Learning in a Probabilistic Framework

These notes are slightly edited from scribe notes in previous years.

1 Introduction: What is Learning?

We gave some examples of this. The goal in the course will be to give precise mathematical models of learning phenomena, algorithms for them and an analysis explaining when and why they succeed.

2 Game Number 0

Consider a learning problem where there are only 3 possible examples, and any binary concept is possible - there are 8 possible concepts, each including a subset of the examples as positive examples and the complement as negative.

Now consider observing a “large sample” of 2 examples in this problem. The only remaining unknown fact is the label (positive or negative) of the third example. What prediction should we make? and what guarantees can we give for our prediction?

We discussed this (in a non precise, hand waving manner) and argued that there is no information about the last label and we cannot do better than random guessing. This type of argument has been formalized in many different contexts, and is sometimes known as “no free lunch theorem”. In our context, the conclusion is that we must make some assumption about the range of possible concepts in order to get any guarantee per generalization. We make such assumptions in the next game.

3 Game Number 1

Examples An example is a number in $X = \{1, 2, ..., 10^{15}\}$.

Hypothesis We fix some representation language. A hypothesis is any subset of $X$ that can be described using a representation of 100 bits in our representation. We assume there is a true concept labeling the examples.

Procedure

1. Take a sample of 300 examples, sampled IID (with any distribution), and store these with their labels as the training set.

2. Find a hypothesis $\hat{h}$ that classifies all 300 examples correctly.
Can we claim that $\hat{h}$ is good in some sense?

- We could take another sample and test on it. If we did and got, say, 90% accuracy, we could show that this observed error was within some range of the true error, based on the distribution we used to select the samples.

- But this is not the question posed. We want to draw conclusions based on only the information known at this point without using additional examples.

- Can this be analyzed? Define: $h$ is good when $\text{Prob}[h(x) \text{ is correct}] \geq 0.75$ and $h$ is bad otherwise.

- Next, we fix any bad hypothesis $h$ (one with at least 25% error). Now what is the probability that $h$ is correct on the training sample?

$$\text{Prob}[h \text{ is correct on 300 IID examples}] \leq 0.75^{300} < 2^{-120}$$

- What, then, is the probability that there is any bad hypothesis that correctly classifies 300 IID examples? We can use a union bound here: $\text{Pr}[A \cup B] \leq \text{Pr}[A] + \text{Pr}[B]$. The probability that there exists any bad $h_i$ still consistent with the data:

$$\text{Pr}[h_1 \text{ is bad & consistent with data}] \bigcup \text{Pr}[h_2\ldots] \bigcup \ldots \bigcup \text{Pr}[h_{2^{100}}\ldots] \leq 2^{100} \times 2^{-120} = 2^{-20}$$

- So with probability $\geq (1 - 2^{-20})$ no bad hypothesis is consistent with 300 IID examples, and therefore $\hat{h}$ cannot be bad.

- The facts used to prove this upper bound are: we get 300 IID examples in the training set; there are only $2^{100}$ hypotheses; the true hypothesis is in this set; the distribution used in training is the same as the one used in testing to evaluate the hypothesis.

4 Game Number 2

This illustrates a case with an infinite set of hypotheses.

Examples An example is a real number.

Hypothesis A hypothesis is any interval $[a, b]$. We assume there is a true concept labeling the examples.

Situation Given 3000 IID examples, we want at most 1% error rate. Below is a visual representation of a subset of these examples, illustrating the domain:

$$\leftarrow - - [ + + + + + + ] \longrightarrow - - - - - -$$

The brackets mark the true concept, in relation to the sample data.

Picking a consistent hypothesis There is a wide range of hypotheses consistent with the sample data, ranging from the smallest interval including all +’s to the widest interval excluding all -’s. Because there is an infinite number of hypotheses, we cannot utilize the same proof from Game 1. In the following, consider $\hat{h}$ = the tightest fit around positive examples.
What claims can we make?

• Suppose we place two points, a+ and b−, on the continuum such that:

\[ a^+ = \text{argmin}_x \Pr[a, x] \geq \frac{0.01}{2} = \frac{1}{200} \]

\[ b^- = \text{argmax}_y \Pr[y, b] \geq \frac{0.01}{2} = \frac{1}{200} \]

Note: both points are defined relative to distribution, not relative to the learner.

• What is the probability that an IID sample of 3000 examples does not hit \([a, a^+]\) even once?

\[ (1 - \frac{1}{200})^{3000} \leq 2^{-21} \]

The same holds for the interval \([b^-, b]\).

• Therefore, the probability that the sample has at least one example in \([a, a^+]\) and at least one example in \([b^-, b]\) is at least \(1 - (2^{-21} + 2^{-21}) = 1 - 2^{20}\).

• In other words, with probability \(\geq 1 - 2^{-20}\) the hypothesis has error smaller than 1% because the probability that the hypothesis is wrong is defined as the event that there is no example in that region:

\[ \text{err}(\text{hyp}) = \Pr[a, x_{\text{min}}] + \Pr[x_{\text{max}}, b] \leq \Pr[a, a^+] + \Pr[b^-, b] = \frac{1}{200} + \frac{1}{200} = \frac{1}{100} \]

5 A Probabilistic Model for Learning

In both cases we can guarantee: when using an IID sample, with high probability the chosen hypothesis has a small error. In the next lecture we will generalize this to define the so called Probably Approximately Correct (PAC) learning model.