ECE595 / STAT598: Machine Learning I

Lecture 25.1: Generalization Bound - M Hypothesis

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Outline

- Lecture 25 Generalization
- Lecture 26 Growth Function
- Lecture 27 VC Dimension

Today's Lecture:

- M Hypothesis
  - PAC framework
  - Guarantee and Possibility
  - The M factor
- Generalization Bound
  - $\mathcal{H}$
  - f
  - Lower and upper limits
- Handling M hypothesis
  - A preview

Probably Approximately Correct

- **Probably**: Quantify error using probability:

$$\mathbb{P}\left[\,|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right] \geq 1 - \delta$$

- **Approximately Correct**: In-sample error is an approximation of the out-sample error:

$$\mathbb{P}\left[|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right] \geq 1 - \delta$$

- If you can find an algorithm $\mathcal{A}$ such that for any ϵ and δ, there exists an N which can make the above inequality holds, then we say that the target function is **PAC-learnable**.

Guarantee VS Possibility

Difference between deterministic and probabilistic learning.

- **Deterministic**:
- "Can $\mathcal{D}$ tell us something *certain* about f outside $\mathcal{D}$?"
- The answer is NO.
- Anything outside $\mathcal{D}$ has uncertainty. There is no way to deal with this uncertainty.
- **Probabilistic**:
- "Can $\mathcal{D}$ tell us something *possibly* about f outside $\mathcal{D}$?"
- The answer is YES.
- If training and testing have the same distribution $p(\boldsymbol{x})$, then training can say something about testing.
- Assume all samples are independently drawn from $p(\boldsymbol{x})$.

One Hypothesis versus the Final Hypothesis

- In this equation

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

  the hypothesis h is *fixed*.
- This h is chosen **before** we look at the dataset.
- If h is chosen **after** we look at the dataset, then Hoeffding is invalid.
- We have to choose a h from $\mathcal{H}$ during the learning process.
- The h we choose depends on $\mathcal{D}$.
- This h is the final hypothesis g.
- When you need to choose g from $h_1, \ldots, h_M$, you need to repeat Hoeffding M times.

The Factor "M"

You can show that

$$|E_{\rm in}(g) - E_{\rm out}(g)| > \epsilon \quad \Longrightarrow \quad \begin{array}{rl} & |E_{\rm in}(h_1) - E_{\rm out}(h_1)| > \epsilon \\ \textbf{or} & |E_{\rm in}(h_2) - E_{\rm out}(h_2)| > \epsilon \\ & \ldots \\ \textbf{or} & |E_{\rm in}(h_M) - E_{\rm out}(h_M)| > \epsilon. \end{array}$$

- To have g, you need to consider $h_1, \ldots, h_M$
- You don't know which h_m to pick; So it is a "OR"
- So there is a sequence of "OR"

The Factor "M"

$$
\begin{aligned}
\mathbb{P}\Big\{ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \Big\} \overset{(a)}{\leq} \mathbb{P}\Big\{ \quad & |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\
\textbf{or} \quad & |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\
& \ldots \\
\textbf{or} \quad & |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon \quad \Big\} \\
\overset{(b)}{\leq} \sum_{m=1}^{M} \mathbb{P}\Big\{ \quad & |E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon \quad \Big\}
\end{aligned}
$$

- We need two identities
- (a) If-statement. $\mathbb{P}[A] \leq \mathbb{P}[B]$ if $A \Rightarrow B$
- (b) Union Bound. $\mathbb{P}[A \textbf{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

The Factor "M"

- Change this equation

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

- to this equation

$$\mathbb{P}\left\{ \left|E_{\text{in}}(g) - E_{\text{out}}(g)\right| > \epsilon \right\} \leq 2Me^{-2\epsilon^2 N}.$$

- So what? M is a constant.
- Bad news: M can be large, or even ∞.
- A linear regression has $M = \infty$.
- Good news: It is possible to bound M.
- We will do it later. Let us look at the interpretation first.