Quiz: Chapter 1+2 definitions

Definition 1. The *in-sample error* is defined to be

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{i=1}^{N}[\![h(\mathbf{x}_i) \neq y_i]\!].$$

Definition 2. The *out-of-sample error* is defined to be

$$E_{\text{out}}(h) = \mathbb{P}\big(h(\mathbf{x}) \neq y\big).$$

Definition 3. The true label function is defined to be

$$f = \operatorname*{arg\,min}_{h \in \mathcal{H}^*} E_{\text{out}}(h),$$

where $\mathcal{H}^*$ is the union of all hypothesis classes.

Definition 4. The *generalization error* of a hypothesis g is defined to be

$|E_{\text{in}}(g) - E_{\text{out}}(g)|$.

Definition 5. 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

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

Definition 6. The *growth function* for a hypothesis class $\mathcal{H}$ is defined to be

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

Definition 7. 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. $|\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)| = 2^N$.

NOTE: You must list all 3 for full credit.

Definition 8. The integer k is said to be a *break point* for hypothesis class $\mathcal{H}$ if

no data set of size k can be shattered by $\mathcal{H}$.

Definition 9. The Vapnik-Chervonenkis dimension (VC dimension) of a hypothesis class $\mathcal{H}$, 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 $d_{\text{VC}} = \infty$.

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

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

NOTE: The more precise, non-asymptotic formulas would also be acceptable.