ECE595 / STAT598: Machine Learning I

Lecture 25.2: Generalization Bound

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

Learning Goal

- The ultimate goal of learning is to make

$$E_{\text{out}}(g) \approx 0.$$

- To achieve this we need

$$E_{\text{out}}(g) \overset{\approx}{\uparrow} E_{\text{in}}(g) \overset{\approx}{\uparrow} 0$$

Hoeffding Inequality Training Error

- Hoeffding inequality holds when N is large
- Training error is small when you train well

Complex $\mathcal{H}$

- Recall Hoeffding inequality

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

- If $\mathcal{H}$ is complex, then M will be large. So the approximation by Hoeffding inequality will be worsen.
- But if $\mathcal{H}$ is complex you have more options during training. So training error is improved.
- So there is a trade-off:

$$E_{\text{out}}(g) \quad \underset{\uparrow}{\approx} \quad E_{\text{in}}(g) \quad \underset{\uparrow}{\approx} \quad 0$$

worse if $\mathcal{H}$ complex — good if $\mathcal{H}$ complex

- You cannot use a very complex model
- Simple models generalize better

Complex f

- Recall Hoeffding inequality

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

- Good news: Hoeffding is not affected by f
- So even if f is complex, Hoeffding remains
- Bad news: If f is complex, then very hard to train
- So training error cannot be small
- There is another trade-off:

$$E_{\text{out}}(g) \quad \underset{\text{no effect by } f}{\approx} \quad E_{\text{in}}(g) \quad \underset{\text{worse if } f \text{ complex}}{\approx} \quad 0$$

- You can make $\mathcal{H}$ to counteract, but complex $\mathcal{H}$ will make Hoeffding worse.

Rewriting the Hoeffding Inequality

- Recall the Hoeffding Inequality

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

- This is the same as

$$\mathbb{P}\Big\{ |E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon \Big\} > 1 - \delta.$$

- Equivalently, we can say: **with probability** $1 - \delta$,

$$E_{\text{in}}(g) - \epsilon \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon.$$

What is δ?

- Move around the terms, then we have

$$2Me^{-2\epsilon^2 N} = \delta \quad \Rightarrow \quad \epsilon = \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}$$

- Plug this result into the previous bound:

$$E_{\text{in}}(g) - \epsilon \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon.$$

- This gives us

$$E_{\text{in}}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}.$$

- This is called the **generalization bound**.

Interpreting the Generalization Bound

$$E_{\text{in}}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}.$$

- N: Training sample.
- More is better.
- δ: The probability tolerance level. "Confidence".
- Small δ: You are very conservative. So you need large N to compensate for $\log\frac{1}{\delta}$
- M: Model complexity.
- Large M: You use a very complicated model. So you need large N to compensate for $\log M$

The Two Sides of the Generalization Bound

- **Upper Limit**

$$E_{\text{in}}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}.$$

- $E_{\text{out}}(g)$ cannot be worse than $E_{\text{in}}(g) + \epsilon$.
- Performance guarantee. $E_{\text{in}}(g) + \epsilon$ is the worst you will have. If you can manage this worst case then you are good.
- **Lower Limit**

$$E_{\text{in}}(g) - \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}} \leq E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2M}{\delta}}.$$

- $E_{\text{out}}(g)$ cannot be better than $E_{\text{in}}(g) - \epsilon$.
- Intrinsic limit of your dataset (which controls N), model complexity (which controls M), and how much you want (which determines δ)