Chapter - 2
Theory of Learning

Aviral Janveja

1 Can We Generalize?

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 is based on the sample of data that we have.

Our hypothesis function g may perform well on the available data points. However, remember that the goal in machine learning is not for g to perform well in-sample but for g to approximate f well, that is $g \approx f$. So, how do we ensure that our hypothesis function generalizes well out of sample for fresh unseen data?

2 Answer : Probably, Approximately.

1. Circumventing the bin example and going directly as per the learning process.

2. The performance of hypothesis g in sample can be formalized in terms of E_{in}. This is the error-rate in-sample or the number of data-points our hypothesis got wrong. E_{out} is what we care about, the error-rate out of sample.

3. Does E_{in} track E_{out} well? Lower Eout means g approximates f well out of sample as well. Hence, Eout is what we care about. We use Ein to get a probabilistic bound on Eout via the Hoeffding inequality(from the law of large numbers in statistics. Adapted for our use-case in ML)

4. Intuitively, if the sample size is big, then it should help. If error tolerance is not too strict, approximation is enough, that should help too. Finally M, which is the number of hypothesis, which is infinite for most relevant models. But this is not our final result in the theory of learning, we will deal with M going forward.

Intuitively. The probability of in sample and out of sample diverging will be low if you have reasonable error tolerance e and a lot of data points N. Model complexity denoted by M = the number of hypothesis.

3 References

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