Foundations of Machine Learning: Assignment 3

1 Rademacher complexity

1.1 Properties

Let G, G_1, G_2 denote arbitrary classes of functions $\mathcal{Z} \to [a, b]$, and let c, d be arbitrary real numbers. Show
(a) $\hat{\mathcal{R}}_S(cG + d) = |c|\hat{\mathcal{R}}_S(G)$, where $cG + d := \{h(z) = cg(z) + d | g \in G\}$.
(b) $\hat{\mathcal{R}}_S(\text{conv}(G)) = \hat{\mathcal{R}}_S(G)$, where $\text{conv}(G) := \{\sum_{i=1}^n \alpha_i g_i | n \in \mathbb{N}, \alpha_i \geq 0, \sum_i \alpha_i = 1, g_i \in G\}$.
(c) $\hat{\mathcal{R}}_S(G_1 + G_2) = \hat{\mathcal{R}}_S(G_1) + \hat{\mathcal{R}}_S(G_2)$, where $G_1 + G_2 := \{g(z) = g_1(z) + g_2(z) | g_1 \in G_1, g_2 \in G_2\}$.

1.2 From Zero-one Loss to Hypothesis

Let H be a family of functions taking values in $\{-1, +1\}$ and let G be the family of loss functions associated to H for the zero-one loss: $G = \{(x, y) \mapsto 1_{h(x) \neq y : h \in H}\}$. For any sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ of elements in $\mathcal{X} \times \{1, +1\}$, let $S_{\mathcal{X}}$ denote its projection over $\mathcal{X} : S_{\mathcal{X}} = (x_1, \ldots, x_m)$. Show that the following relation holds between the empirical Rademacher complexities of G and H:

$$\hat{\mathcal{R}}_S(G) = \frac{1}{2}\hat{\mathcal{R}}_{S_{\mathcal{X}}}(H).$$

2 Growth function and VC dimension

2.1 Growth Function of Intervals

(Exercise 3.1 in *Foundations of Machine Learning*) Growth function of intervals in $\mathbb{R}$. Let H be the set of intervals in $\mathbb{R}$. The VC-dimension of H is 2. Compute its growth function $\Pi_H(m)$, $m \geq 0$. Compare your result with the general bound for growth functions.

2.2 Tightness of Sauer's Lemma

(Exercise 3.2 in *Foundations of Machine Learning*) Prove that Sauer's lemma (theorem 3.5 in *Foundations of Machine Learning*) is tight, i.e., for any set X of $m > d$ elements, show that there exists a hypothesis class H of VC-dimension d such that $\Pi_H(m) = \sum_{i=0}^d \binom{m}{i}$.

2.3 VC-dimension of closed balls

(Exercise 3.9 in *Foundations of Machine Learning*) VC-dimension of closed balls in $\mathbb{R}^n$. Show that the VC-dimension of the set of all closed balls in $\mathbb{R}^n$, i.e., sets of the form $\{x \in \mathbb{R}^n : ||x - x_0||^2 \leq r^2\}$ for some $x_0 \in \mathbb{R}^n$ and $r \geq 0$, is less than or equal to $n + 2$.

3 Error bound

Let D be a distribution over $X \times \{0,1\}$, and let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ be a random sample from D. Let

$$err(h) = Pr_{(x,y)\sim D}[h(x) \neq y]$$
$$\widehat{err}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{h(x_i)\neq y_i}$$

For simplicity, we will assume that $\mathcal{H}$ is finite, although the results of this problem can be carried over to the infinite case. Note that none of the results depend on $|\mathcal{H}|$. Let $\hat{h}$ and h^* be the hypotheses in $\mathcal{H}$ with minimum training error and generalization error, respectively:

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \widehat{err}(h)$$
$$h^* = \arg\min_{h\in\mathcal{H}} err(h)$$

Be sure to keep in mind that, unlike h^*, $\hat{h}$ is a random variable that depends on the random sample S.

(a) Prove that

$$\mathbb{E}\left[\widehat{err}(\hat{h})\right] \leq err(h^*) \leq \mathbb{E}\left[err(\hat{h})\right].$$

(b) Prove that, with probability at least $1-\delta$,

$$\left|\widehat{err}(\hat{h}) - \mathbb{E}\left[\widehat{err}(\hat{h})\right]\right| \leq O\left(\sqrt{\frac{ln(1/\delta)}{m}}\right).$$

Give explicit constants, and be sure to end up with a result that does not depend on $|\mathcal{H}|$.

(c) Explain in words the meaning of what you proved in both of the preceding parts, and how we would expect training error to compare to test error when using a machine learning algorithm on actual data.