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.

But, how can we be sure that a hypothesis function that performs well on the available data points, will generalize well out-of-sample too? Remember, the goal in machine learning is not just to perform well in-sample, but to approximate f as well, that is $h_f \approx f$. Hence, an important question remains, does our hypothesis function perform well out-of-sample, on fresh unseen data points?

Combine this with this

First, let us formalize what we mean by performance of a candidate hypothesis function h on the available data points :

For each point $\mathbf{x}_n$ in our data-set. If $h(\mathbf{x}_n) \neq f(\mathbf{x}_n)$, that is, if our hypothesis function disagrees with the given output, then it constitutes an **error**. Now, 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, which is what actually matters for learning?

2 Probability to the Rescue

Instead of bin example first and connection to learning later, connect the bin example with learning example from the start. Then go on to connection to real learning directly. Instead of current image, add images from slide 9, 11, 12

In line with the Perceptron example from chapter 1, think of the bin representing the general population and the sample representing the data-points that come from the very same population distribution. The green marble is when our hypothesis function gets it right and the red marble represents an error.

In order to answer the above question, let us borrow a simple example from the probability domain. Consider a bin filled with 100 marbles, 90 green and 10 red. Suppose you randomly pick 10 marbles out of the bin. Is it possible that all of them or most of them turn out red? Well yes, it is possible, but is it probable? Not really.

Computing the exact values, shows that the probability of getting 5 or more reds in a sample of 10, taken from the above bin is :

$$P[Red \geq 5] \approx 0.0016$$

Therefore, getting a mostly red sample from a mostly green bin is highly unlikely.

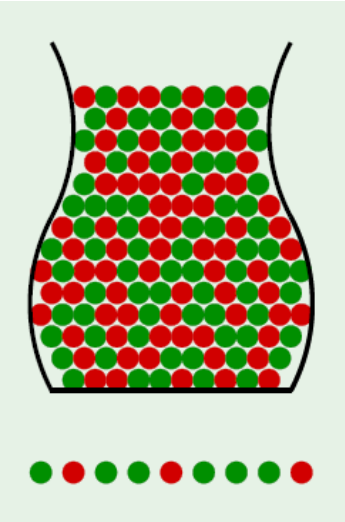

Figure 1: Bin Example

We can similarly compute the probabilities of getting 1 red marble, 2 red marbles and so on in our sample of 10. We can also compute the above results for a different sample size of 20 marbles. Computing these results, clearly shows us that the bin frequency and the sample frequency are likely to track each other, not perfectly but closely, especially as the sample size increases.

Therefore, answering the question raised above, we can safely claim as per the laws of probability, that the in-sample error E_{in} is likely to track the out-of-sample error E_{out}, especially when the sample-size is sufficiently large. Formally, E_{in} is related to E_{out} through the Hoeffding Inequality :

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

Let us clarify the different terms mentioned in the above expression. $E_{in}(h)$ is the in-sample error of our candidate hypothesis, $E_{out}(h)$ is the out-of-sample error, ϵ is the tolerance value, that is the difference between E_{in} and E_{out} that is acceptable to us. N is the sample size, that is the number of data points available to us. Hence, the above expression can be put into simple words as follows :

> The probability of E_{in} and E_{out} diverging beyond our tolerance value ϵ will be low, if we have a reasonable ϵ and a lot of data points N.

It can be seen in the above expression that, having a lot of data points (large N) reduces the value on the right-hand-side, thus giving us a stronger probability bound. Whereas, a small tolerance (ϵ) tends to increase the right-hand-side value, hence weakening our probability bound.

3 Connection to Learning

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.