MLT Homework 2

Ana Borovac
Jonas Haslbeck
Bas Haver

September 2018

Question 1

Subquestion 1.1

Monotonicity of Sample Complexity: Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot, \cdot)$. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters.

Solution

Firstly we are going to show that $m_{\mathcal{H}}$ is monotonically nonincreasing in ϵ. In order to do that, let recall the PAC learnability definition:

$$P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon \right) < \delta, \quad \text{if } m \geq m_{\mathcal{H}}(\epsilon, \delta)$$

Now we fix $\delta \in (0, 1)$ and pick $0 < \epsilon_1 < \epsilon_2 < 1$:

$$\begin{aligned} P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon_1 \right) < \delta, \quad \text{if } m \geq m_{\mathcal{H}}(\epsilon_1, \delta) \\ P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon_2 \right) < \delta, \quad \text{if } m \geq m_{\mathcal{H}}(\epsilon_2, \delta) \end{aligned}$$

Since $\epsilon_1 < \epsilon_2$, if follows:

$$\{L_{\mathcal{D}}(A(S)) > \epsilon_1\} \supset \{L_{\mathcal{D}}(A(S)) > \epsilon_2\}$$

$$P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon_1 \right) > P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon_2 \right)$$

From that we can conclude, that every $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$ also satisfies the inequality $m \geq m_{\mathcal{H}}(\epsilon_2, \delta)$, but not the other way around. So:

$$m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$$

For the other part of the proof we fix $\epsilon \in (0,1)$ and pick $0 < \delta_1 < \delta_2 < 1$:

$$P_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) > \epsilon\right) < \delta_1, \quad \text{if } m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$$
$$P_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(A(S)) > \epsilon\right) < \delta_2, \quad \text{if } m \geq m_{\mathcal{H}}(\epsilon, \delta_2)$$

Conclusion is similar to the previous conclusion. Since $\delta_1 < \delta_2$, every $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$ satisfies the inequality $m \geq m_{\mathcal{H}}(\epsilon, \delta_2)$, but not the other way around. So:

$$m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$$

Subquestion 1.2

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0,1\}$, and let $\mathcal{H}$ be the class of concentric circles in the plane, that is, $\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbb{1}_{[||x|| \leq r]}$. Prove that $\mathcal{H}$ is PAC learnable (assume realizability), and its sample complexity in bounded by

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{\log(1/\delta)}{\epsilon}$$

.

Solution

By realizability assumption we know that there exists a circle which separates samples classified as 1 and samples classified as 0. Let denote the radius of that circle with r^*.

Let suppose that our algorithm A returns h_r, where r is the smallest radius of circle enclosing all samples classified as 1.

Next, we define one more radius s. We define s for which it holds:

$$P(\{x : s \leq ||x|| \leq r^*\}) = \epsilon$$

Now, we would like to prove that, if $s < r < r^*$ there is little chance of error. First, we know that $r < r^*$, otherwise perfect circle would not classify all training samples correctly. Second, the errors that our algorithm makes are in the space between circles with radii r and r^*. From the definition of s and $s \geq r$ we know that:

$$P(\{x : r \leq ||x|| \leq r^*\}) < \epsilon$$

But, what is the probability that we make a bigger error. In other words, what is the probability of $r < s$?

$$P(error \geq \epsilon) = (1-\epsilon)^m \leq e^{-\epsilon m}$$

We would like that this probability is small, so:

$$\begin{aligned}
e^{-\epsilon m} &< \delta \\
-\epsilon m &< \log \delta \\
\epsilon m &> \log \frac{1}{\delta} \\
m &> \frac{\log 1/\delta}{\epsilon}
\end{aligned}$$

Question 2

Let $\mathcal{H}$ be a hypothesis class of binary classifiers. Show that if $\mathcal{H}$ is agnostic PAC learnable, then $\mathcal{H}$ is PAC learnable as well.

Solution

Because $\mathcal{H}$ is agnostic PAC learnable, we know that $\forall \mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and $\forall \epsilon, \delta \in (0,1)$, it holds:

$$P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \min_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) + \epsilon) \right) < \delta$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

We would like to prove that for $\forall \mathcal{D}$ over $\mathcal{X}$ and $\forall f \in \mathcal{H}$ and $\forall \epsilon, \delta \in (0,1)$, it holds:

$$P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon \right) < \delta$$

whenever $m \geq m_{\mathcal{H}}(\epsilon, \delta)$. When proving PAC learnability, we assume that there exists a perfect labeling function. From that, we can conclude:

$$\min_{h \in \mathcal{H}} (L_{\mathcal{D}}(h)) = 0$$

It follows:

$$\begin{aligned}
P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \min_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) + \epsilon) \right) &< \delta \\
P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \min_{h \in \mathcal{H}} (L_{\mathcal{D}}(h)) + \epsilon \right) &< \delta \\
P_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) > \epsilon \right) &< \delta
\end{aligned}$$

Question 3

The Bayes optimal predictor: Show that for every probability distribution $\mathcal{D}$, the Bayes optimal predictor $f_{\mathcal{D}}$ is optimal, in the sense that for every classifier g from $\mathcal{X}$ to $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Solution

The solution was written with the help of [**?**].

Let's recall the definition of Bayes classifier:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } P(y=1|x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

and the definition of true error of a prediction rule h:

$$L_{\mathcal{D}}(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y) = \mathcal{D}(\{(x,y) : h(x) \neq y\})$$

We would like to prove that $\forall g : \mathcal{X} \to \{0,1\} : \; L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$ or written differently:

$$L_{\mathcal{D}}(g) - L_{\mathcal{D}}(f_{\mathcal{D}}) \geq 0$$

Now, we are going to transform the definition of true error:

$$\begin{aligned} P(error) &= P(h(x) \neq y|x) \\ &= 1 - P(h(x) = y|x) \\ &= 1 - [P(h(x) = 1 \wedge y = 1|x) + P(h(x) = 0 \wedge y = 0|x)] \\ &= 1 - [P(h(x) = 1)P(y = 1|x) + P(h(x) = 0)P(y = 0|x)] \\ &= 1 - [P(h(x) = 1)P(y = 1|x) + P(h(x) = 0)(1 - P(y = 1|x))] \end{aligned}$$

Furthermore:

$$\begin{aligned} L_{\mathcal{D}}(g) - L_{\mathcal{D}}(f_{\mathcal{D}}) &= (1 - [P(g(x) = 1)P(y = 1|x) + P(g(x) = 0)(1 - P(y = 1|x))]) \\ &\quad - (1 - [P(f_{\mathcal{D}}(x) = 1)P(y = 1|x) + P(f_{\mathcal{D}}(x) = 0)(1 - P(y = 1|x))]) \\ &= -[P(g(x) = 1)P(y = 1|x) + P(g(x) = 0)(1 - P(y = 1|x))] \\ &\quad + [P(f_{\mathcal{D}}(x) = 1)P(y = 1|x) + P(f_{\mathcal{D}}(x) = 0)(1 - P(y = 1|x))] \\ &= P(y = 1|x)(-P(g(x) = 1) + P(f_{\mathcal{D}}(x) = 1)) \\ &\quad + (1 - P(y = 1|x))(-P(g(x) = 0) + P(f_{\mathcal{D}}(x) = 0)) \\ &= P(y = 1|x)(-P(g(x) = 1) + P(f_{\mathcal{D}}(x) = 1)) \\ &\quad + (1 - P(y = 1|x))(-1 + P(g(x) = 1) + 1 - P(f_{\mathcal{D}}(x) = 1)) \\ &= P(y = 1|x)(P(f_{\mathcal{D}}(x) = 1) - P(g(x) = 1)) \\ &\quad + (1 - P(y = 1|x))(P(g(x) = 1) - P(f_{\mathcal{D}}(x) = 1)) \\ &= (P(f_{\mathcal{D}}(x) = 1) - P(g(x) = 1))(P(y = 1|x) - (1 - P(y = 1|x))) \\ &= (P(f_{\mathcal{D}}(x) = 1) - P(g(x) = 1))(2P(y = 1|x) - 1) \end{aligned}$$

Next, we are going to divide our problem into two parts:

- $P(y = 1|x) \geq 1/2$:

$$L_{\mathcal{D}}(g) - L_{\mathcal{D}}(f_{\mathcal{D}}) = \underbrace{(1 - \underbrace{P(g(x) = 1)}_{\in[0,1]})}_{\geq 0}\underbrace{(2\underbrace{P(y = 1|x)}_{\geq 1/2} - 1)}_{\geq 0} \geq 0$$

- $P(y=1|x) < 1/2$:

$$L_{\mathcal{D}}(g) - L_{\mathcal{D}}(f_{\mathcal{D}}) = (\underbrace{-P(g(x)=1)}_{\leq 0})\underbrace{(2\underbrace{P(y=1|x)}_{<1/2}-1)}_{<0} \geq 0$$

Now we can conclude:

$$L_{\mathcal{D}}(g) - L_{\mathcal{D}}(f_{\mathcal{D}}) \geq 0$$
$$L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g); \quad \forall g$$

Question 4

Subquestion 4.1

Prove that the following two statements are equivalent (for any learning algorithm A, any probability distribution $\mathcal{D}$, and any loss function whose range is $[0,1]$*):*

1. *For every $\epsilon, \delta > 0$, there exists $m(\epsilon,\delta)$ such that $\forall m \geq m(\epsilon,\delta)$*

$$P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > \epsilon) < \delta$$

2.

$$\lim_{m\to\infty} \mathbb{E}_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S))) = 0$$

Solution

Subquestion 4.2

Bounded loss functions: Prove that if the range of the loss function is $[a,b]$ the in sample complexity satisfies

$$m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2,\delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil$$

Solution

First, let's recall Hoeffding's Inequality:

$$\forall \epsilon > 0: \; P\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right) \leq 2e^{-2m\epsilon^2/(b-a)^2}$$

where $\theta_1, \ldots, \theta_m$ are i.i.d. random variables and $\mathbb{E}[\theta_i] = \mu$ $(i = 1, \ldots, m)$ and $P(a \leq \theta_i \leq b) = 1$ $(i = 1, \ldots, m)$. On lectures we proved the inequality for example, where $a = 0$ and $b = 1$. We can do the same proof here:

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} 2e^{-2m\epsilon^2/(b-a)^2} \\ &= 2|\mathcal{H}|e^{-2m\epsilon^2/(b-a)^2}\end{aligned}$$

Finally, if we choose:

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)(b-a)^2}{2\epsilon^2}$$

then:

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta$$

Furthermore:

$$\begin{aligned}m_{\mathcal{H}}^{UC}(\epsilon, \delta) &\leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)(b-a)^2}{2\epsilon^2} \right\rceil \\ m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) &\leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil\end{aligned}$$

Question 5

Prove that when the expected losses $L_{\mathcal{D}}(h)$ are bounded, we have

$$L_{\mathcal{D}}(h_S) - \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \leq 2 \sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)|$$

Solution

References

[1] R. Nowak. Lecture 2: Introduction to Classification and Regression. `nowak.ece.wisc.edu/SLT09/lecture2.pdf`, May 2009. Online; accessed 23 September 2018.