ECE595 / STAT598: Machine Learning I
Lecture 28 Sample and Model Complexity

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

- Lecture 28 Sample and Model Complexity
- Lecture 29 Bias and Variance
- Lecture 30 Overfit

Today's Lecture:

- Generalization Bound using VC Dimension
    - Review of growth function and VC dimension
    - Generalization bound
- Sample and Model Complexity
    - Sample complexity
    - Model complexity
    - Trade off

VC Dimension

Definition (VC Dimension)

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by d_{VC}, is the largest value of N for which $\mathcal{H}$ can shatter all N training samples.

- You give me a hypothesis set $\mathcal{H}$, e.g., linear model
- You tell me the number of training samples N
- Start with a small N
- I will be able to shatter for a while, until I hit a bump
- E.g., linear in 2D: $N = 3$ is okay, but $N = 4$ is not okay
- So I find the **largest** N such that $\mathcal{H}$ can shatter N training samples
- E.g., linear in 2D: $d_{\mathrm{VC}} = 3$
- If $\mathcal{H}$ is complex, then expect large d_{VC}
- Does not depend on $p(\boldsymbol{x})$, $\mathcal{A}$ and f

Linking the Growth Function

Theorem (Sauer's Lemma)

Let d_{VC} be the VC dimension of a hypothesis set $\mathcal{H}$, then

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i}. \tag{1}$$

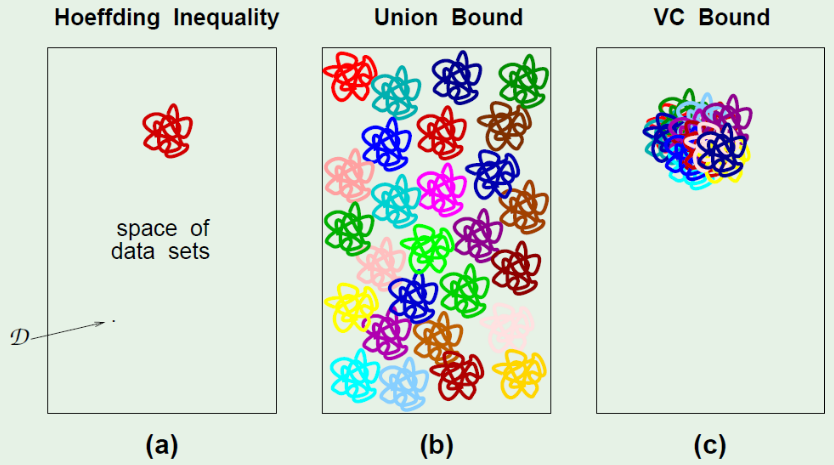
- I skip the proof here. See AML Chapter 2.2 for proof.
- What is more interesting is this:

$$\sum_{i=0}^{d_{\mathrm{VC}}} \binom{N}{i} \leq N^{d_{\mathrm{VC}}} + 1.$$

  This can be proved by simple induction. Exercise.
- So together we have

$$m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{VC}}} + 1.$$

Difference between VC and Hoeffding



Generalization Bound Again

- Recall the generalization bound

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

- Substitute M by $m_{\mathcal{H}}(N)$, and then $m_{\mathcal{H}}(N) \le N^{d_{\text{VC}}} + 1$:

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\log\frac{2(N^{d_{\text{VC}}}+1)}{\delta}}.$$

- Wonderful!
- Everything is characterized by δ, N and d_{VC}
- d_{VC} tells us the expressiveness of the model
- You can also think of d_{VC} as the effective number of parameters

Generalization Bound Again

- If $d_{\mathrm{VC}} < \infty$,
- Then as $N \to \infty$,
$$\epsilon = \sqrt{\frac{1}{2N} \log \frac{2(N^{d_{\mathrm{VC}}} + 1)}{\delta}} \to 0.$$
- If this is the case, then the final hypothesis $g \in \mathcal{H}$ will generalize.
- $d_{\mathrm{VC}} = \infty$,
- Then $\mathcal{H}$is as diverse as it can be
- It is not possible to generalize
- Message 1: If you choose a complex model, then you need to pay the price of training sample
- Message 2: If you choose an extremely complex model, then it may not be able to generalize regardless the number of samples

Generalizing the Generalization Bound

Theorem (Generalization Bound)

For any tolerance $\delta > 0$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{\mathbf{8}}{N}\log\frac{\mathbf{4}m_{\mathcal{H}}(\mathbf{2N})}{\delta}},$$

with probability at least $1-\delta$*.*

- Some small subtle technical requirements. See AML chapter 2.2
- How tight is this generalization bound? Not too tight.
- The Hoeffding inequality has a slack. The inequality is too general for all values of E_{out}
- The growth function $m_{\mathcal{H}}(N)$ gives the **worst case** scenario
- Bounding $m_{\mathcal{H}}(N)$ by a polynomial introduces slack

Outline

- Lecture 28 Sample and Model Complexity
- Lecture 29 Bias and Variance
- Lecture 30 Overfit

Today's Lecture:

- Generalization Bound using VC Dimension
  - Review of growth function and VC dimension
  - Generalization bound
- Sample and Model Complexity
  - Sample complexity
  - Model complexity
  - Trade off

Sample and Model Complexity

Sample Complexity

- What is the smallest number of samples required?
- Required to ensure training and testing error are close
- Close = within certain ϵ, with confidence $1-\delta$
- Regardless of what learning algorithm you use

Model Complexity

- What is the largest model you can use?
- Refers to the hypothesis set
- With respect to the number of training samples
- Largest = measured in terms of VC dimension
- Can use = within certain ϵ, with confidence $1-\delta$
- Regardless of what learning algorithm you use

Sample Complexity

- The generalization bound is

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \log \frac{4 m_{\mathcal{H}}(2N)}{\delta}}.$$

- If you want the generalization error to be at most ϵ, then

$$\sqrt{\frac{8}{N} \log \frac{4 m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon.$$

- Rearrange terms and use VC dimension,

$$N \geq \frac{8}{\epsilon^2} \log\left(\frac{4(2N)^{d_{\text{VC}}} + 1}{\delta}\right).$$

- Example. $d_{\text{VC}} = 3$. $\epsilon = 0.1$. $\delta = 0.1$ (90% confidence). Then the number of samples we need is

$$N \geq \frac{8}{0.1^2} \log\left(\frac{4(2N)^3 + 4}{0.1}\right).$$

Sample Complexity

- How to solve for N in this equation?

$$N \geq \frac{8}{0.1^2} \log \left(\frac{4(2N)^3 + 4}{0.1} \right).$$

- Put $N = 1000$ to the right hand side

$$N \geq \frac{8}{0.1^2} \log \left(\frac{4(2 \times 1000)^3 + 4}{0.1} \right) \approx 21,193.$$

- Not enough. So put $N = 21,193$ to the right hand side. Iterate.
- Then we get $N \approx 30,000$.
- So we need at least 30,000 samples.
- However, generalization bound is not tight. So our estimate is over-estimate.
- Rule of thumb, $10 \times d_{\mathrm{VC}}$.

Error Bar

- The generalization bound is

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \log\left(\frac{4\left((2N)^{d_{\text{VC}}} + 1\right)}{\delta}\right)}.$$

- What error bar can we offer?
- Example. $N = 100$. $\delta = 0.1$ (90% confidence). $d_{\text{VC}} = 1$.
-

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{100} \log\left(\frac{4\left((2 \times 100) + 1\right)}{0.1}\right)} \approx E_{\text{in}}(g) + 0.848.$$

- Close to useless.
- If we use $N = 1000$, then

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + 0.301.$$

- Somewhat more respectable estimate.

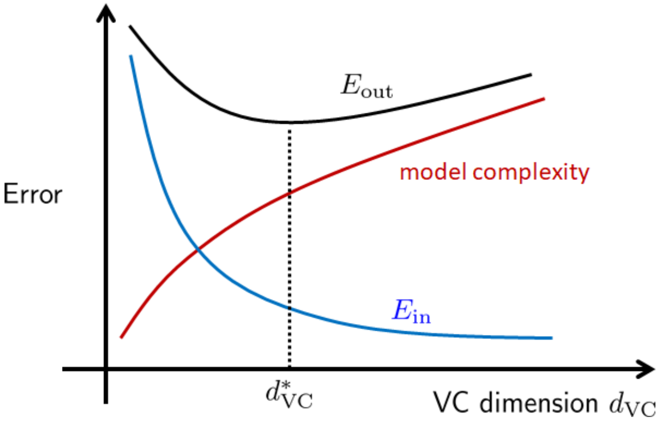
Model Complexity

- The generalization bound is

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \underbrace{\sqrt{\frac{8}{N} \log \frac{4\left((2N)^{d_{\text{VC}}} + 1\right)}{\delta}}}_{=\epsilon(N,\mathcal{H},\delta)}$$

- $\epsilon(N, \mathcal{H}, \delta) =$ penalty of the model complexity
- If d_{VC} is large, then $\epsilon(N, \mathcal{H}, \delta)$ is big
- So the generalization error is large
- There is a trade-off curve

Trade-off Curve



Generalization Bound for Testing

- Testing Set: $\mathcal{D}_{\text{test}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L\}$.
- The final hypothesis g is already determined. So no need to use Union bound.
- The Hoeffding is as simple as

$$\mathbb{P}\Big\{ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \Big\} \le 2e^{-2\epsilon^2 L},$$

- The generalization bound is

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2L} \log \frac{2}{\delta}}.$$

- If you have a lot of testing samples, then $E_{\text{in}}(g)$ will be good estimate of $E_{\text{out}}(g)$
- Independent of model complexity
- Only δ and L

Reading List

- Yasar Abu-Mostafa, Learning from Data, chapter 2.1
- Mehrya Mohri, Foundations of Machine Learning, Chapter 3.2
- Stanford Note http://cs229.stanford.edu/notes/cs229-notes4.pdf