Machine Learning Exercises (chapter 3)

2. Part 1: We define S as a training set of $X \times \{0,1\}$. If the training set contains a posetive instance (x^+), the algorithm returns h_{x^+}, otherwise, it returns h^- which is an ERM (we assume the fact that there can be at most a point with lable 1) $\longrightarrow L_S(h_S) = 0$

2. Part 2: (PAC learnability) - Assume that $x^+ \notin S$, if $D(\{x^+\}) \leq \varepsilon$ then $L_{(D,f)}(h) \leq \varepsilon$ for all h in H. Suppose that $D(\{x^+\}) > \varepsilon$, then: $\forall x' \in X$ st. $x' \neq x \rightarrow D(\{x'\}) \leq 1-\varepsilon$. then we have:

$$\{S|_x : L_{(D,f)}(h_s) > \varepsilon\} = \{S|_X : x^+ \notin S|_X \cdot D(\{x^+\}) > \varepsilon\} =$$

$$= \{S|_X : \forall x' \in S|_X \; D(\{x'\}) \leq 1-\varepsilon\}$$

$$\rightarrow D^m(\{S|_X : L_{(D,f)}(h_S) > \varepsilon\}) = D^m(\{S|_X : \forall x' \in S|_X \; D(\{x'\}) \leq 1-\varepsilon\}) \leq$$

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

Let $\delta \in (0,1)$ st. $e^{-\varepsilon m} \leq \delta \Longrightarrow m \geq \frac{\log(\frac{1}{\delta})}{\varepsilon}$

Hence $\mathcal{H}$ is PAC leanable with $m_H \leq \frac{\log(\frac{1}{\delta})}{\varepsilon}$.

3. Assume that ERM algorithm A takes S as training set and it returns the tightest circle that contains all posetive instances. Define h_S as output function and its radius by r_S, we having Realizability assumption (existing a $h^* \in H$ that its radius is r^*)

Now Let $\varepsilon, \delta \in (0,1)$, we suppose that there is a r which $r \leqslant r^*$

then: $D_X(\{x : r \leqslant \|x\| \leqslant r^*\}) = \varepsilon$ / $E = \{x \in \mathbb{R}^2 ; r \leqslant \|x\| \leqslant r^*\}$

we suppose E then we have:

$$P(L_D(h_s) > \varepsilon) = P(x_i \in S \text{ s.t. } x \notin E) = \prod_{i=0}^{m} (1 - P(x_i \in E))^m =$$

$$= (1-\varepsilon)^m \leqslant e^{m\varepsilon} \leqslant \delta$$

Now Let $\delta \in (0,1)$, the we have:

$e^{m\varepsilon} \leqslant \delta$

or

$m \geqslant \dfrac{\log(\frac{1}{\delta})}{\varepsilon}$

Hence H is PAC-leanable and $m_H(\varepsilon, \delta) \leqslant \dfrac{\log(\frac{1}{\delta})}{\varepsilon}$

5- H_B : Set of bad hypotheses /

we suppose the $h \in H_B$ s.t. $L_{(\bar{D}_m, f)}(h) \geqslant \varepsilon$ then we have:

$$\mathbb{P}[L_{(S,f)}(h)=0] = \prod_{i=1}^{m} \mathbb{P}_{X\sim D_i}[f(x)=h(x)] =$$

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

(by the geometric-arithmetic)

$$=\left(\mathbb{P}_{x\sim D_{\bar{m}}}[f(x)=h(x)]\right)^m \leqslant (1-\varepsilon)^m \leqslant e^{-\varepsilon m}$$

Therefore we have:

$$\mathbb{P}[\exists h \in H: L_{(\bar{D}_m, f)}(h) > \varepsilon ,\ L_{(S,f)}(h)=0] \leqslant |H_B| e^{-\varepsilon m} \leqslant |H| e^{-\varepsilon m}$$

6. We define $h_S := A(S)$, D: distribution over $X \times \{0,1\}$ then we have: $D^m(\{S|_X : L_D(h_S) > \min_{h \in H} L_D(h) + \varepsilon\}) < \delta$

Suppose $f \in H$: f is a true labeling function.

(determine probability of x over y.)

If $m > m_H(\varepsilon, \delta)$, $A(S)$ will perform and output h. then:

$$D_X^m(\{S : L_{(D_X, f)}(h) > \varepsilon\}) < \delta$$

H is PAC-learnable

with realizability: $\min L_D(h) = 0$, hence A is a PAC learner of H

7. Suppose $x \in \mathbf{X}$ and $\mathbf{P}$: conditional probability of a positive lable given x, then we have:

$$\mathbb{P}[f_D(X) \neq Y \mid X=x] = \mathbb{P}[Y=0 \mid X=x]\,\mathbb{1}_{[P \geq \frac{1}{2}]} + \mathbb{P}[Y=1 \mid X=x]\cdot \mathbb{1}_{[P<\frac{1}{2}]}$$

$$= (1-P)\,\mathbb{1}_{[P \geq \frac{1}{2}]} + P \cdot \mathbb{1}_{[P<\frac{1}{2}]}$$

$$= \min\{P, 1-P\}$$

suppose that g is classifier $g: X \to \{0,1\}$:

$$\mathbb{P}[g(X) \neq Y \mid X=x] = \mathbb{P}[g(x)=0 \mid \mathbf{X}=x]\,\mathbb{P}[Y=1 \mid X=x] + \mathbb{P}[g(X)=1 \mid \mathbf{X}=x]$$
$$+ \mathbb{P}[Y=0 \mid X=x] =$$

$$= \mathbb{P}[X=x]\,p + \mathbb{P}[g(X)=1 \mid \mathbf{X}=x](1-p) \geq$$

$$\geq \mathbb{P}[g(X)=0 \mid X=x]\min\{p, 1-p\} + \mathbb{P}[g(X)=1 \mid X=x]\min\{p, 1-p\} =$$

$$= \min\{p, 1-p\}$$

Therefore:

$$L_D(g) = E_{(X,Y)\sim D}[\mathbb{1}_{[g(X) \neq Y]}] = E_{X\sim D_x}\left[E_{Y\sim D_{Y|X}}[\mathbb{1}_{[g(X) \neq Y]} \mid X=x]\right] \geq$$

$$\geq E_{x\sim D_x}[E[\mathbb{1}_{[f_D(x) \neq Y]} \mid X=x]] = E[\mathbb{1}_{[f_D(X) \neq Y]}] = L_D(f_D)$$

Hence:

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

Machine Learning Exercises (Chapter 5)

5.2 _ a) pros and cons: (Algorithm A)

Pros: It has small ε_{est}, because of less complexity (less prone to overfiting)

This Algorithm can easily interpret models in a plot (dim: 2d)

Cons: Inductive bias might be too large (high ε_{app}) and we can't use A, P, I in our model

(Algorithm B)

Pros: B has smaller ~~than~~ inductive bias, reducing risk of underfitting.

It has small ε_{app}.

Cons: It has larger ε_{est}. because of that our complex model may lead to overfitting

.2_b) Increasing size of s (traing data set) leads to $L_S(h_S)$ is a better estimate of $L_D(h_S)$. It means lower ε_{est}, becase of that B is better than A.

ε_{app} can be reduced by choosing Algorithm B (Complexity)

ε_{est} can be decreased with the size of S.