Chapter 2: Training vs Testing

Motivation

The Finite Hypothesis Class Generalization (FHCG) Theorem from the previous lecture notes gives us the bound

$$E_{\text{out}} \leq E_{\text{in}} + O\left(\sqrt{\frac{\log M - \log \delta}{N}}\right). \tag{1}$$

This bound is not useful for infinite hypothesis classes because the resulting bound is trivial.

The goal of this chapter is to introduce a new way to measure the "size" of a hypothesis class called the *VC dimension* (d_{VC}). It turns out that many infinite hypothesis classes have finite VC dimension, and the *fundamental theorem of machine learning* states

$$E_{\text{out}} \leq E_{\text{in}} + O\left(\sqrt{\frac{d_{\text{VC}} - \log \delta}{N}}\right). \tag{2}$$

Proving the above bound is rather technical. The textbook proves the slightly weaker bound

$$E_{\text{out}} \leq E_{\text{in}} + O\left(\sqrt{\frac{d_{\text{VC}} \log N - \log \delta}{N}}\right). \tag{3}$$

There are unfortunately a lot of technical details needed to understand the VC dimension. This chapter is the most mathematically difficult part of this course.

Note 1. The textbook contains sections labeled "safe skip" that contain the full proof of the theorem above. You are not responsible for the "safely skippable" portions of the chapter. You are responsible for everything else.

Section 2.1.1 Effective Number of Hypotheses

Definition 1. Let $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathcal{X}$. The *dichotomies* generated by a hypothesis class $\mathcal{H}$ on these points are defined by

$$\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N) = \Big\{ \big(h(\mathbf{x}_1), ..., h(\mathbf{x}_N)\big) : h \in \mathcal{H} \Big\} \tag{4}$$

Example 1. Consider the dataset of 4 points defined by

$$\begin{aligned} \mathbf{x}_1 &= (+1, +1) \\ \mathbf{x}_2 &= (-1, +1) \\ \mathbf{x}_3 &= (+1, -1) \\ \mathbf{x}_4 &= (-1, -1) \end{aligned}$$

and the dataset of 4 points defined by

$$\begin{aligned} \mathbf{x}'_1 &= (+1, +1) \\ \mathbf{x}'_2 &= (-1, -1) \\ \mathbf{x}'_3 &= (+2, +2) \\ \mathbf{x}'_4 &= (-2, -2) \end{aligned}$$

What are the dichotomies generated by $\mathcal{H}_{\text{axis}}$ hypothesis class on these two datasets? Recall that

$$\mathcal{H}_{\text{axis}} = \Big\{ \mathbf{x} \mapsto \text{sign}(x_i) : i \in [d] \Big\}.$$

Definition 2. The *growth function* for a hypothesis class $\mathcal{H}$ is defined to be

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,...,\mathbf{x}_N \in \mathcal{X}} \big|\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)\big|. \tag{5}$$

Problem 1 (Example 2.1, page 43)**.** Let $\mathcal{H}$ be the perceptron hypothesis class in 2 dimensions. What is $m_{\mathcal{H}}(3)$ and $m_{\mathcal{H}}(4)$?

Problem 2. Either prove the following statement, or find a counterexample: For all datasets and all hypothesis classes, $m_{\mathcal{H}}(N) \leq 2^N$.

Definition 3. We say that a hypothesis class $\mathcal{H}$ can *shatter* a dataset $\mathbf{x}_1, ..., \mathbf{x}_N$ if any of the following equivalent statements are true:

1. $\mathcal{H}$ is capable of generating all possible dichotomies of $\mathbf{x}_1, ..., \mathbf{x}_N$.
2. $\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N) = \{-1, +1\}^N$.
3. $|\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)| = 2^N$.

Definition 4. If no data set of size k can be shattered by $\mathcal{H}$, then k is said to be a *break point* for $\mathcal{H}$.

Section 2.1.2: Bounding the Growth Function

Theorem 1. If $m_{\mathcal{H}}(k) < 2^k$ for some value k, then

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{k} = O(N^{k-1}). \tag{6}$$

This implies that, $m_{\mathcal{H}}$ grows exponentially before its first breakpoint, and polynomially thereafter.

Section 2.1.3 / 2.1.4: The VC Dimension

Definition 5. The Vapnik-Chervonenkis dimension (VC dimension) of a hypothesis set $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$ or simply d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{\mathrm{VC}} = \infty$.

Fact 1 (Equation 2.9/2.10, page 50)**.** For all hypothesis classes $\mathcal{H}$, we have that

$$m_{\mathcal{H}}(N) \leq N^{d_{\mathrm{VC}}} + 1 \tag{7}$$

Theorem 2 (VC generalization bound)**.** For any tolerance $\delta > 0$, we have that with probability at least $1 - \delta$,

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

Substituting the bound from Fact 1 above, we get that

$$E_{\text{out}} \leq E_{\text{in}} + \sqrt{\frac{8}{N} \log \frac{4(2N)^{d_{\text{VC}}} + 1}{\delta}} = O\left(\sqrt{\frac{d_{\text{VC}} \log N - \log \delta}{N}}\right). \tag{9}$$

1) constants are very loose

2) why care? if $d_{vc} < \infty$
then learning possible
$N \to \infty \Rightarrow E_{in} \to E_{out}$

Problem 3. What is the VC dimension of the perceptron hypothesis class?

$d=2$ $m_{\mathcal{H}}(3) = 2^3 = 8$ exp.

$m_{\mathcal{H}}(4) = 14 \neq 2^4$ not exp.

$\Rightarrow d_{vc} = 3$

generic d

$\overbrace{d_{vc} \geq k}^{①}$ and $\overbrace{d_{vc} \leq k}^{②} \Rightarrow d_{vc} = k$

general procedure for showing VC dimension

① $d_{vc} \geq d+1$

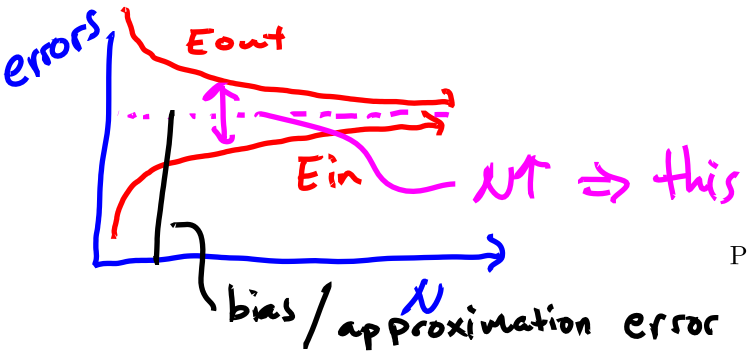
Find any dataset of size $N=d+1$ that that the perceptron shatters $\Rightarrow m_{\mathcal{H}}(N) = 2^N$

② $d_{vc} \leq \boxed{d+1}\ k$

all datasets of size $\underset{=k+1}{\boxed{d+2}}$ cannot be shattered.

claim $d_{vc} = d+1 = \underbrace{\Theta(d)}$

↑ super important



$N\uparrow \Rightarrow$ this gap shrinks

Not in any textbook because it is "too obvious"

Application (not in textbook)

Problem 4. You are a bank using the perceptron to learn a formula for whether or not to issue a loan.

1. You have successfully learned a model on the dataset $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ where each $\mathbf{x}_i$ has d features. Unfortunately, the training error of the model is too high. Management has allocated money to create a new dataset. Your choices are to either spend that money to add new features to the existing dataset, (1) or to add more data points that all have the same features. According to VC theory, which action makes the most sense? (2)

   # data points = N, # dim/features = d

   E_{in} too high

   two choices:

   1) make d larger

   2) add more data $\Rightarrow N$ larger

   only possibility that makes sense, $d\uparrow \Rightarrow E_{in}\downarrow\leftrightarrow$

   $\Rightarrow N\uparrow$ cannot help $\Rightarrow E_{in}\leftrightarrow$ $E_{in}\uparrow$

   the only way that N plays a role is $|E_{in} - E_{out}| \le O\left(\sqrt{\frac{d_{vc}}{N}}\right)$

   $$E_{in} = \frac{1}{N}\sum_{i=1}^{N}[\![h(x_i)\neq y_i]\!]$$

2. You decided to augment the dataset so that it now has $2d$ features instead of only d features. Now the generalization error is too high. According to VC theory, how many more data points will you need in order to achieve the same generalization error that you had before?

   $|E_{in} - E_{out}|$ too high why?

   $d\uparrow \Rightarrow d_{vc}\uparrow \Rightarrow |E_{in} - E_{out}| < O\left(\sqrt{\frac{d_{vc}}{N}}\right)\uparrow$

   if $d_{vc}\uparrow 2x \Rightarrow \boxed{N\uparrow 2x}$ for $O\sqrt{\frac{d_{vc}}{N}}$ to $\leftrightarrow$

   $\mathcal{H}_{perceptron}\{ — : w\in\mathbb{R}^d\}$

   dim=d   dim=2d

   $$\mathcal{H}_1 \le \mathcal{H}_2 \Rightarrow E_{in}(g_2) \le E_{in}(g_1)$$