Telecom Paris
A. Tchamkerten

Introduction to Machine Learning (SI221/MICAS911)

Assignment 3 - Solutions

Exercise 1 (Error decomposition)**.** Let h_S be an ERM$_{\mathcal{H}}$ predictor for some function class $\mathcal{H}$. Write the prediction error $L_P(h_s) = \mathbb{E}_{Z\sim P}(\ell(Z, h_S))$ as

$$L_P(h_s) = \varepsilon_{\text{app}} + \varepsilon_{\text{est}}$$

where $\varepsilon_{\text{app}} := \min_{h\in\mathcal{H}} L_P(h)$ and $\varepsilon_{\text{est}} := L_P(h_s) - \varepsilon_{\text{app}}$. Interpret this error decomposition.

Solution. ε_{app} represents the lowest error probability that can be achieved by any predictor in $\mathcal{H}$ if the data distribution P is known. Since it depends on $\mathcal{H}$, it is sometimes referred to as *inductive bias*, this is the error/bias due to the learner choice of the class of predictors $\mathcal{H}$. The larger the class $\mathcal{H}$ the lower ε_{app}.

On the other hand the estimation error ε_{est} refers to the "error overhead" due to the fact that ERM relies on empirical samples, and is only an approximation of the true (minimal) risk (notice that $\varepsilon_{\text{est}} \leq L_P(h_s)$ as $h_s \in \mathcal{H}$). By constrast with ε_{app}, ε_{est} depends also on the number of samples. Typically, the more the number of samples, the lower ε_{est}.

Therefore, for a given number of samples, reducing the bias implies considering a rich class $\mathcal{H}$. But a rich class is also more prone to overfitting and therefore may increase ε_{est}. Conversely, reducing ε_{est} increases ε_{app}, a scenario sometimes referred to as "underfitting." As an extreme case, if $\mathcal{H}$ consists of all functions, then $\varepsilon_{\text{app}} = 0$ but the NFL theorem implies a huge number of samples if we ever want to achieve a small ε_{est}. $\square$

Exercise 2 (VC dimension, parity)**.** Let $\mathcal{X} = \{0,1\}^n$. Given $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ let

$$h_{\mathcal{I}}(x) = (\sum_{i\in\mathcal{I}} x_i) \mod 2$$

denote the parity of x over the coordinates in $\mathcal{I}$. Show that the VC dimension of the set of all such functions, that is

$$\mathcal{H}_{parity} = \{h_{\mathcal{I}} : \mathcal{I} \subseteq \{1, 2, \ldots, n\}\},$$

is n.

Solution. As an upper bound we have

$$VCdim(\mathcal{H}_{parity}) \leq \log_2(|\mathcal{H}_{parity}|).$$

To show that this bound is tight it suffices to consider the set composed of the basis vectors in $\{0,1\}^n$ $\square$

Exercise 3 (VC dimension, signed intervals)**.** Consider the class of signed intervals over $\mathcal{X} = \mathbb{R}$

$$\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$$

where $h_{a,b,s}(x) = s$ if $x \in [a, b]$ and $h_{a,b,s}(x) = -s$ if $x \notin [a, b]$. Show that VCdim($\mathcal{H}$)=3.

Solution. We first show that there exists a set of cardinality 3 that can be shattered by $\mathcal{H}$. Let $\mathcal{A} = \{1, 2, 3\}$. The following table describes one way (specific choices of a and b) to shatter all possible ways of shattering $\mathcal{A}$ with $\mathcal{H}$:

1	2	3	a	b	s
-	-	-	0.5	3.5	-1
-	-	+	2.5	3.5	1
-	+	-	1.5	2.5	1
-	+	+	1.5	3.5	1
+	-	-	0.5	1.5	1
+	-	+	1.5	2.5	-1
+	+	-	0.5	2.5	1
+	+	+	0.5	3.5	1

Hence, $VCdim(\mathrm{H}) \geq 3$. Now pick any set of cardinality 4 $\mathcal{A} = \{x_1, x_2, x_3, x_4\}$ which we assume, without loss of generality to satisfy $x_1 < x_2 < x_3 < x_4$. Any such set cannot be completely shattered as the labeling $y_1 = y_3 = -1$ and $y_2 = y_4 = 1$ cannot be obtained. □

Exercise 4 (VC dimension, halfspaces)**.** A homogeneous halfspace is specified by a vector $\mathbf{w}$ in $\mathbb{R}^d$ which defines a binary function

$$\mathbf{x} \mapsto h_{\mathbf{w}}(\mathbf{x}) := \text{sign}\langle \mathbf{w}, \mathbf{x} \rangle$$

Show that the VCdimension of the class of homogeneous halfspaces in $\mathbb{R}^d$ is equal to d. Show that the VCdimension of the class of non-homogeneous halfspaces defined by

$$\mathbf{x} \mapsto h_{\mathbf{w},b}(\mathbf{x}) := \text{sign}\langle \mathbf{w}, \mathbf{x} \rangle + b$$

with $\mathbf{w}$ in $\mathbb{R}^d$ and b in $\mathbb{R}$ is $d + 1$.

Solution. See Linear predictors chapter, Theorem 9.2, 9.3 UML book (hardcopy) □

Exercise 5 (VC dimension, bounds)**.** In class we established the upper bound $VCdim(\mathcal{H}) \leq \log(|\mathcal{H}|)$. Here we will show that this bound can be quite loose.

1. Find an example of a class $\mathcal{H}$ of functions on the unit interval $[0, 1]$ such that $VCdim(\mathcal{H}) < \infty$ while $|\mathcal{H}| = \infty$.

2. Find an example of a finite class $\mathcal{H}$ of functions on the unit interval $[0, 1]$ where $VCdim(\mathcal{H}) < \log(|\mathcal{H}|)$.

Solution.

- The class $\mathcal{H}$ of indicator function $1\{x \geq t\}$ with $t \in \mathbb{R}$ is infinite while its VC dimension equals 1.

- The class of functions composed of only $1\{x \geq 1\}$ and $1\{x \leq 1/2\}$ has vcdimension equal to zero.

□