Learning in a Probabilistic Framework

These notes are slightly edited from previous scribe notes (in Spring 2006) taken by Noah Smith.

1 Introduction: What is Learning?

Learning may be defined as changing behavior in order to perform a task better over time. Several important talks in machine learning are:

- Supervised Classification
- Regression
- Unsupervised Learning
- Mixed Supervised & Unsupervised
- Active Learning

The focus of the course is on supervised learning but we will touch on other models as well. 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 1

Examples Every 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 (or buy or ...) a hypothesis \( \hat{h} \) that classifies all 300 examples correctly.
Can we claim that $h$ 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?

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

- 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:

\[
\text{Pr}[h_1 \text{ is bad & consistent with data}] \cup \text{Pr}[h_2...] \cup \text{Pr}[h_{2100}...] \leq 2^{100} \times 2^{-120} = 2^{-20}
\]

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

- The facts used to prove this upper bound are: there were exactly 300 IID examples in the training set, and only $2^{100}$ hypotheses exist. If there were an infinite number of hypotheses, or if the examples were not selected IID, this proof cannot be used.

3 Game Number 2

Examples Every example is a single 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:

\[
\leftarrow - - [ + ++ + +++ + + ] --- - - - - - - - \rightarrow
\]

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 $\hat{h} = \text{the tightest fit around positive examples.}$
What claims can we make?

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

\[
\begin{align*}
    a^+ &= \arg\min_x Pr[a, x] \geq 0.01 2 = \frac{1}{200} \\
    b^- &= \arg\max_y Pr[y, b] \geq 0.01 2 = \frac{1}{200}
\end{align*}
\]

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?

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

Ditto 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:

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

4 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.