

Foundations of Machine Learning — Homework Assignment 1

Anirudhan J Rajagopalan
N18824115
ajr619

October 11, 2015

A. PAC Learning

1

Algorithm A: Given a sample S , the algorithm returns the tightest interval I' containing all points labeled with 1 in sample S .

Error intervals: We define error regions as the intervals that lie between $[a, b]$ but outside I' .

Proof of PAC learnability :

Let I be the target concept — $[a, b]$.

Let $\epsilon > 0$ be the error.

Let m be the number of samples in S .

Let I_s be the tightest interval formed by the the points labelled 1 from sample S .

Let $\Pr[I_s]$ denote the probability mass of the interval defined by I_s .

$$\Pr[I_s] > \epsilon$$

Lets assume that the interval I_s (denoted by $[a', b']$) has error ϵ . The error can be found in intervals $[a, a']$ and $(b', b]$ denoted by I_1 and I_2 respectively. If we assume that the error is equally distributed across the two regions, we can denote the error to be $\epsilon/2$ for each of I_1 and I_2 . Each point in the error region has a probability of $(1 - \frac{\epsilon}{2})$. So Probability of the error in sample being greater than ϵ can be written as

$$\begin{aligned} \Pr_{S \sim \mathcal{D}^m} [R(I_S) > \epsilon] &\leq \sum_{i=1}^2 \Pr_{S \sim \mathcal{D}^m} [I \cap I_i = \emptyset] \\ &\leq 2(1 - \epsilon/2)^m \\ &\leq 2e^{-m\epsilon/2} \end{aligned}$$

Equating the RHS to δ gives us the sample complexity.

$$\begin{aligned} 2e^{-m\epsilon/2} &= \delta \\ \frac{2}{\delta} &= e^{m\epsilon/2} \\ \log \frac{2}{\delta} &= \frac{m\epsilon}{2} \\ m &= \frac{2}{\epsilon} \log \frac{2}{\delta} \end{aligned}$$

2

Algorithm A: Given a sample \mathcal{S} for target concept C containing p closed intervals,

1. If there are p separate sequence of positively labeled points in the training data, then return the union of the p tightest intervals containing the positive points.
2. Otherwise, return $(p - i)$ tightest intervals, each containing a sequence of positive labels separated by i negative labels. i can take the values 0 to $(p - 1)$.

Error intervals: Let $[a_1, b_1] \cup [a_2, b_2] \cup \dots [a_p, b_p]$ be the target concept C . Let $\epsilon > 0$. We can assume that $\Pr[a_i, b_i] > \epsilon/(p + 1)$. The actual error on the training set is gap between the target concept $[a_i, b_i]$ and the learned concept $[a'_i, b'_i]$. Each of the error regions r_i can occur with a probability $\epsilon/2(p + 1)$. (Factor of 2 for $[a_i, a'_i]$ and $[b'_i, b_i]$).

Proof of PAC learnability :

If $\text{error}(h_S) > \epsilon$, then either the union of the intervals misses one of the regions r_i or $\Pr[b_i, a_{i+1}] > \epsilon/(p + 1)$. Thus by union bound, we have

$$\begin{aligned} \Pr[\text{error}(h_S)] &\leq \Pr[\exists i \in [1, p] : h_S \text{ misses } r_i] + e^{-m\epsilon/(p+1)} \\ &\leq 2p(1 - \frac{\epsilon}{2(p+1)})^m + e^{-m\epsilon/(p+1)} \\ &\leq 2pe^{-m\epsilon/2(p+1)} + e^{-m\epsilon/(p+1)} \\ &\leq (2p + 1)e^{-m\epsilon/2(p+1)} \end{aligned}$$

Setting $\delta = 0$ and solving the RHS will give us a bound for m .

$$(2p + 1)e^{-m\epsilon/2(p+1)} = \delta \quad (1)$$

$$\frac{2p + 1}{\delta} = e^{m\epsilon/2(p+1)} \quad (2)$$

$$\log \frac{2p + 1}{\delta} = \frac{m\epsilon}{2(p+1)} \quad (3)$$

$$m = \frac{2(p+1)}{\epsilon} \log \frac{2p + 1}{\delta} \quad (4)$$

Therefore, for probability of error to be less than ϵ m should be greater than the value obtained in (4) (Sample complexity).

When p is 2 :

Substituting $p = 2$ in the equation (4) above gives us $m \geq \frac{6}{\epsilon} \log \frac{5}{\delta}$ as the sample complexity.

Time complexity :

The time complexity is $O(p)$.

B. Rademacher complexity, growth function

1

The upper bound on the growth function can be found by finding out the number of dichotomies possible using the given hypothesis set for a given m . For a given value of θ , the functions in H can classify either the points to the left of θ as '+' ($x \mapsto 1_{x \leq \theta} : \theta \in \mathbb{R}$) or as '-' ($x \mapsto 1_{x \geq \theta} : \theta \in \mathbb{R}$). Similarly H can classify points to the right of θ as '+' or '-'.

If we have a sample of size m , we have a total of $(m + 1)$ ways of classifying the sample. Each classification can be done in two ways (as seen in the previous paragraph). Hence we have a total of $2(m+1)$ ways of classification. Of these classifications, two classification will be repeated (the one at the extremities). So we have to subtract 2 from the total ways of classification. Hence total ways of classification is $2(m + 1) - 2 = 2m$.

$$\prod_H(m) = 2m$$

Upper bound on $\mathfrak{R}_m(H)$

The upper bound is given by

$$\begin{aligned} \mathfrak{R}_m(H) &\leq \sqrt{\frac{2 \log \prod_H(m)}{m}} \\ &\leq \sqrt{\frac{2 \log(2m)}{m}} \end{aligned}$$

2

$$\mathfrak{R}_S(H) = \frac{1}{m} E_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h_1(x_i) h_2(x_i) \right] \quad (5)$$

$$= \frac{1}{m} E_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i (\max(0, h_1(x_i) + h_2(x_i) - 1)) \right] \quad (6)$$

$$\leq \frac{1}{m} E_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i (h_1(x_i) + h_2(x_i)) \right] \quad (7)$$

$$\leq \frac{1}{m} E_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h_1(x_i) + \sum_{i=1}^m \sigma_i h_2(x_i) \right] \quad (8)$$

$$\leq \frac{1}{m} E_{\sigma} \left[\sup_{h \in H} \sum_{i=1}^m \sigma_i h_1(x_i) + \sum_{i=1}^m \sigma_i h_2(x_i) \right] \quad (9)$$

$$\leq \frac{1}{m} E_{\sigma} \left[\sup_{h \in H_1} \sum_{i=1}^m \sigma_i h_1(x_i) + \sup_{h \in H_2} \sum_{i=1}^m \sigma_i h_2(x_i) \right] \quad (10)$$

$$\leq \frac{1}{m} E_{\sigma} \left[\sup_{h \in H_1} \sum_{i=1}^m \sigma_i h_1(x_i) \right] + \frac{1}{m} E_{\sigma} \left[\sup_{h \in H_2} \sum_{i=1}^m \sigma_i h_2(x_i) \right] \quad (11)$$

$$\leq \mathfrak{R}_S(H_1) + \mathfrak{R}_S(H_2) \quad (12)$$

The equations and its explanations are as follows:

6 — by rewriting $h_1(x_i)h_2(x_i)$ in 1-lipshitz function form. The lipshitz form is valid as $h_1(x_i)$ and $h_2(x_i)$ take values 0,1.

7 — due to Talagrand's lemma

10 — Since $\sup(f + g) \leq \sup(f) + \sup(g)$

8, 9, 11, 12 — Expanding the equation and replacing with Rademacher complexity terms.