Formulation of the Learning Problem

I. Concept and function learning (realizable case)

$\mathcal{X}$ – feature space

$\mathcal{C}$ – concept class (collection of subsets of $\mathcal{X}$)

$\boxed{a}$ $\boxed{a}$ ... all images of lowercase 'a'

$\boxed{o}$ $\boxed{n}$ ... negative examples

$X_1, \ldots, X_n \overset{iid}{\sim} P$ (instances)

$P \in \mathcal{P}$ (class of distributions)

$c^* \in \mathcal{C}$ (unknown concept)

training data: (X_i, Y_i) $\quad i = 1, \ldots, n$

$$Y_i = 1_{\{X_i \in c^*\}}$$

$Z_i = (X_i, Y_i)$ – instance-label pair

learning algo: $Z^n = (Z_1, \ldots, Z_n) \xrightarrow{A_n} \hat{C}_n \in \mathcal{C}$

$$\mathcal{A} = \{A_n\}_{n=1}^{\infty} \qquad A_n : (\mathcal{X} \times \{0,1\})^n \to \mathcal{C}$$

$\hat{C}_n = A_n(Z_1, \ldots, Z_n)$ – hypothesis returned by $\mathcal{A}$ on Z^n

– training

$X \sim P$ (indep. of $X_1, \dots, X_n$)

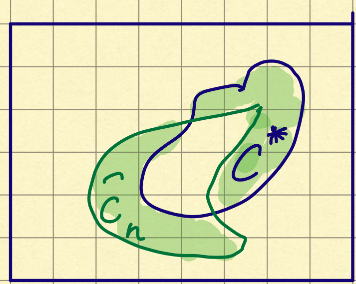
$\hat{Y} = 1_{\{X \in \hat{C}_n\}}$

Misclassification: $X \notin C^*, \; X \in \hat{C}_n$

$X \in C^*, \; X \notin \hat{C}_n$

$\{x \in \mathcal{X} : x \in C^* \cap \hat{C}_n^c \text{ or } x \in (C^*)^c \cap \hat{C}_n\}$

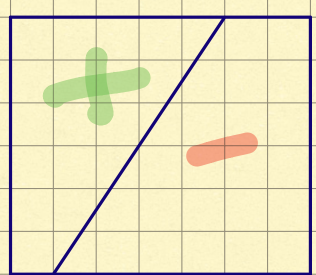
$=: \hat{C}_n \Delta C^*$ (sym. diff)



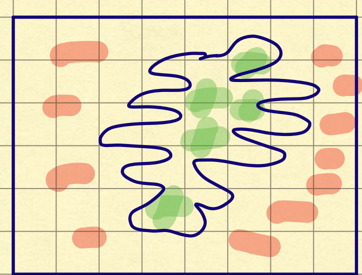
$P[\hat{C}_n \Delta C^*]$ - classification error

↳ r.v. (depends on Z^n)

"Learnability" ⟷ complexity / richness of concepts



"simple"



"complex"

Probably Approximately Correct (PAC):

property of a learning algo for a given task

$$\text{data } (Z^n) \xrightarrow{A_n} \hat{C}_n$$

$P[\hat{C}_n \Delta C^*] \leq \varepsilon$ (approx. correct)

P_C^n - dist. of n iid labeled pts when $C^* = C$

$$P^n_{C^*}\left\{ z^n : P[\hat{C}_n \Delta C^*] > \varepsilon \right\} < \delta$$

– probably approx. correct

(L. Valiant, D. Angluin 1980s

<u>Notation</u>: fix $C \in \mathcal{C}$

(X,Y): $X \sim P$, $Y = 1_{\{X \in C\}}$

P_C = dist. of (X,Y)

P^n_C = dist. of $(X_1,Y_1), \dots, (X_n,Y_n) \overset{iid}{\sim} P_C$

<u>PAC learnability</u>: property of a learning task $(\mathcal{P}, \mathcal{C})$

PAC learnable $\Rightarrow$ $\exists$ a PAC learning algo $\mathcal{A}$ for $(\mathcal{P}, \mathcal{C})$

Fix $\mathcal{A}$: $r_{\mathcal{A}}(n, \varepsilon, P) := \sup_{C \in \mathcal{C}} P^n_C \left\{ P(\hat{C}_n \Delta C) > \varepsilon \right\}$

$$\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) := \sup_{P \in \mathcal{P}} r_{\mathcal{A}}(n, \varepsilon, P)$$

$\mathcal{A}$ is PAC if $\lim_{n \to \infty} \bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) = 0.$
(at accuracy ε)

To prove PAC learnability: produce PAC learning algo

Examples:

1) Any finite concept class is PAC-learnable

$\mathcal{C} = \{C_1, C_2, \dots, C_M\}$ $\quad M < \infty$

Claim: $\exists$ a learning algo $\mathcal{A}$ s.t.

$$\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \leq M(1-\varepsilon)^n$$

where $\mathcal{P}$ is the class of all distributions on $\mathcal{X}$

$n \geq \frac{1}{\varepsilon} \log \frac{M}{\delta}$ examples suffices to give

$$\bar{r}_{\mathcal{A}}(n, \varepsilon, \mathcal{P}) \leq \delta.$$

Proof sketch: $\quad$ data Z^n

$$\mathcal{F}(Z^n) := \left\{ m \in [M] : \; Y_i = 1_{\{X_i \in C_m\}} \; \forall i \right\}$$

$$\hat{m}_n := \min \{ m : m \in \mathcal{F}(Z^n) \}$$

$$\hat{C}_n = C_{\hat{m}_n}$$

"bad set": $\mathcal{B} := \{ m : P(C_m \Delta C_{m^*}) > \varepsilon \}$

$$P^n_{C^*}[\hat{m}_n \in \mathcal{B}] = \sum_{m \in \mathcal{B}} P^n_{C^*}[\hat{m}_n = m]$$

$\hat{m}_n \in \mathcal{F}(Z^n)$: $\quad C_{\hat{m}_n}$ consistent w/ data

$\hat{m}_n \in \mathcal{B}$ $\quad$: $\quad P[C_m \Delta C_{m^*}] > \varepsilon$

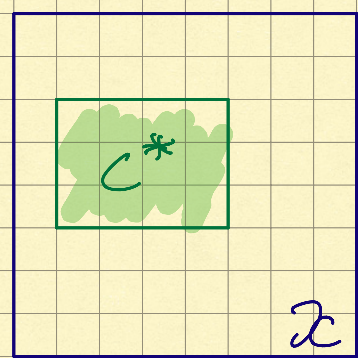
$$P^n_{c^*}[\hat{m}_n = m] \leq P[X_i \notin c_{m^*} \Delta c_m, \forall i]$$
$(m \in B)$
$$\leq (1-\varepsilon)^n$$

$\blacksquare$

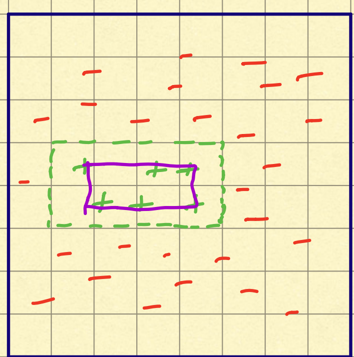
2) Axis-aligned rectangles are PAC learnable

$\mathcal{X} = [0,1]^2$ (unit square)

$\mathcal{C} = \{ [a_1, b_1] \times [a_2, b_2] :$
$a_1, b_1, a_2, b_2 \in [0,1] \}$



Algo:



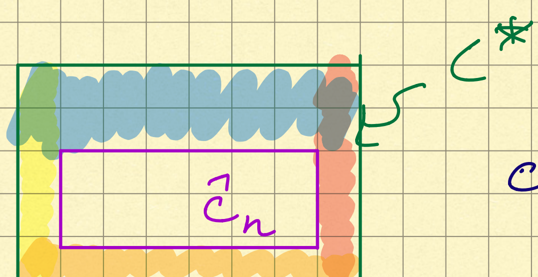
A_n : data $\longrightarrow \hat{c}_n$:
smallest $c \in \mathcal{C}$ that contains all positive examples

Claim: $\bar{r}_A(n, \varepsilon, \mathcal{P}) \leq 4(1 - \varepsilon/4)^n$

all dist. on $[0,1]^2$

(Blumer, Ehrenfeucht, Haussler, Warmuth, 1980s)

idea:



$c^* \Delta \hat{c}_n = \square \cup \square \cup \square \cup \square$

random sets

show that, w.p. $\geq 1 - 4(1-\varepsilon/4)^n$,

each of these random rectangles has $P(\cdot) \leq \varepsilon/4$

$\blacksquare$

$$\bar{r}_A(n, \varepsilon, P) \leq 4(1-\varepsilon/4)^n \leq 4e^{-n\varepsilon/4} \leq \delta$$

if $n \geq \frac{4}{\varepsilon} \log \frac{4}{\delta}$