Assignment 1

(Topics in Computing)

PRIYANSH AGRAHARI

CS20B1014

1. We are given that $\mathcal{H}$ is a hypothesis class for binary classification, and it is PAC learnable with sample complexity $m_{\mathcal{H}}$. We have to show that:

(i) For any fixed δ, and $0 < \epsilon_1 \leq \epsilon_2 < 1$, we must have $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$:

→ Since by definition we have that for a distribution D over X, ~~& target~~ hypothesis f, and algorithm A that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}$,

For ϵ_1, $L_{(D,f)}(h) \leq \epsilon_1$ and $Pr\{L_{(D,f)}(h) \leq \epsilon_1\} \geq 1-\delta$,

when sample size $m \geq m_{\mathcal{H}}(\epsilon_1, \delta) = m_1$

For ϵ_2, $L_{(D,f)}(h) \leq \epsilon_2$ and $Pr\{L_{(D,f)}(h) \leq \epsilon_2\} \geq 1-\delta$,

when sample size $m \geq m_{\mathcal{H}}(\epsilon_2, \delta) = m_2$

$\Rightarrow L_{(D,f)}(h) \leq \epsilon_1 \leq \epsilon_2 \Rightarrow m_2 \leq m_1$

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

(ii) For any fixed ϵ, and $0 < \delta_1 \leq \delta_2 < 1$, we must have $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$:

→ Following the definitions made in (i),

For δ_1, $Pr\{L_{(D,f)}(h) \leq \epsilon\} \geq 1-\delta_1$, when $m = m_1 \geq m_{\mathcal{H}}(\epsilon, \delta_1)$

For δ_2, $Pr\{L_{(D,f)}(h) \leq \epsilon\} \geq 1-\delta_2$, when $m = m_2 \geq m_{\mathcal{H}}(\epsilon, \delta_2)$

since $\delta_1 \leq \delta_2 \Rightarrow 1-\delta_2 \leq 1-\delta_2$.

$\Rightarrow Pr\{L_{(D,f)}(h) \leq \epsilon\} \geq 1-\delta_2 \geq 1-\delta_1 \Rightarrow m_1 \geq m_2$

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

2. We have an interval classifier $h_{[a,b]}$ given by

$$h_{[a,b]} = \begin{cases} 1 & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } a, b \in \mathbb{R}$$

And the Hypothesis class $\mathcal{H} = \{h_{[a,b]} \mid a, b \in \mathbb{R}\}$

Under the realizability assumption, we have to show the following:

(i) Consider the algorithm A that, when given a sample $S = \{x_1, \ldots, x_n\}$ outputs the smallest (tightest) interval that encloses all points in S that have label 1. Show that A minimizes empirical risk:

→ Considering A, the error of the prediction rule $h_{[a,b]}$ will comprise of the following:

(a) $h(x_i) = 1$ but $y_i \neq 1$: this occurs when $x_i \in [a,b]$ and $y_i = 0$

(b) $h(x_i) = 0$ but $y_i \neq 0$: this occurs when $x_i \notin [a,b]$ and $y_i = 1$

But since A only chooses $[a,b]$ such that all $y_i = 1$ are enclosed in it, case (b) cannot occur. Hence the empirical risk of h over S translates to

$$L_S(h) = \frac{|\{x_i : h(x_i) \neq 0 \mid y_i = 0 \;\&\; x_i \in [a,b]\}| + 0}{n}$$

By the realizability assumption, we assume that there exists $h^* \in \mathcal{H}$ such that $L_{(D,f)}(h^*) = 0$, which implies that when S is sampled over D and labelled by f, we have $L_S(h^*) = 0$. This represents the case when all points labelled 1 are adjacent, in which case $[a,b]$ will tightly contain only 1- labelled points, giving $L_S(h) = 0$.

Hence the empirical risk is minimized by A.

(ii) Show that $\mathcal{H}$ is PAC learnable via algorithm A and find the sample complexity:

We must find a polynomial bound on n such that $h_{[a,b]}$ has an error of at most ϵ, with a probability at least $1-\delta$, to show that $\mathcal{H}$ is PAC learnable. Knowing algorithm A, the only erroneous labels made by $h_{[a,b]}$ will be when sample points within $[a,b]$ have label 0.

Hence, $L_{(D,f)}(h)$ represents the probability of having a 0-labelled data point in $[a,b]$.

$$\therefore \quad L_{(D,f)}(h) \leq \epsilon$$

Then the probability of not finding any points labelled 0, or finding that all points are labelled 1 is $\overset{\text{at most}}{\wedge}(1-\epsilon)$. For n sample points, this becomes $\overset{\text{at most}}{\wedge}(1-\epsilon)^n$.

$$\therefore \quad (1-\epsilon)^n \leq \delta$$

Using the approximation $(1-\epsilon) \leq e^{-\epsilon}$, we get

$$e^{-\epsilon n} \leq \delta \quad \text{or} \quad n \leq \frac{\log(1/\delta)}{\epsilon}$$

$\therefore$ $\mathcal{H}$ is PAC learnable via algorithm A, and the sample complexity, $m_{\mathcal{H}} \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$

4. For any joint distribution D over $X \times \{0,1\}$, the Bayes optimal predictor is defined as:

$$f_D \triangleq \begin{cases} 1 & \text{if } Pr[y=1|x] \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

We have to show that this is optimal, i.e., show that for any $g: X \to \{0,1\}$, it must be the case that $L_D(f_D) \leq L_D(g)$:

For some $x \in X$, let α_x be the probability of a label 1 given x, by Bayes predictor, that is, $\alpha_x = Pr[f_D(x) = 1 \mid X = x]$

Considering $Pr[f_D(x) \neq y \mid X = x] = 1_{[at\ \alpha_x \geq 1/2]} \cdot Pr[Y=0|X=x] + 1_{[at\ \alpha_x < 1/2]} \cdot Pr[Y=1|X=x]$

$= 1_{(at\ \alpha_x \geq 1/2)} \cdot (1-\alpha_x) + 1_{(at\ \alpha_x < 1/2)} \cdot \alpha_x$ (using disjoint union and independent events)

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

For $g: X \to \{0,1\}$, considering

$Pr[g(X) \neq y \mid X=x] = Pr[g(X)=0|X=x] \cdot Pr[y=1|X=x]$

$+ Pr[g(x)=1|X=x] \cdot Pr[y=0|X=x]$

$= Pr[g(X)=0|X=x] \cdot \alpha_x + Pr[g(X)=1|X=x] \cdot (1-\alpha_x)$

$\geq Pr[g(X)=0|X=x] \cdot \min\{\alpha_x, 1-\alpha_x\} + Pr[g(X)=1|X=x] \cdot \min\{\alpha_x, 1-\alpha_x\}$

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

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