ECE595 / STAT598: Machine Learning I
Lecture 28.1: Sample and Model Complexity - Generalization Bound using VC Dimension

Spring 2020
Stanley Chan

School of Electrical and Computer Engineering
Purdue University
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 $d_{\text{VC}}$, 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: $d_{\text{VC}} = 3$
- If $\mathcal{H}$ is complex, then expect large $d_{\text{VC}}$
- Does not depend on $p(x)$, $\mathcal{A}$ and $f$
Linking the Growth Function

Theorem (Sauer’s Lemma)

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

\[
m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\text{VC}}} \binom{N}{i}.
\]  

(1)

I skip the proof here. See AML Chapter 2.2 for proof.

What is more interesting is this:

\[
\sum_{i=0}^{d_{\text{VC}}} \binom{N}{i} \leq N^{d_{\text{VC}}} + 1.
\]

This can be proved by simple induction. Exercise.

So together we have

\[
m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1.
\]
Difference between VC and Hoeffding

(a) Hoeffding Inequality
(b) Union Bound
(c) VC Bound

space of data sets
Generalization Bound Again

- Recall the generalization bound

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

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

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

- Wonderful!
- Everything is characterized by \( \delta, N \) and \( d_{\text{VC}} \)
- \( d_{\text{VC}} \) tells us the expressiveness of the model
- You can also think of \( d_{\text{VC}} \) as the effective number of parameters
Generalization Bound Again

- If $d_{VC} < \infty$,
- Then as $N \to \infty$,
  \[ \epsilon = \sqrt{\frac{1}{2N} \log \frac{2(N^{d_{VC}} + 1)}{\delta}} \to 0. \]
- If this is the case, then the final hypothesis $g \in \mathcal{H}$ will generalize.
- $d_{VC} = \infty$,
- 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 \( \delta > 0 \)

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

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 \( E_{\text{out}} \).
- The growth function \( m_{\mathcal{H}}(N) \) gives the \textbf{worst case} scenario
- Bounding \( m_{\mathcal{H}}(N) \) by a polynomial introduces slack