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 a set of observations that will be used to approximate the unknown target function. Finally, our approximation of the target function, called the hypothesis function g, based on the sample of data that we have.

Now, our hypothesis function performs well on the available data points. However, remember the goal in machine learning is not for g to perform well in-sample, but for g to approximate f well, that is $g \approx f$. Hence, the important question remains, does our hypothesis function perform well out-of-sample on fresh unseen data points?

2 Answer : Probably, Approximately.

First, let us formalize the performance of a candidate hypothesis h in terms of its agreement or disagreement with f on the available data points.

For each point $\mathbf{x}$ in our data sample. If $h(\mathbf{x}) \neq f(\mathbf{x})$, that is, if our candidate hypothesis disagrees with the given output, then it constitutes an **error**. Now, the sum-total of errors that our hypothesis h makes in-sample, would then be 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?

2.1 Possible Vs Probable

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, from the above bin is :

$$P[Red \geq 5] \approx 0.000672$$

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

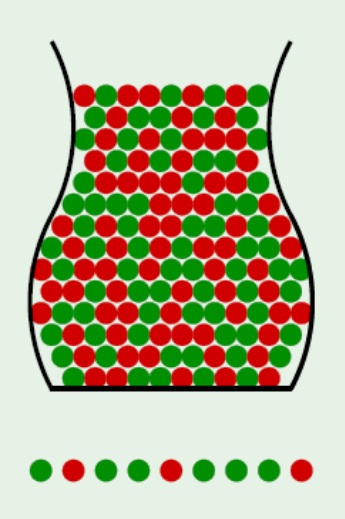

Figure 1: Bin Example

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.

2.2 Hoeffding Inequality

Connecting the bin example to the learning scenario :

The bin represents the general population and the sample represents our 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. **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. It is one of the laws of large numbers, from the field of statistics :

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

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 Real Learning

Ch 5 It is unfeasible to do away with the M Term for number of hypotheses. Add multiple hypotheses for the real learning scenario. Add Union bound in a general way to multiply the M term. Then we can stick to the perceptron example to How do we capture the actual leaning scenario which involves looking through multiple hypothesis before finding the best one. Especially fora model like perceptron with infinitely many possible hypotheses ? First : Perceptron has infinitely many possible hypotheses. How do we deal with it. Dichotomies, data set shattering, break-point and growth function.

4 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.

5 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.