CS156 Yaser Abu-Mostafa Notes

Brad Burkman

3 February 2021

Contents

1 Lecture 1

Lecture 1 of 18 of Caltech's Machine Learning Course - CS 156 by Professor Yaser Abu-Mostafa
`https://www.youtube.com/watch?v=mbyG85GZ0PI&hd=1`

1.1 Requirements for Machine Learning

- A pattern exists
- We cannot pin it down mathematically
- We have data on it

1.2 Components of Learning

- Input: $\mathbf{x}$, a d-dimensional vector
- Output: $y \in \{-1, 1\}$
- Target Function: $f : \mathcal{X} \to \mathcal{Y}$ (unknown)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ (historical records), with N as the number of data points.
- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (known, because we created it. The goal is to find a hypothesis g that approximates f well, $g \approx f$.
- Learning Algorithm $\mathcal{A}$
- Hypothesis Set $\mathcal{H}$ Set of candidate formulas.

2 Lecture 2

`https://www.youtube.com/watch?v=MEG35RDD7RA&t=3s`

Review of Requirements for Machine Learning

A pattern exists	Whether or not a pattern exists, we can apply learning algorithms. At the end of the process, we will have methods for determining whether we have actually learned something, *i.e.* whether there actually was a pattern and whether we actually found a good approximation for it. Not knowing whether there's a pattern should not keep us from applying learning methods. If we know there isn't a pattern, however, we would be wasting our time.
We cannot pin it down mathematically	If we could pin it down mathematically, machine learning is not the recommended technique, but it will work, approximately.
We have data on it	Can't do machine learning without data.

"In science and in engineering, you go a huge distance by settling not for absolutely certain, but almost certain. It opens a world of possibilities."

ϵ is tolerance

Hoeffding's Inequality $\qquad \mathbb{P}[|\nu - \mu| < \epsilon] \leq 2e^{-2\epsilon^2 N}$

We will use this inequality throughout the course.

P.A.C. Probably Approximately Correct

Hoeffding's Inequality is true for all N and ϵ. Not asymptotic. Belongs to the Laws of Large Numbers, but doesn't just apply to large numbers, although it's only relevant for large numbers to get any useful results in practice.

Bound (right side) does not depend on μ, which is unknown.

Difference between *verification* and *learning*. Learning is searching the spa ce $\mathcal{H}$ to deliberately find a hypothesis h that fits the data well.

2.1 Vocabulary

ν is *in sample* denoted by E_{in}

μ is *out of sample* denoted by E_{out}

Given a hypothesis h, we have $E_{\text{in}}(h)$ and $E_{\text{out}}(h)$.

Hoeffding becomes $\qquad \mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}$

There's a problem, that if we have M hypotheses, and g is one of them, then

$$
\begin{aligned}
\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] &\le \mathbb{P}\Big[|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\
&\qquad \text{or } |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\
&\qquad \dots \\
&\qquad \text{or } |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\Big] \\
&= \sum_{m=1}^{M} \mathbb{P}[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon] \\
&= \sum_{m=1}^{M} 2e^{-2\epsilon^2 N} \\
&= 2Me^{-2\epsilon^2 N}
\end{aligned}
$$