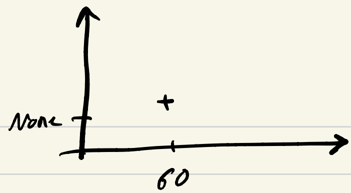
Lecture 13

PAC Learning.

Example 1: Running base on temperature precipitation

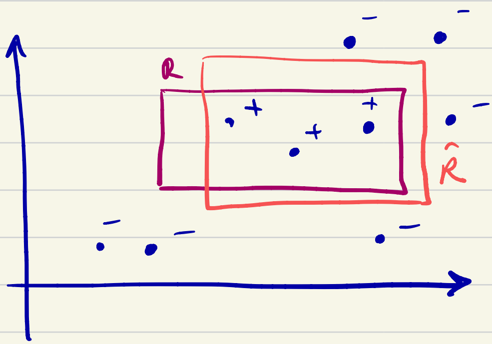


Learning an axis-aligned rectangle R in $\mathbb{R}^2$

Samples: points $p_1, \ldots, p_n \sim D$ over $\mathbb{R}^2$

label $y_1, \ldots, y_n$

$$y_i = \begin{cases} +1 & \text{if } p_i \in R \\ -1 & \text{Otherwise} \end{cases}$$



Goal: output $\hat{R}$ s.t. error of $\hat{R}$ is small (say ϵ) with high probability (say $1-\delta$)

$$\mathrm{err}(\hat{R}) = \Pr_{p \sim D}[\hat{R} \text{ mislabel } p]$$

$$= \Pr_{p \sim D}\left[\begin{array}{lll} (p \in R & \text{and} & p \notin \hat{R}) \\ & \text{or} & \\ (p \notin R & \text{and} & p \in \hat{R}) \end{array}\right]$$

D is arbitrary but fix.

while D can be potentially unusual / irregular, the notion of error is also defined based on the same D.

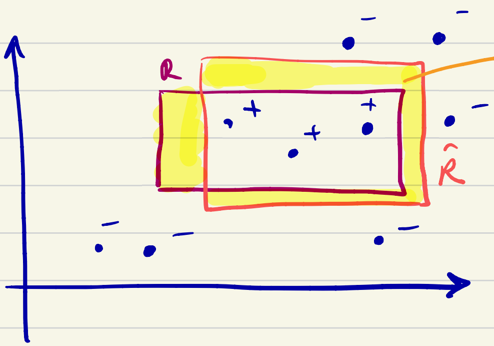
Solution:

Algorithm:

1- Draw m samples (for sufficiently large)

2- set $\hat{R}$ to be a rectangle that

correctly label all the sample points



$A := R \Delta \hat{R}$

the area that we mislabel points

$$err(\hat{R}) = \Pr_{p \sim D}[p \in A] = D(A)$$

by our definition of $\hat{R}$, there is no sample point in $A := R \Delta \hat{R}$

If $err(\hat{R}) > \varepsilon \Rightarrow D(A) > \varepsilon$

How likely it is to not see any sample from A ?

Ideally, we want:

$$\Pr_{D}[\text{ \# samples in } A = 0] \overset{?}{\leq} \delta$$

$$= (1 - D(A))^m \leq (1-\varepsilon)^m \quad \text{(independent samples)}$$

$$\leq e^{-\varepsilon m}$$

set $m = \frac{\log 1/\delta}{\varepsilon}$

$$\leq \delta$$

$\Rightarrow$ Hence, with probability at least $1-\delta$

$err(\hat{R}) \leq \varepsilon$.

efficient
- # samples $= O\left(\frac{\log 1/\delta}{\varepsilon}\right)$
- time $O(m)$

Well behaved target class

Probably Approximately Correct (PAC)

X instance space set of all instances
(input space / domain)

$c : X \to \{+1, -1\}$ concept a function to label elements

C concept class a collection of labeling functions

c^* target concept $c^* \in C$ and label all instances correctly

D target distribution distribution over instances

sample / training data set $\begin{cases} \langle x_1, c^*(x_1) \rangle \\ \langle x_2, c^*(x_2) \rangle \\ \vdots \\ \langle x_n, c^*(x_n) \rangle \end{cases}$

+ "distribution free" setting

samples drawn from an arbitrary distribution. but error is measured according to the same distribution.

Some papers focus on specific class of distributions such as Gaussians.

+ We say we are in the realizable case if there exists a concept $c^* \in C$ that label all the instances in the domain perfectly

+ The goal is to find an unknown target concept c in a known concept class using labeled samples.

- find $\hat{c}$ in C with small error w.h. prob.
- Efficiency: # samples & time

PAC learning (Probably Approximately Correct)

Suppose that we have a concept class C over X. We say that C is PAC learnable if there exists an algorithm A s.t:

$\forall\ c \in C$, $\forall\ D$ over X, $\forall\ \varepsilon, \delta \in (0, 0.5]$

A receives ε, δ, and samples $\langle x_1, c(x_1) \rangle, \ldots, \langle x_n, c(x_n) \rangle$ where x_i's are iid samples from D.

Then, w. p. $\geq 1-\delta$, A outputs $\hat{c}$ [proper $\in C$] s.t.

$$\text{err}(\hat{c}) \leq \varepsilon.$$

The probability is taken over the randomness in the samples and any internal coin flips of A.

+ Usually efficiency means :

sample complexity & time complexity

$= O(\text{ploy}(1/\varepsilon, 1/\delta))$

+ ε = error parameter

δ = confidence parameter

These two parameters capture two kinds of error :

ε: small discrepancy between concepts is not detectable.

δ: with some small probability, the sample set is not representative of reality.