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. The goal of this chapter is to replace the dependence on M above with a dependence on the "VC dimension" (d_{VC})

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

We will not prove this result. But in order to use this result in practice, we must define the VC dimension and be able to calculate the VC dimension of our hypothesis classes.

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

$$\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N) = \left\{ \big(h(\mathbf{x}_1), ..., h(\mathbf{x}_N)\big) : h \in \mathcal{H} \right\} \tag{3}$$

Example 1. Consider the dataset of 4 points defined by

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

What are the dichotomies generated by the following hypothesis classes on this dataset?

$$\mathcal{H}_{\text{axis}} = \left\{ \mathbf{x} \mapsto \text{sign}(x_i) : i \in [d] \right\}$$
$$\mathcal{H}_{\text{perceptron}} = \left\{ \mathbf{x} \mapsto \text{sign}(\mathbf{w}^T \mathbf{x} + b) : b \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^d \right\}$$

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

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,...,\mathbf{x}_N \in \mathcal{X}} \left|\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)\right|. \tag{4}$$

Definition 3. We say that a hypoothesis 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 $\mathcal{H}$.

Example 2 (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 1. Example 2.2 in the textbook (43-45) contains many more examples of computing the growth function. You should work through and understand all of these examples.

Fact 1. For all datasets and all hypothesis classes, $m_{\mathcal{H}}(N) \leq 2^N$.

Section 2.1.2: Bounding the Growth Function

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

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{k} = O(N^{k-1}). \tag{5}$$

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 $d_{\text{VC}}(\mathcal{H})$ 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 $d_{\text{VC}} = \infty$.

Fact 2 (Equation 2.9/2.10, page 50)**.** For all hypothesis classes $\mathcal{H}$, we have that

$$m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1 \tag{6}$$

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

$$E_{\text{out}} \leq E_{\text{in}} + \sqrt{\frac{8}{N} \log \frac{4 m_{\mathcal{H}}(2N)}{\delta}}. \tag{7}$$

Substituting the bound from Fact 2 above, we get that

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

Problem 2. What is the VC dimension of the perceptron hypothesis class?

Problem 3. 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, or to add more data points that all have the same features. According to VC theory, which action makes the most sense?

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?