CS 260: Machine Learning	Spring 2020
Homework 2	
Hand Out: April.15	*Due: April.29*

1. **Exercise 5.1**

   Based on the definition and hint (Lemma B.1),

$$\begin{aligned}\mathbb{P}_{S\sim\mathcal{D}^m}\Big[L_{\mathcal{D}}(A(S)) \geq 1/8\Big] &= \mathbb{P}_{S\sim\mathcal{D}^m}\Big[L_{\mathcal{D}}(A(S)) \geq 1 - 7/8\Big] \\ &\geq \frac{\mathbb{E}[L_{\mathcal{D}}(A(S))] - (1 - 7/8)}{7/8} \\ &\geq \frac{1/4 - 1/8}{7/8} \\ &= \frac{1}{7},\end{aligned}$$

   which concludes our proof. ■

2. **Exercise 6.1**

   Given the condition that two hypothesis classes $\mathcal{H}'$, $\mathcal{H}$ satisfy $\mathcal{H}' \subseteq \mathcal{H}$. Then for any subset $C = \{c_1, ..., c_m\} \subseteq \mathcal{X}$, we have $\mathcal{H}'_C \subseteq \mathcal{H}_C$. Suppose C is shattered by $\mathcal{H}'$, then C can be also shattered by $\mathcal{H}$. As a consequence, $\text{VCdim}(\mathcal{H}'_C) \leq \text{VCdim}(\mathcal{H}_C)$ for any set C, therefore, $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$. ■

3. **Exercise 6.2**

   (a) We first show that

$$\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{=k}) \leq k.$$

   Suppose a subset $C \subseteq \mathcal{X}$ satisfies $|C| = k+1$. Then there does not exist $h \in \mathcal{H}^{\mathcal{X}}_{=k}$ such that

$$h(x) = 1, \quad \text{for all } x \in C,$$

   which indicates that $\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{=k}) \leq k$. Similarly, we can find a subset $C' \subseteq \mathcal{X}$ with $|C'| = |\mathcal{X}| - k + 1$ and there is no $h \in \mathcal{H}^{\mathcal{X}}_{=k}$ such that $h(x) = 0$ for all $x \in C$. This implies that $\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{=k}) \leq |\mathcal{X}| - k$. And then we obtain

$$\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{=k}) \leq \min\{k, |\mathcal{X}| - k\}.$$

   Next we will show that $\text{VCdim}(\mathcal{H}^{\mathcal{X}}_{=k}) \geq \min\{k, |\mathcal{X}|-k\}$. Let a subset $C = \{x_1, x_2, ..., x_m\} \subseteq \mathcal{X}$ with size $|C| = m$ and $m \leq \min\{k, |\mathcal{X}| - k\}$, and denote the corresponding labels as $(y_1, y_2, ..., y_m) \in \{0, 1\}^m$. We denote s as

$$s = \sum_{i=1}^{m} y_i.$$

We select a subset $C' \subseteq \mathcal{X} \setminus C$ such that $|C'| = k - s$. Then we select a hypothesis h such that

$$h(x_i) = y_i, \quad x_i \in C$$

and

$$h(x) = 1\{x \in C'\}, \quad x \in \mathcal{X} \setminus C.$$

Thus we can derive that C can be shattered by $\mathcal{H}_{=k}^{\mathcal{X}}$, which indicates that $\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) \geq \min\{k, |\mathcal{X}| - k\}$. Together with the upper bound on $\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}})$, we conclude that $\mathrm{VCdim}(\mathcal{H}_{=k}^{\mathcal{X}}) = \min\{k, |\mathcal{X}| - k\}$.

(b) When we have a set S with the size $2k+2$, there exists a partition that has $k+1$ positive labels and $k+1$ negative labels. For this case, there does not exist a hypothesis in the hypothesis class that correctly predicts the partition. For a set of size $2k+1$, for any partition, there must be either at most k positive instances or at most k negative instances. Thus, we can find a hypothesis in the hypothesis class that predicts the partition correctly. We conclude that the VC dimension is $\min\{2k+1, |\mathcal{X}|\}$. $\blacksquare$

4. **Exercise 6.3** To begin with, we can derive that $|\mathcal{H}_{\text{n-parity}}| = 2^n$. Based on the definition, we can get that

$$\mathrm{VCdim}(\mathcal{H}_{\text{n-parity}}) \leq \log(|\mathcal{H}_{\text{n-parity}}|) = n.$$

Next we need to show that $\mathrm{VCdim}(\mathcal{H}_{\text{n-parity}}) \geq n$. For the basis vectors $\{\mathbf{e}_i\}_{i=1}^n$, and define the corresponding labels $(y_1, ..., y_n) \in \{0,1\}^n$. Define $S = \{i \in [n] | y_i = 1\}$, and let the hypothesis h be $h_S(\mathbf{e}_i) = y_i$. Then we have $\{\mathbf{e}_i\}_{i=1}^n$ can be shattered by $\mathcal{H}_{\text{n-parity}}$, thus $\mathrm{VCdim}(\mathcal{H}_{\text{n-parity}}) = n$. $\blacksquare$

5. **Exercise 6.4**

Throughout this question, we use $\mathcal{X} = \mathbb{R}^d$. We will illustrate the concrete cases: $(<,=)$, $(=,<)$, $(=,=)$ and $(<,<)$.

- $(<,=)$. We consider the hypothesis class $\mathcal{H} = \{1_{[\|\mathbf{x}\|_2 \leq r]} | r \geq 0\}$. The VC-dimension of $\mathcal{H}$ is 1, since there exists $\mathbf{x} \in \mathbb{R}^d$ such that $\{\mathbf{x}\}$ can be shattered by $\mathcal{H}$, and there exist $\{\mathbf{x}_1, \mathbf{x}_2\}$, $\|\mathbf{x}_1\| \leq \|\mathbf{x}_2\|$, such that can not be shattered by $\mathcal{H}$. Let $A = \{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_i$ are standard basis in R^d, then we have $\mathcal{H}_A = \{(0,0), (1,1)\}$, and $\{B \subseteq A | \mathcal{H} \text{ shatters } B\} = \{\phi, \{\mathbf{e}_1\}, \{\mathbf{e}_2\}\}$. Futhermore, we have $\sum_{i=0}^d \binom{|A|}{i} = 3$, where $d = 1$.
- $(=,<)$. Here we consider axis-aligned rectangles in $\mathbb{R}^2$ in this chapter, whose VC-dimension is 4. Then we construct $A = \{(0,0), (1,0), (2,0)\}$, and all the labelings except $(1,0,1)$ can be obtained. Then we have $|\mathcal{H}_A| = 7$, $|\{B \subseteq A | \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^d \binom{|A|}{i} = 8$.
- $(<,<)$. Consider the class $\mathcal{H} = \{\mathrm{sign}\langle \mathbf{w}, \mathbf{x} \rangle | \mathbf{w} \in \mathbb{R}^d\}$ where $d \geq 3$. Suppose $A = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, and A can be shattered, therefore $\mathrm{VCdim}(\mathcal{H}) \geq 3$. Then we construct A as $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = \mathbf{e}_1$, $\mathbf{x}_2 = \mathbf{e}_2$, $\mathbf{x}_3 = (1,1,0,...,0)$, then we can derive that all the labelings except $(1,1,-1)$ and $(-1,-1,1)$ are obtained, which indicates that $|\mathcal{H}_A| = 6$, $|\{B \subseteq A | \mathcal{H} \text{ shatters } B\}| = 7$, and $\sum_{i=0}^d \binom{|A|}{i} = 8$.

- $(=,=)$. Consider $d=1$, and the class $\mathcal{H}=\{1_{[x\geq t]}\,|t\in\mathbb{R}\}$, then the VC-dimension is 1. Construct a finite set $A\subseteq\mathbb{R}$, then each of the three terms in "Sauer's inequality" equals $|A|+1$.

■

6. **Exercise 6.7**

   (a) The hypothesis class $\mathcal{H}=\{1_{[x\geq s]}\,|s\in\mathbb{R}\}$ is infinite, where $\text{VCdim}(\mathcal{H})=1$.

   (b) Consider the hypothesis class $\mathcal{H}=\{1_{[x\leq 1]},1_{[x\leq 1/3]}\}$, where $\mathcal{H}$ is finite and $\text{VCdim}(\mathcal{H})=\log_2(|\mathcal{H}|)$. ■

9. **Exercise 11.1**

   Let S be an i.i.d. sample. Let h be the output of the described learning algorithm. Note that (independently of the identity of S), $L_D(h)=1/2$ (since h is a constant function). Let us calculate the estimate $L_V(h)$. Assume that the parity of S is 1. Fix some fold $\{(\mathbf{x},y)\}\subseteq S$. We distinguish between two cases:

   - The parity of $S\backslash\{\mathbf{x}\}$ is 1. It follows that $y=0$. When being trained using $S\backslash\{\mathbf{x}\}$, the algorithm outputs the constant predictor $h(\mathbf{x})=1$. Hence, the leave-one-out estimate using this fold is 1.
   - The parity of $S\backslash\{\mathbf{x}\}$ is 0. It follows that $y=1$. When being trained using $S\backslash\{\mathbf{x}\}$, the algorithm outputs the constant predictor $h(\mathbf{x})=0$. Hence, the leave-one-out estimate using this fold is 1.

   Averaging over the folds, the estimate of the error of h is 1. Consequently, the difference between the estimate and the true error is $1/2$. The case in which the parity of S is 0 is analyzed analogously. ■

10. **Exercise 11.2**

    Consider for example the case in which $\mathcal{H}_1\subseteq\mathcal{H}_2\subseteq\cdots\subseteq\mathcal{H}_k$, and $|\mathcal{H}_i|=2^i$ for every $i\in k$. Learning $\mathcal{H}_k$ in the agnostic-PAC model provides the following bound for an ERM hypothesis h:

    $$L_D(h)\leq\min_{h\in\mathcal{H}_k}L_D(h)+\sqrt{\frac{2(k+1+\log(1/\delta))}{m}}\tag{1}$$

    Alternatively, we can use model selection as we describe next. Assume that j is the minimal index which contains a hypothesis $h^*\in\operatorname{argmin}_{h\in\mathcal{H}}L_D(h)$. We first train $\mathcal{H}_i$ on the $(1-\alpha)m$ training examples using ERM rule with respect to $\mathcal{H}_i$. Denote $\hat{h}_i$ as the hypothesis returned by ERM rule. Then we apply the ERM rule with respect to the finite class $\{\hat{h}_1,\cdots,\hat{h}_k\}$ on the αm examples. Denote $\hat{h}$ as the final hypothesis returned by this approach.

    Since $\{\hat{h}_1,\cdots,\hat{h}_k\}$ is a finite class with size k, with probability of at least $1-\delta/2$, we have:

    $$L_D(\hat{h})\leq L_D(\hat{h}_j)+\sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}$$

Now we consider each of the hypotheses in H_j, since $\hat{h}_j$ is obtained using ERM rule on $(1-\alpha)m$ training data, we obtain that with probability at least $1-\delta/2$,

$$L_D(\hat{h}_j) \leq L_D(h^*) + \sqrt{\frac{2}{(1-\alpha)m} \log \frac{4|\mathcal{H}_j|}{\delta}}$$

Combining the two last inequalities with union bound, we obtain that with probability at least $1-\delta$,

$$\begin{aligned} L_D(\hat{h}) &\leq L_D(h^*) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m} \log \frac{4|\mathcal{H}_j|}{\delta}} \\ &= L_D(h^*) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m} (j + \log \frac{4}{\delta})} \end{aligned} \tag{2}$$

Comparing the two bounds (inequaliy 1 and inequality 2), we see that when the "optimal index" j is significantly smaller than k, the bound achieved using model selection is much better. Being even more concrete, if j is logarithemic in k, we achieve a logarithmic improvement. ■

7. **Exercise 7.1**

(a) Denote $n = \max_{h \in \mathcal{H}} \{|d(h)|\}$.Since each $h \in \mathcal{H}$ has a unique description, then we can derive that

$$|\mathcal{H}| \leq \sum_{i=0}^{n} 2^i = 2^{n+1} - 1,$$

which indicates that $\text{VCdim}(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor \leq n + 1 \leq 2n$.

(b) Denote $n = \max_{h \in \mathcal{H}} \{|d(h)|\}$. For $\mathbf{x}, \mathbf{y} \in \cup_{k=0}^{n} \{0,1\}^k$, we say $\mathbf{x} \sim \mathbf{y}$ if $\mathbf{x}$ is a prefix of $\mathbf{y}$ or $\mathbf{y}$ is a prefix of $\mathbf{x}$, which is a symmetric relation. Suppose d is prefix-free, then we can bound the size of $\mathcal{H}$ by the number of equivalence classes. Since there exists a one-to-one mapping from $\{0,1\}^n$ to the set of equivalence classes. Then we can derive that $|\mathcal{H}| \leq 2^n$, which concludes our proof. ■

8. **Exercise 7.2**

Suppose there exists k such that $w(h_k) > 0$ and denote $w^* = w(h_k) > 0$, then according to the nondecreasing, we have

$$\sum_{i=1}^{\infty} w(h_i) \geq \sum_{i=k}^{\infty} w(h_i) \geq \sum_{i=k}^{\infty} w^* = \infty,$$

which is contradict to the condition that $\sum_{i=1}^{\infty} w(h_i) \leq 1$. ■