Foundations of Machine Learning — Homework Assignment 1

Anirudhan J Rajagopalan
N18824115
ajr619

October 11, 2015

A. PAC Learning

1

Algorithm $\mathcal{A}$: Given a sample $\mathcal{S}$, Let the algorithm consists of returning the tightest interval I' = I_s containing all points labeled with 1 in sample S.

Error regions: 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

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

Equating the RHS to δ gives us the sample complexity.

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