Computational Learning Theory

What is ML?

→ "Supervised machine learning"

(active learning, online learning, privacy...)

What will we study in CLT?

→ Develop a formal model for learning

→ How much data do we need?

→ What are the fastest (fast) algorithms for learning?

→ Are there tradeoffs between space, time, "data"?

→ Are there inherently hard problems?

→ Information theory, game theory, optimization.

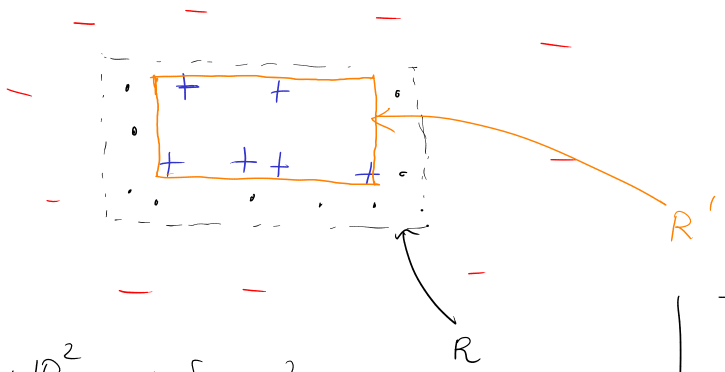
→ Definition; theorem; proof

Administration:

- Weeks 5-8 Hilary Term; Weeks 1-4 Trinity Term
- 4 problem classes. (Friday noon before class)
- Mini-project: (Week 4 trinity (Monday) → Week 6 (Tuesday))

- Probability
- Standard algorithms
- Basic Complexity
-

(Axis-aligned Rectangles)



$C_R : \mathbb{R}^2 \to \{+, -\}$

"Learn" this function?

Tightest fitting rectangle.

How much predictive power do we have?

"Test distribution" vs "Training distribution"

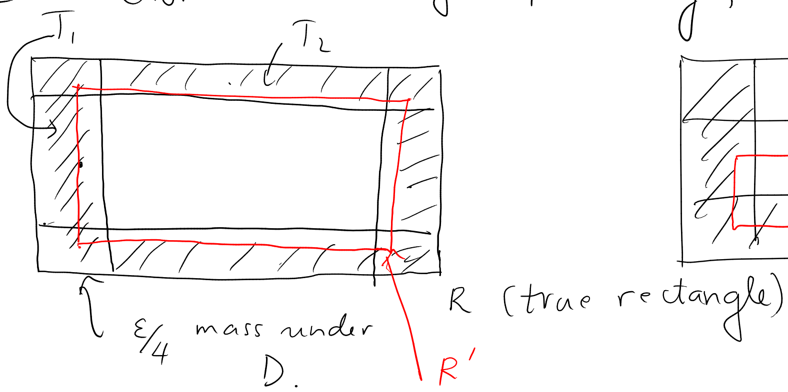
We'll be tested on similar data as we had at the time of training.

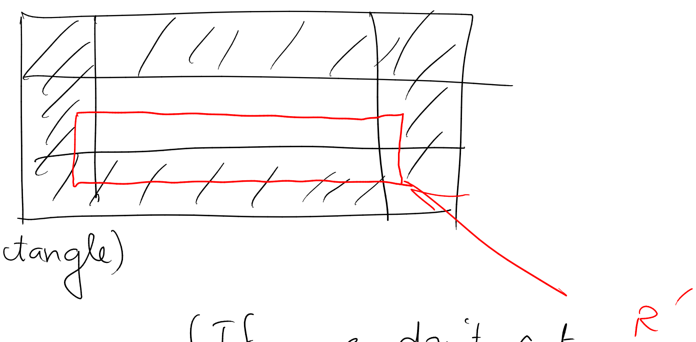
Let D be the distribution over $\mathbb{R}^2$ from which we get data.

Unknown Rectangle (R), s.t. all +ve points inside R and negative points outside R.

Algorithm: Get m points from D labelled according to R, and find the tightest fitting rectangle R'.

(Low error with high probality)





(If we don't get any point from T_2)

At least one point from each of $T_1, \dots, T_4$

High prob

very low probability

A_i is the event that after drawing m points i.i.d. from D and there is no point in T_i.

$$\mathbb{P}(A_i) = \left(1 - \frac{\varepsilon}{4}\right)^m$$

$$\mathcal{E} = A_1 \cup A_2 \cup A_3 \cup A_4 \quad \text{(Bad event)}.$$

if $\mathcal{E}$ does not occur, then error $(R'; R) = \mathbb{P}_D(\text{making a mistake}) \leq \varepsilon$

The Union Bound (Boole's Inequality)

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

$$\mathbb{P}(\mathcal{E}) = \sum_i \mathbb{P}(A_i) \leq 4 \cdot \left(1 - \varepsilon/4\right)^m$$

$$\leq 4 \cdot e^{-m\varepsilon/4} \leq \delta \quad (*)$$

$$1 - x \leq e^{-x}$$

If $\boxed{m \geq \frac{4}{\varepsilon} \log(4/\delta)}$, then $(*)$ is satisfied.

ε: Accuracy

δ: Confidence

m: #data

Probably Approximately Learning (PAC) Learning

- Algorithm doesn't know the target rectangle R, but it does know that data labelled according some rectangle.
- Algorithm has access to training data; tested from the same dist.
- Want "<u>statistically efficient</u>" (# data) & "<u>computationally efficient</u>"

<u>Instance Space</u>: X all possible instances or example seen by the algorithm.

e.g. Rectangle learning game: $X = \mathbb{R}^2$

e.g. $\{0,1\}^{10\times 10}$ 10x10 black and white images

$[0,1]^{100\times 100\times 3}$ 100x100 colour images

<u>Concept Class</u>: A concept c over X is a boolean function $c : X \to \{0,1\}$ (binary classification)

A concept class C over X is a collection of concepts c over X.

$c_R : \mathbb{R}^2 \to \{0,1\}$ which can be represented by some axis-parallel rectangle

---

Learning algorithm has knowledge of the concept class C, but not the specific <u>target</u> concept c.

<u>Distribution</u>: We have some unknown but fixed distribution over X.

"Draw $x \sim D$, label $c(x)$, output $(x, c(x))$."

$EX(c, D)$: example oracle.

<u>Error</u>: Let $h : X \to \{0,1\}$ be some <u>hypothesis</u>,

For target $c \in C$, and distribution D

$$err(h) = err(h; c, D) = \mathbb{P}_{x\sim D}[c(x) \neq h(x)]$$

Sample Complexity: # calls to the oracle $EX(c, D)$

Running Time: Running of the learning algorithm

Accuracy (ε): error of the algorith

Confidence (δ): failure probality of the algorithm

PAC Learning (Take I): Let $\mathcal{C}$ be a concept class over X. We say that $\mathcal{C}$ is PAC learnable if there exists a learning algorithm L, with the following property: $\forall$ target concept $c \in C$, $\forall$ distribution D over X, for all $0 < \varepsilon < 1/2$ & $0 < \delta < 1/2$, if L is given access to $EX(c, D)$ and inputs ε, δ, then with probability at least $1-\delta$, L outputs $h \in C$ satisfying $err(h) \leq \varepsilon$.

- The probability is over $EX(c, D)$ calls + internal randomization.

If L runs in time polynomial in $1/\varepsilon$, $1/\delta$, we say that $\mathcal{C}$ is efficiently PAC-learnable.

Theorem: The class of axis-aligned rectangles is PAC (Take I) learnable. (even efficiently).
