{S}'] \ \ (\textit{union bound}) \\
&\leq |\mathcal{H}'| 2^{-m\epsilon/2} \ \ (\textit{step 4}) \\
&\leq \Pi_{\mathcal{H}}(2m) 2^{-m\epsilon/2}
\end{aligned}$$

## Step 6: $Pr[B''] \leq \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2}$

By marginalization $Pr[a] = E_x[Pr[a|x]]$, we have

$$Pr[B''] = E_{\mathcal{S}, \mathcal{S}'}[Pr[B''|\mathcal{S}, \mathcal{S}']]$$

## Step 7: Back to theorem

Putting the above 6 steps altogether,

$$\begin{aligned} Pr[B] &\leq 2 \cdot Pr[B'] \\ &= 2 \cdot Pr[B''] \\ &\leq 2 \cdot \Pi_{\mathcal{H}}(2m) \cdot 2^{-m\epsilon/2} \\ &\leq \delta \end{aligned}$$

Solving for $\epsilon$ we have,

$$\epsilon = \frac{2}{m}(\lg \Pi_{\mathcal{H}}(2m) + \lg(1/\delta) + 1)$$

which is obviously

$$O\left(\frac{\lg \Pi_{\mathcal{H}}(2m) + \lg 1/\delta}{m}\right).$$