CS 260: Machine Learning **Spring 2020**

Homework 2

Student Name: Andrew Choi *UID: 205348339*

- Feel free to talk to other students in the class when doing the homework. You should, however, write down your solution yourself. You also must indicate on each homework with whom you collaborated and cite any other sources you use including Internet sites.
- You will write your solution in LaTeX and submit the **pdf file through Gradescope**. You also need to submit the **zipped LaTeX files to CCLE**. We will grade your homework based on the final version of the pdf file submitted to Gradescope. We will not grade the zipped Latex files on CCLE. However, failure to submitting your LaTax files to CCLE will incur 2 points penalty out of 100 points.
- The homework (both pdf and zipped Latex source files) is due at **1:59 PM before the class**.

1. Exercise **5.1**

   The following can be proved by using Lemma B.1.

   $$\mathbb{P}[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$$

   As $L_{\mathcal{D}}(A(S)) \in [0,1]$ and $\mathbb{E}[L_{\mathcal{D}}(A(S)) = 1/4$, we can substitute these values into Lemma B.1 to complete the proof.

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

2. Exercise **6.1**

   If hypothesis class $\mathcal{H}' \subseteq \mathcal{H}$ then that means that if $\mathcal{H}'$ shatters a set $C \subset \mathcal{X}$, $\mathcal{H}$ also shatters the same set C as it includes all the hypotheses contained in $\mathcal{H}'$. Therefore, for every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$ then $\text{VCdim}(\mathcal{H}') \leq \text{VCdim}(\mathcal{H})$.

3. Exercise **6.2**

   a. Two conditions must be explored, $k \leq |\mathcal{X}|/2$ and $k > |\mathcal{X}|/2$, where the VCdim is k and $|\mathcal{X}| - k$, respectively. Therefore, the VCdim of $\mathcal{H}_{=k} = \min\{k, |\mathcal{X}| - k\}$. First, we show that a set of size $k+1$ cannot be shattered. Simply consider the case where all $k+1$ instances are positive. Then, since the hypothesis class can only assign the value of 1 to exactly k elements, $\text{VCdim}(\mathcal{H}_{=k}) \leq k$. The same can be done for a set of size $|\mathcal{X}| - k + 1$. Now, we tighten the bound. Let there be a set $S \subset \mathcal{X}$ of size $m \leq \min\{k, |\mathcal{X}| - k\}$. Let a equal the sum of the labels of set S. Then, if we pick a subset $C \subset \mathcal{X} \setminus S$ of $k - a$ elements where $h \in \mathcal{H}_{=k}$ satisfies all labels from S and produces a 1 for all values $\in C$, S is shattered, concluding the proof.

b. The VCdim of $\mathcal{H}_{at-most-k} = k$. First, we show that $\mathcal{H}_{at-most-k}$ cannot shatter a set of size $k+1$. Simply consider a set of size $k+1$ in which all instances are positive examples. Then, since the $\mathcal{H}_{at-most-k}$ can only classify at most k positive examples, $\text{VCdim}(\mathcal{H}_{at-most-k}) \leq k$. Now, if we consider a set S of all possible 2^k binary vectors of size k, we can see that $\mathcal{H}_{at-most-k}$ is capable of classifying all of them by definition, thus shattering S, which concludes the proof.

4. Exercise **6.3**

The VC-dimension of $\mathcal{H}_{n-parity} = n$. First, we can show the upper bound as we know that $|\mathcal{H}_{n-parity}| = 2^n$.

$$\text{VCdim}(\mathcal{H}_{n-parity}) \leq \log_2(|\mathcal{H}_{n-parity}|) = n$$

Next, we can show that $\mathcal{H}_{n-parity}$ can shatter a set of size n. Consider all 2^n possible binary vectors of size n. As the hypothesis class $\mathcal{H}_{n-parity}$ is capable of calculating the correct parity of any vector up to size n, we can conclude that $\text{VCdim}(\mathcal{H}_{n-parity}) = n$.

5. Exercise **6.4**

a. First, we consider the $(=, =)$ case shown below.

$$|\mathcal{H}_A| = |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| = \sum_{i=0}^{d} \binom{|A|}{i}$$

Consider the threshold hypothesis class whose VCdim is 1. If we choose any finite set of points $A \subset \mathbb{R}$, then it can be seen that Sauer's lemma equates to $|A| + 1$ for each element.

b. Next, we consider the $(<, =)$ case.

$$|\mathcal{H}_A| < |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| = \sum_{i=0}^{d} \binom{|A|}{i}$$

Consider a 2D threshold hypothesis class in which the threshold is a horizontal line on a 2D plane. This class has a VCdim of 1 for the same reason as the 1D threshold hypothesis class. Then, for $d = 2$, if we have two points horizontally aligned, then the hypothesis class can produce labels (0, 0) and (1, 1) only. This results in the inequality $2 < 3 = 3$.

c. Next, we consider the $(=, <)$ case.

$$|\mathcal{H}_A| = |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| < \sum_{i=0}^{d} \binom{|A|}{i}$$

Consider the interval hypothesis class whose VCdim is 2. Then, for $d = 3$ (3 points on a line), the interval hypothesis can produce all labels except for (1, 0, 1). This results in the inequality $7 = 7 < 8$.

d. Finally, we consider the $(<, <)$ case.

$$|\mathcal{H}_A| < |\{B \subseteq A : \mathcal{H} \text{ shatters } B\}| < \sum_{i=0}^{d} \binom{|A|}{i}$$

Could not come up with a case for this.

6. Exercise **6.7**

a. One example is the threshold hypothesis class.

$$\mathcal{H} = \{h(x) = \mathbb{1}[x \geq t] : t \in \mathbb{R}\}$$

As the domain $\mathcal{X}$ includes all real numbers in the range of 0 to 1, the number of possible hypotheses is infinite, thus, the size of the hypothesis class $\mathcal{H}$ is infinite. At the same time, the threshold function cannot shatter a set of size 2 which proves that $\text{VCdim}(\mathcal{H})$ is 1.

b. One example is a hypothesis class containing two indicator functions,

$$\mathcal{H} = \{\mathbb{1}[0 \leq x < 0.5], \mathbb{1}[0.5 \leq x \leq 1]\}$$

It can intuitively be seen that $\mathcal{H}$ shatters any set of size 1 but is unable to shatter a set of size 2. Therefore, $\text{VCdim}(\mathcal{H}) = \log_2(|\mathcal{H}|) = 1$.

7. Exercise **11.1**

First, we know that $L_{\mathcal{D}}(h) = 1/2$, as h is a constant function. Next, we must calculate $L_V(h)$ when using leave-one-out k-fold cross validation. Let S be a set of i.i.d. samples. We then fix one hold $\{(x, y)\} \subset S$. Now, we must solve for the two cases in which the parity of S is 1 or 0. We then analyze two different scenarios within each of these cases where the parity of $S \setminus \{x\}$ is 1 or 0 for a total of four possible cases. Below we analyze the two cases for when the parity of S is 1. Note that the two remaining cases in which the parity of S is 0 can be solved similarly.

a. The parity of S is 1 and the parity of $S \setminus \{x\} = 1$. This indicates that $y = 0$. The learning algorithm then outputs the constant hypothesis $h(x) = 1$ which results in $L_V(h) = 1$.

b. The parity of S is 1 and the parity of $S \setminus \{x\} = 0$. This indicates that $y = 1$. The learning algorithm then outputs the constant hypothesis $h(x) = 0$ which results in $L_V(h) = 1$.

Averaging over all k folds, we get $L_V(h) = 1$. Therefore, we conclude the following:

$$|L_V(h) - L_{\mathcal{D}}(h)| = 1/2$$

8. Exercise **11.2**

We denote $h* = argmin_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ as the predictor with the minimal true error. We denote $ERM_{\mathcal{H}}$ as the predictor returned by applying ERM to all data with size m. We denote $\hat{h}_i$ as the predictor returned by applying ERM to $\mathcal{H}_i$ to training set with size $(1-\alpha)m$. We denote $\hat{h}$ as the predictor returned by applying ERM to hypothesis class $\mathcal{H}_r = \{\hat{h}_1, \hat{h}_2, ...\hat{h}_k\}$ on αm validation data.

a. For learning $\mathcal{H}$ with ERM on m examples, we can find the error as this is agnostic PAC learnable. With probability $1-\delta$,

$$|L_{\mathcal{D}}(ERM_{\mathcal{H}}) - L_{\mathcal{D}}(h^*)| \leq \epsilon$$
$$\epsilon = \sqrt{\frac{2\log(2|\mathcal{H}|/\delta)}{m}}$$

b. For the model validation case, we can solve for the error by the following derivation. First, we bind the error of $\hat{h}$ in terms of $\hat{h}_x$ where we assume that $h^* \in \mathcal{H}_x$. Then with probability $1-\delta$,

$$\begin{aligned} L_{\mathcal{D}}(\hat{h}) &\leq L_V(\hat{h}) + \sqrt{\frac{\log(2|\mathcal{H}_r|/\delta)}{2\alpha m}} \\ &\leq L_V(\hat{h}_i) + \sqrt{\frac{\log(2|\mathcal{H}_r|/\delta)}{2\alpha m}} \ \forall i \in [k] \\ &\leq L_D(\hat{h}_i) + \sqrt{\frac{2\log(2|\mathcal{H}_r|/\delta)}{\alpha m}} \end{aligned}$$

Next, we bound the error of $\hat{h}_x$.

$$L_{\mathcal{D}}(\hat{h}_x) \leq L_{\mathcal{D}}(h^*) + \sqrt{\frac{2\log(2|\mathcal{H}_x|/\delta)}{(1-\alpha)m}}$$

If we combine these two statements we get the following error bound.

$$|L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(h^*)| \leq \sqrt{\frac{2\log(2|\mathcal{H}_x|/\delta)}{(1-\alpha)m}} + \sqrt{\frac{2\log(2|\mathcal{H}_r|/\delta)}{\alpha m}}$$

Comparing the two error bounds, we can see that ERM on the hypothesis class $\mathcal{H}$ will have lower error than the model selection method when $|\mathcal{H}|$. Likewise, if $|\mathcal{H}_x|$ and $|\mathcal{H}_r|$ are $<<$ $|\mathcal{H}|$, then model selection will be a better choice.

9. Exercise **7.1**

Each hypothesis $h \in \mathcal{H}$ has a unique description length, therefore, we must sum up all possible descriptions up to the maximum length to bound the VC dimension. Denote the maximum description length as $m = \max_{h \in \mathcal{H}} |d(h)|$. Then, we can show the

following.

$$|\mathcal{H}| \leq \sum_{i=0}^{m} 2^i \approx 2^{m+1}$$
$$\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|) \leq m + 1 \leq 2m$$

For prefix-free descriptions, $|\mathcal{H}|$ can be upper bounded by the number of equivalence classes pertaining to matching prefixes which equals 2^m. Therefore,

$$|\mathcal{H}| \leq 2^m$$
$$\text{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|) \leq m$$

10. Exercise **7.2**

As the weights are monotonically nondecreasing, for all j such that $j \geq 1$, this means that

$$\sum_{i=1}^{\infty} w(h_i) \geq \sum_{i=j}^{\infty} w(h_j)$$

Due to this property of monotonicity, it can then be proven that this hypothesis class cannot be nonuniformly learned due the violation of the weight condition.

$$\sum_{i=1}^{\infty} w(h_i) = \infty$$