Chapter - 2
Theory of Learning

Aviral Janveja

1 Can We Learn?

Revisiting the process of learning introduced in chapter one : We have an unknown target function f, a hypothesis set H and the final hypothesis function h_f chosen by the learning algorithm using the data points available to us. However, an important question arises : How can we be sure that a hypothesis function that performs well on the available data points, will perform well out-of-sample, on fresh unseen data points?

Before trying to answer that question, let us formalize what we mean by performance of a hypothesis function on the available data points in terms of **error**. 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 hence called the 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 one and frame it as a simple probability experiment :

Think of the general population of a city as a pot of marbles, where each marble represents an individual customer. Grey marbles represent fresh unseen data points, Green marbles represent correct predictions by our hypothesis function and the Red marbles represent errors. Finally, the data-set corresponds to a sample drawn from that bin.

In line with the above example, E_{in} will then correspond to the fraction of red marbles in our sample, while E_{out} will stand for the unknown fraction of red marbles in the pot.

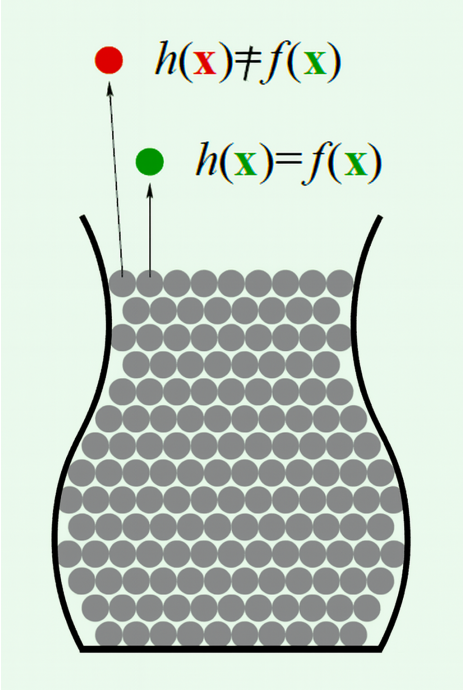

Figure 1: The Pot of Marbles

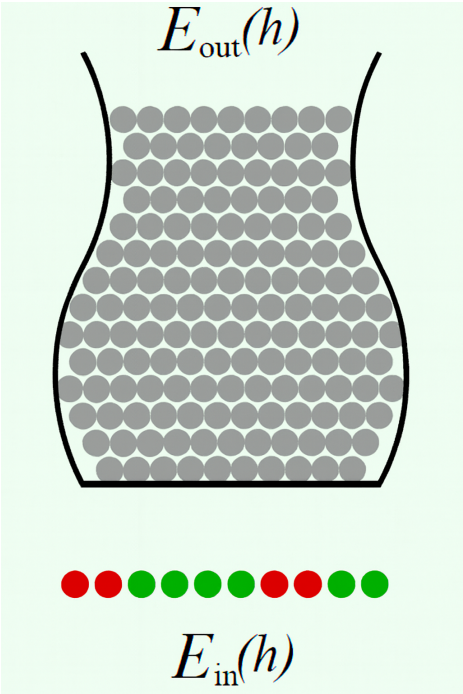

Figure 2: Possible Vs Probable

Consider a specific bin consisting of 100 marbles, 90 green and 10 red. Suppose we pick a sample of 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 in a sample of 10 from that bin, is only about 0.16%. In other words, obtaining a mostly red sample from a mostly green bin is highly unlikely.

Hence, as per the laws of probability, we can expect E_{in} and E_{out} to track each other, especially as the sample-size grows, that is :

$$E_{in}(h) \approx E_{out}(h)$$

This probabilistic understanding is 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} differing from each other will be low, if we 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 probabilistic bound tighter, giving us more confidence for out-of-sample performance, whereas a **small** ϵ makes the bound looser, since such precision is harder to guarantee.

3 Connection to Learning

In the previous section, we have established that E_{out} is likely to track E_{in} for a given hypothesis h. However, in a real learning scenario, we are not just given a final hypothesis function directly. We need to search through the hypothesis set, based on the data points in order to find the best hypothesis function. The real learning scenario is depicted in the following image :

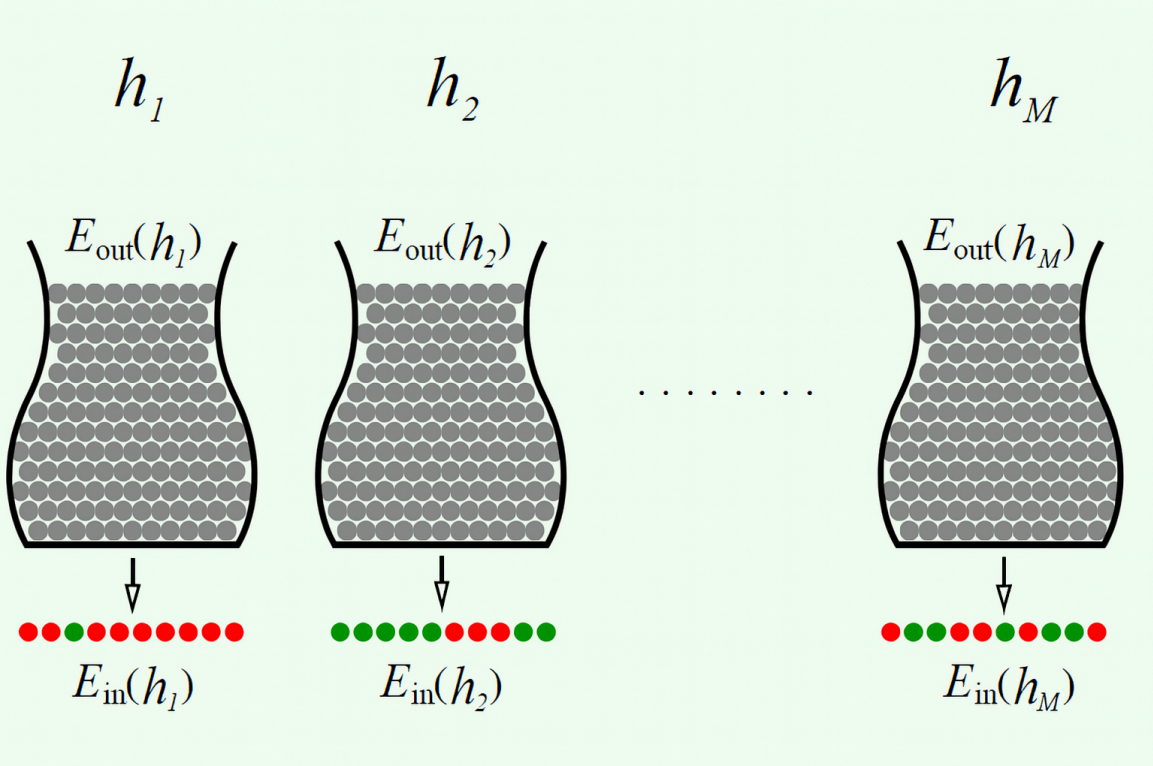

Figure 3: Multiple Hypothesis

Each hypothesis defines it own pot as the fraction of red marbles depends on h, both Ein(h) and Eout(h). So does Hoeffding inequality apply to the multiple pot scenario as well. Let us continue with our probability example. If we consider 20 bins, each mostly green. What is the probability of finding a sample that has 5 or more reds. We see that the chances have increased markedly. This is because when considering multiple hypothesis, the chances of one of them going wayward by fluke increases as we increase the number of hypothesis under consideration. Hence as per the laws of probability, we will need to account for that in the Hoeffding inequality. Finding the exact probability in the learning situation will require us to look closely at hypothesis set and the data points to figure how different hypothesis functions overlap with each other. For now, we look at a simpler solution, we can ignore the overlaps between different hypotheses and use the union bound to bound the overall probability of the learning process.

This adds an M term. Which might be infinite for simple models like perceptron. We will see that there is a way to bound probability even when possible hypothesis are infinitely many. For that we will have to consider the overlap between the different hypothesis functions and their performance as visible to us on the available data points.

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.