Chapter - 2
Theory of Learning

Aviral Janveja

1 Can We Learn?

Revisiting the machine learning process described in Chapter 1 :

- We have an unknown target function f, which represents the underlying pattern that we would like to uncover.
- Next, we have the data-set, which is a set of observations that will be used to approximate the unknown target function.
- Finally, our approximation of the target function, called the final hypothesis function h_f, which is selected by the learning algorithm from the hypothesis set $H = \{h_1, h_2, h_3...h_m\}$, using the sample of data that we have.

However, an important question remains, how can we be sure that a hypothesis function that performs well on the available data points, will generalize well out-of-sample too, on fresh unseen data points? Before trying to answer that question, let us formalize what we mean by performance of a hypothesis function h on the available data points :

- For each point $\mathbf{x}_n$ in our data-set, if $h(\mathbf{x}_n) \neq y_n$, then it constitutes an **error**.
- The sum-total of errors that h makes in-sample, is called the total in-sample error or E_{in}.
- Naturally, we try to minimize E_{in} as much as possible. But does a small in-sample error E_{in} imply a small out-of-sample error E_{out} as well?

2 Probability to the Rescue

In order to tackle the fundamental question raised above, let us continue with the loan example from Chapter 1 and frame it in a simple probabilistic setting.

Think of the general population of a city as a bin of red and green marbles, where each marble represents an individual customer. The green marbles represent correct predictions by our hypothesis, while the red marbles represent errors. Finally, our data-set corresponds to a sample drawn from that bin.

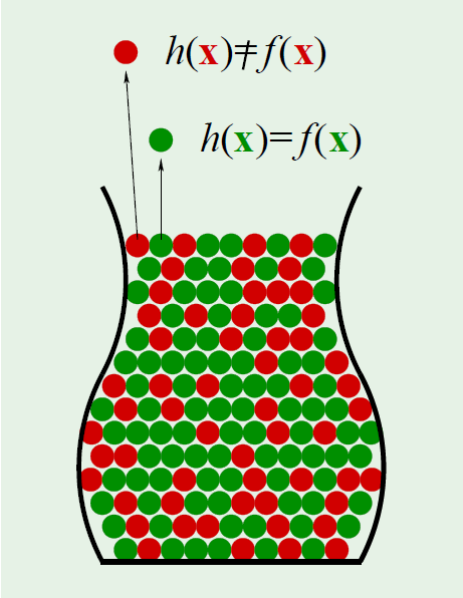

Figure 1: The Bin Example

Now, consider a specific bin consisting of 100 marbles, 90 green and 10 red. Suppose we pick 10 marbles out of that bin independently. Is it possible that all of them or most of them turn out to be red?

Well, it is possible, but is it probable? Not really!

Computing the exact probability shows that the chances of getting 5 or more reds is only about 0.16%. In other words, obtaining a mostly red sample from a mostly green bin is highly unlikely. **Therefore**, as per the laws of probability, we can expect E_{in} to track E_{out}, especially as the sample-size grows.

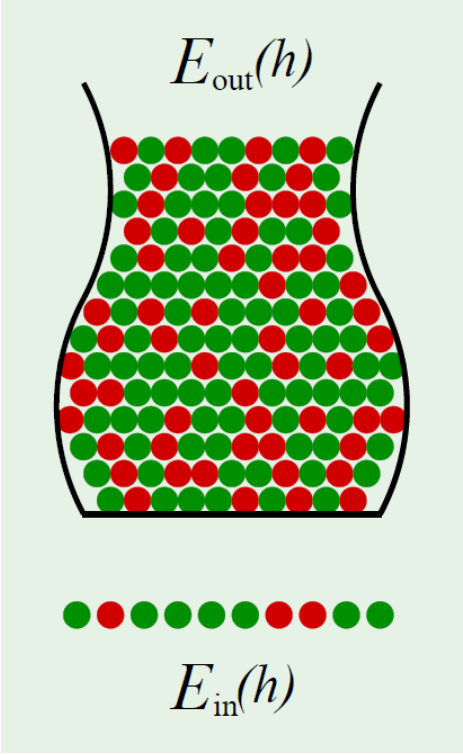

Figure 2: Possible Vs Probable

This intuition is indeed formalized by the Hoeffding Inequality as follows :

$$P[\ |E_{in}(h) - E_{out}(h)| > \epsilon\ ] \leq 2e^{-2\epsilon^2 N}$$

In simple words :

> The probability of E_{in} and E_{out} diverging will be low, if we have accept a reasonable approximation and have a lot of data points.

Let us clarify the terms in the above inequality :

- $E_{in}(h)$: in-sample error of a hypothesis.
- $E_{out}(h)$: out-of-sample error of a hypothesis.
- ϵ : The tolerance value, which is the maximum acceptable difference between E_{in} and E_{out}.
- N : Number of available data points.

It can be observed from the above expression that a **large** N makes the bound tighter, giving more confidence in generalization, whereas a **small** ϵ makes the bound looser, since such precision is harder to guarantee.

3 Connection to Learning

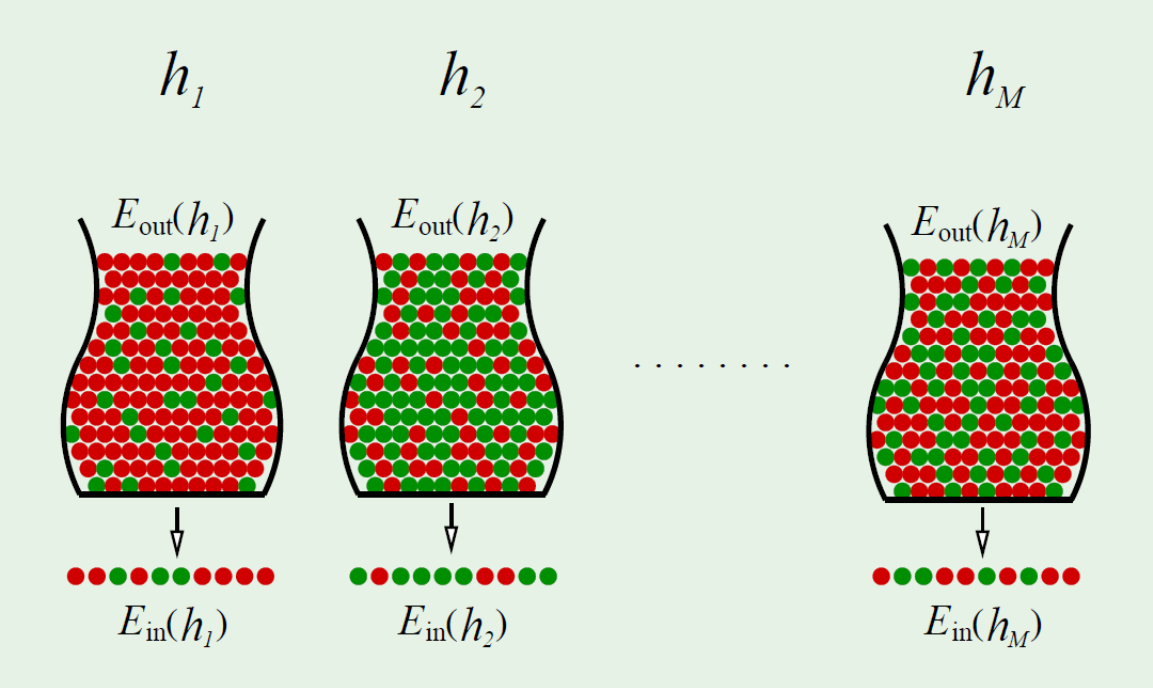

Figure 3: Enter Caption

The M Term via the union bound. We are bounding a hypothesis set H, out of which we select one hypothesis. Not a single pre-fixed hypothesis h. Otherwise that would be verification of h, not learning which is about selecting g from a set of h(s).

Plan

How to structure the chapter going forward.

1. Till now, we have established that in-sample and out-of-sample errors are likely to track each other as per the laws of Probability, formalized via the Hoeffding Inequality.

2. Now, Hoeffding deals with a single hypothesis h. However, learning involves exploring multiple candidate hypothesis h to arrive at the final hypothesis g. So, how do we modify Hoeffding to capture the real learning scenario?

3. Lecture 5 : Taking care of multiple hypotheses. Now, possible hypothesis can be infinitely many. But we need to first look at which are the relevant hypothesis that need to be counted = Dichotomies, Growth Function and Break Point.

4. Lecture 6 : Through which we will arrive at the the VC inequality. Which is a modification to the Hoeffding Inequality that captures the real learning scenario.

5. Lecture 7 : Finally, we finish with the final theory notations of VC dimension and Generalization Bound. Lecture 8 is filler as well just like lecture 4, might just add a few points from that.

4 References

1. CalTech Machine Learning Course - CS156, Lecture 2.

2. CalTech Machine Learning Course - CS156, Lecture 4.

3. CalTech Machine Learning Course - CS156, Lecture 5.