Subject:

Date: No:

A Formal Learning Model - Exercises

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. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \leq \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

$D \rightsquigarrow$ unknow distribution over $\mathcal{X}$

$f \in \mathcal{H} \rightsquigarrow$ target hypothesis

$A \rightsquigarrow$ learning algorithm with sample complexity of $m_{\mathcal{H}}(\cdot,\cdot)$

Monotonicity of parameter ϵ:

fix δ, we have to show $\epsilon_1 \leq \epsilon_2 \Rightarrow m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$

Given an i.i.d. training sequence of size $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$

we have $L_{D,f}(h) \leq \epsilon_1 \leq \epsilon_2 \xrightarrow[\text{by minimality of } m_{\mathcal{H}}(\epsilon_2,\delta)]{} m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$

Monotonicity of parameter δ:

fix ϵ, we have to show $\delta_1 \leq \delta_2 \Rightarrow m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$

Given an i.i.d. training sequence of size $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$

we have $L_{D,f}(h) \leq$

PONIX notebook

Subject:

Date: No:

2. Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{singleton} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where for each $z \in \mathcal{X}$, h_z is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in \mathcal{X}$, $h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{singleton}$ in the realizable setup.

Algorithm
- → if a positive instance like x_+ appears in S then return h_{x_+}
- → if S doesn't have any positive instances then return h^-

2. Show that $\mathcal{H}_{singleton}$ is PAC learnable. Provide an upper bound on the sample complexity.

$\epsilon \in (0,1)$, D is a distribution over $\mathcal{X}$

i) if h^- is the true hypothesis ✓

ii) suppose there exists a x_+ instance.

if x_+ apears in $S \rightsquigarrow$ ✓

if $D(\{x_+\}) \leq \epsilon \rightsquigarrow$ the generalization error is at most ϵ

if x doesn't appear in S and $D(\{x_+\}) > \epsilon$ then

$$D(\, D(\{x_+\}) > \epsilon \mid x \text{ doesn't appear in } S) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

$\Rightarrow \mathcal{H}_{singleton}$ is PAC learnable and

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil .$$

Subject:

Date: No: 2

3. 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 $h_r(x) = 1_{[\|x\| \leq r]}$. Prove that $\mathcal{H}$ is PAC learnable (assume realizability) and its sample complexity is bounded by $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log 1/\delta}{\epsilon} \rceil$

Suppose $\hat{h}$ be a hypothesis corresponding to the tightest circle that contains all positive instances. Let $\hat{h}$ has the radius $\hat{r}$. Let h^* be a circle with generalization risk of 0 and its radius is r^*.

Let r be a scalar such that $D_1\{x : r \leq \|x\| \leq r^+\} = \varepsilon$.

The probability that $L_D(h_S) \geq \varepsilon$ is bounded by the probability that no points belong to $\{x : r \leq \|x\| \leq r^+\}$ and it is bounded by $(1-\varepsilon)^m$ which is less that $e^{-\varepsilon m}$

$$\Rightarrow m_{\mathcal{H}}(\varepsilon, \delta) \leq \lceil \frac{\log 1/\delta}{\varepsilon} \rceil.$$

4. In this question, we study the hypothesis class of Boolean conjuctions defined as follows. The instance space is $\mathcal{X} = \{0,1\}^d$ and the label set is $\mathcal{Y} = \{0,1\}$. A literal over the variables $x_1, \ldots, x_d$ is a simple Boolean function that takes the form $f(x) = x_i$, for some $i \in [d]$, or $f(x) = 1 - x_i$ for some $i \in [d]$. We use the notation $\bar{x}_i$ as a shorthand for $1 - x_i$. A conjuction is any product of literals. In Boolean logic, the product is denoted using the $\wedge$ sign. For example, the function $h(x) = x_1 \cdot (1 - x_2)$ is written as $x_1 \wedge \bar{x}_2$.

We consider the hypothesis class of all conjuctions of literals over the d variables. The empty conjuction is interpreted as the all-positive hypothesis (namely, the function that returns $h(x) = 1$ for

PONIX notebook

Subject:

Date: No:

all x). The conjuction $x_1 \wedge \bar{x}_2$ (and similarly any conjuction involving a literal and its negation) is allowed and interpreted as the all-negative hypothesis (namely, the conjuction that return $h(x)=0$ for all x). We assume realizability: Namely, we assume that there exists a Boolean conjuction that generates the labels. Thus, each example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ consists of an assignment to the d Boolean variables $x_1, \ldots, x_d$, and its truth value (0 for false and 1 for true).

For instance, let $d=3$ and suppose that the true conjuction is $x_1 \wedge \bar{x}_2$. Then, the training set S might contain the following instances:

$((1,1,1),0), ((1,0,1),1), ((0,1,0),0), ((1,0,0),1)$.

Prove that the hypothesis class of all conjuctions over d variable is PAC learnable and bound its sample complexity. Propose an algorithm that implements the ERM rule, whose runtime is polynomial in $d \cdot m$.

First we compute the size of $\mathcal{H}$

x_i
- x_i appears in the corresponding conjuction
- $\bar{x}_i$ appears in the conjuction
- not x_i nor $\bar{x}_i$ appear in the ~

$$\Rightarrow |\mathcal{H}| = 3^d + \underbrace{1}_{\text{all-negative hypothesis}}$$

$\mathcal{H}$ is PAC learnable and

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{d \log 3 + \log(1/\delta)}{\varepsilon} \right\rceil$$

Subject:

Date: No:

$h_0 = x_1 \cap \bar{x}_1 \cap \ldots \cap x_d \cap \bar{x}_d$ ⇝ h_0 is always minus hypothesis

$((a_1, y_1), \ldots, (a_m, y_m))$ ⇝ i.i.d. training sequence of size m

For :

negative examples ⇝ our algorithm always neglect them

positive examples if $a_i = 1$ we remove $\bar{x}_i$ from h

if $a_i = 0$ we remove x_i from h

⇝ h_i labels positively all the positive examples among $a_1, a_2, \ldots, a_i$

and

the set of literals in h_i contains the set of literals in the target hypothesis

⇝ $\Rightarrow$ h_i classifies correctly the negative elements among $a_1, a_2, \ldots, a_m$

$\Rightarrow$ h is ERM

since we have m examples and for each of them it takes linear time d to process it the algorithm running time is $O(m.d)$

Subject:

Date: No:

5. Let $\mathcal{X}$ be a domain and let $D_1, D_2, \ldots, D_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in \mathcal{H}$. Suppose we are getting a sample S of m examples, such that the instances are independent but are not identically distributed; the ith instance is sampled from D_i and then y_i is set to be $f(x_i)$. Let $\bar{D}_m$ denote the average, that is, $\bar{D}_m = (D_1 + D_2 + \cdots + D_m)/m$.

Fix an accuracy parameter $\varepsilon \in (0,1)$. Show that

$P[\exists h \in \mathcal{H}$ s.t. $L_{\bar{D}_m, f}(h) > \varepsilon$ and $L_{S,f}(h) = 0] \leq |\mathcal{H}| e^{-\varepsilon m}$

Fix $h \in \mathcal{H}$ s.t. $L_{\bar{D}_m, f}(h) > \varepsilon$

We have

$$\frac{P_{x \sim D_1}(h(x) = f(x)) + P_{x \sim D_2}(h(x) = f(x)) + \cdots + P_{x \sim D_m}(h(x) = f(x))}{m}$$

$$\leq 1 - \varepsilon$$

$$\underset{\substack{\text{Sample} \\ \text{Sequence}}}{P} \sim D_1, \ldots, D_m \, (L_S(h) = 0) = \prod_{i=1}^{m} P_{x \sim D_i}(h(x) = f(x))$$

$$= \left(\left(\prod_{i=1}^{m} P_{x \sim D_i}(h(x) = f(x)) \right)^{\frac{1}{m}} \right)^m$$

geometric-arithmetic mean inequality

$$\leq \left(\frac{P_{x \sim D_1}(h(x) = f(x)) + \cdots + P_{x \sim D_m}(h(x) = f(x))}{m} \right)^m$$

$$\leq (1-\varepsilon)^m \leq e^{-\varepsilon m}$$

6. 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. Furthermore, if A is a successful agnostic PAC learner

for $\mathcal{H}$, then A is also a successful PAC learner for $\mathcal{H}$.

$\mathcal{H}$ is agnostic PAC learnable

$A \rightsquigarrow$ a learning algorithm that learns $\mathcal{H}$

$D \rightsquigarrow$ unknown distribution over $\mathcal{X}$

$f \rightsquigarrow$ target function

Assume D is a joint distribution over $\mathcal{X} \times \{0,1\}$

the conditional probability of y given x is determined deterministically by f

$\inf_{h \in \mathcal{H}} L_D(h) = 0 \rightsquigarrow$ because of realizability

Let A consists of a training set S with m i.i.d. instances which are labeled by f $\Rightarrow$ with probability at least $1-\delta$, A returns an h with

$$L_D(h) \leq \inf_{h' \in \mathcal{H}} L_D(h') + \varepsilon$$
$$= 0 + \varepsilon = \varepsilon.$$

7. The Bayes Optimal predictor: Show that for every probability distribution D, the Bayes optimal predictor f_D is optimal, in the sense that for every classifier g from $\mathcal{X}$ to $\{0,1\}$,

$L_D(f_D) \leq L_D(g)$.

$$P(f_D(x) \neq y \mid X = x) = \mathbb{1}_{P(y=1|X=x) \geq \frac{1}{2}} \cdot P(y=0 \mid X=x) +$$
$$\mathbb{1}_{P(y=1|X=x) < \frac{1}{2}} \cdot P(y=1 \mid X=x)$$
$$= \mathbb{1}_{P(y=1|X=x)}(1 - P(y=1 \mid X=x)) +$$
$$\mathbb{1}_{P(y=1|X=x)}(P(y=1 \mid X=x))$$
$$= \min\{P(y=1 \mid X=x), P(y=0 \mid X=x)\}$$

Subject:

Date: No:

$g : \mathcal{X} \longrightarrow \{0,1\}$

$$P(g(x) \neq y \mid X=x) = P(g(x)=0 \mid X=x) \cdot P(y=1 \mid X=x)$$
$$+ P(g(x)=1 \mid X=x) \cdot P(y=0 \mid X=x)$$
$$\geq P(g(x)=0 \mid X=x) \cdot \min\{P(y=1 \mid X=x), P(y=0 \mid X=x)\}$$
$$+ P(g(x)=1 \mid X=x) \cdot \min\{P(y=1 \mid X=x), P(y=0 \mid X=x)\}$$
$$= \min\{ P(Y=1 \mid X=x), P(Y=0 \mid X=x)\}$$

By the law of total expectation:

$$L_D(f_D) = E_{(x,y)\sim D}[1_{f_D(x) \neq y}]$$

$$= E_{x\sim D_x}[E_{y\sim D_{y|x}}(1_{f_D(x)\neq y} \mid X=x)]$$

$$= E_{x\sim D_x}(P(y=1 \mid X=x))$$

$$\leq E_{x\sim D_x}[E_{y\sim D_{y|x}}(1_{g(x)\neq y} \mid X=x)]$$

$$= L_D(g).$$

ect:

Date: No:

9. Consider a variant of the PAC model in which there are two example oracles: one that generates positive examples and one that generates negative examples, both according to the underlying distribution D on $\mathcal{X}$. Formally, given a target function $f : \mathcal{X} \to \{0,1\}$ let D^+ be the distribution over $\mathcal{X}^+ = \{x \in \mathcal{X} : f(x) = 1\}$ defined by $D^+(A) = D(A) / D(\mathcal{X}^+)$, for every $A \subset \mathcal{X}^+$. Similarly, D^- is the distribution over $\mathcal{X}^-$ induced by D.

The definition of PAC learnability in the two-oracle model is the same as the standard definition of PAC learnability expect that here the learner has access to $m_{\mathcal{H}}^+(\varepsilon, \delta)$ i.i.d. examples from D^+ and $m^-(\varepsilon, \delta)$ i.i.d. examples from D^-. The learner's goal is to output h s.t. with probability at least $1-\delta$ (over the choice of the two training sets, and possibly over the nondeterministic decisions made by the learning algorithm), both $L_{(D^+, f)}(h) \leq \varepsilon$ and $L_{(D^-, f)}(h) \leq \varepsilon$.

1. (*) Show that if $\mathcal{H}$ is PAC learnable (in the standard one-oracle model), then $\mathcal{H}$ is PAC learnable in the two-oracle model.

2. (**) Define h^+ to be the always-plus hypothesis and h^- to be always-minus hypothesis. Assume that $h^+, h^- \in \mathcal{H}$. Show that if $\mathcal{H}$ is PAC learnable in the-oracle model, then $\mathcal{H}$ is PAC learnable in the standard one-oracle model.

$A \rightsquigarrow$ an algorithm that learns $\mathcal{H}$

$\mathcal{H} \longrightarrow$ PAC learnable in one-oracle model

we want to prove that $\mathcal{H}$ is PAC learnable in two-oracle model

$D \rightsquigarrow$ distribution over $\mathcal{X} \times \{0,1\}$

drawing points from the negative and positive oracles with

PONIX notebook

equal probability $\equiv$ drawing i.i.d. examples from a distribution D' with equal probability for positive and negative examples

For every $A \subseteq \mathcal{X}$ we have

$$D'(A) = \frac{1}{2} D^{+}(A) + \frac{1}{2} D^{-}(A).$$

$$\Rightarrow D'(\{x : f(x) = 1\}) = D'(\{x : f(x) = 0\}) = 0.5$$

if we let the algorithm have access to a training set that is drawn i.i.d. from D' with size $m_{H}(\varepsilon/2, \delta)$ then with probability at least $1-\delta$ the algorithm return h with $L_{(D', f)}(h) \leq \frac{\varepsilon}{2}$

$$L_{(D', f)}(h) = P_{x \sim D'}(f(x) \neq h(x))$$

$$= P_{x \sim D'}[f(x) = 1, h(x) = 0] + P_{x \sim D'}[f(x) = 0, h(x) = 1]$$

$$= P_{x \sim D'}(f(x) = 1) \cdot P_{x \sim D'}(h(x) = 0 \mid f(x) = 1) + P_{x \sim D'}(f(x) = 0) \cdot P_{x \sim D'}(h(x) = 1 \mid f(x) = 0)$$

$$= P_{x \sim D'}(f(x) = 1) \cdot P_{x \sim D}(h(x) = 0 \mid f(x) = 1) + P_{x \sim D'}(f(x) = 0) \cdot P_{x \sim D}(h(x) = 1 \mid f(x) = 0)$$

$$= \frac{1}{2} L_{(D^{+}, f)}(h) + \frac{1}{2} L_{(D^{-}, f)}(h)$$

it implies that with probability at least $1-\delta$

$$L_{(D^{+}, f)}(h) \leq \varepsilon$$

$$L_{(D^{-}, f)}(h) \leq \varepsilon$$