Active learning for Binary classification

Consider example space $\mathcal{X}$ - space of all images (e.g. satellite photos)

(binary) label space $\{0,1\}$ - contains "human-made object" or not.

Assume: for every $x \in \mathcal{X}$ $\exists$ a corresponding label $y_x \in \{0,1\}$

Hypothesis class $\mathcal{H}$: $\forall h \in \mathcal{H}$ $h: \mathcal{X} \to \{0,1\}$.

"Traditional" passive learning (or supervised learning)

$\exists$ distribution $\mathcal{D}$ over $\mathcal{X}$ and we observe

$$\{(x_t, y_t)\}_{t=1}^{n} \overset{iid}{\sim} \mathcal{D} \qquad x_t \in \mathcal{X},\ y_t \in \{0,1\}$$

Given dataset learn $\hat{h}_n = \underset{h \in \mathcal{H}}{\arg\min} \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}\{h(x_t) \neq y_t\}$

Reason about true risk $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}\{\hat{h}_n(X) \neq Y\}]$.

Active learning selects examples to be labelled

and we evaluate an algorithm based on

both # labels requested, # unlabelled looked at

and $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}\{\hat{h}_n(X) \neq Y\}]$

Question: Can active learning achieve same accuracy as passive learning w/ far fewer labels?

Settings of interest

	Separable	Agnostic
Pool-based setting	Today	
Streaming setting		

Separable setting: $\exists h^* \in \mathcal{H}: \quad Y_x = h^*(x) \quad \forall x \in \mathcal{X}$

i.e. $\exists$ exists a hypothesis that perfectly labels all data

Agnostic setting: Not the separable setting.

When the label for $X \in \mathcal{X}$ is requested, we observe

$Y \sim \text{Bernoulli}(\eta_x)$ where $\eta: \mathcal{X} \to [0,1]$ is arbitrary.

Pool-based setting Example space $\mathcal{X}$ is finite and fixed.

Game proceeds in rounds

Input: $\mathcal{H}, \mathcal{X}$

for $t = 1, 2, \ldots, n$

Learner chooses $x_t \in \mathcal{X}$

Nature reveals y_t

Learner outputs $\hat{h}_n \in \mathcal{H}$ and recieves loss

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{h}_n(x_i) \neq y_i\} \text{ over entire pool } \mathcal{X}$$

Streaming setting $\mathcal{X}$ can be uncountable. Exists a distribution $\mathcal{D}_X$ over $\mathcal{X}$.

for $t = 1, 2, \ldots, n$

Nature reveals $x_t \overset{iid}{\sim} \mathcal{D}_X$

Learner decides request label or not

If yes, nature reveals y_t, else round ends.

Learner recieves loss $\mathbb{E}_{(X,Y)\sim \mathcal{D}_{XY}}[\mathbb{1}\{\hat{h}_n(X) \neq Y\}]$.

Note An algorithm for streaming setting can always be applied to the pool-based setting.

Separable, pool-based setting (Exact setting: identify h^*)

$\mathcal{X}$ is finite, and can be enumerated $\mathcal{X} = \{1, \dots, n\}$
$n = |\mathcal{X}|$

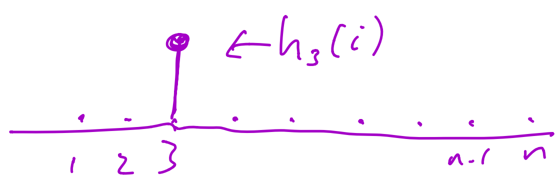
$\mathcal{X}$ finite $\Rightarrow$ $\mathcal{H}$ is finite wLOG

$\exists h^* \in \mathcal{H}: \quad h^*(i) = h^*(x_i) = y_i \quad \forall x_i \in \mathcal{X} \ (i \in \mathcal{X})$
$(i \in [n])$

Ex. Thresholds $\quad h_j(x_i) = \mathbb{1}\{ i \leq j \}, \ \mathcal{H} = \{ h_j : j \in [n] \}$

For this class I can use bisection search to learn $h^* \in \mathcal{H}$ using just $\lceil \log_2 |\mathcal{H}| \rceil$ labels.

Ex. Needle in a haystack $\quad h_j(x_i) = \mathbb{1}\{ i = j \}$



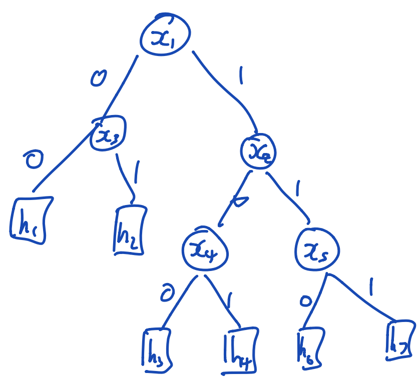
$|\mathcal{H}| - 1$ queries suffice

To identify h^*, nothing is better than exhaustive search

Question: Given arbitrary hypothesis class $\mathcal{H}$, how many queries are necessary and sufficient to identify $h^* \in \mathcal{H}$?

For a deterministic algorithm A let $\varsigma(\mathcal{X}, \mathcal{H}, A, h)$ be the number of labels requested under $h^* = h$, until all other hypotheses are ruled out

WLOG any deterministic algorithm A is a binary tree:



w/ leaves as hypotheses $\mathcal{H}$.

$\#$ leaves $= |\mathcal{H}|$

$\Longrightarrow$ depth of tree $\geq \lceil \log_2 |\mathcal{H}| \rceil$

$\Longrightarrow$ Some $h \in \mathcal{H}$ requires $\geq \lceil \log_2 |\mathcal{H}| \rceil$ labels

Proposition: For any hypothesis class $\mathcal{H}$ we have.

$$\min_{A} \max_{h \in \mathcal{H}} \varsigma(\mathcal{X}, \mathcal{H}, A, h) \geq \lceil \log_2 |\mathcal{H}| \rceil .$$

Extended teaching dimension (Hegedus 1995)

Def We say $S \subset \mathcal{X}$ is a specifying set for $b \in \{0,1\}^n$ wrt $\mathcal{H}$

if $|\{h \in \mathcal{H} : h(x) = b(x)\ \forall x \in S\}| \leq 1$.

Note: b is not necessarily in $\mathcal{H}$.

When $b \in \mathcal{H}$ then a specifying set is sufficient to "teach" the concept $b \in \mathcal{H}$.
When $b \notin \mathcal{H}$ " " to prove that $b \notin \mathcal{H}$

Def For any $\mathcal{X}$, $\mathcal{H}$ define the extended teaching dimension

$$\text{EXT-TD}(\mathcal{H}) = \min\{k : \forall b \in \{0,1\}^n, \exists \text{ spec. set } S \text{ for } b \text{ w/ } |S| \leq k\}.$$

Theorem For any $\mathcal{H}$ we have

$$\text{EXT-TD}(\mathcal{H}) \leq \min_{A} \max_{h \in \mathcal{H}} S(\mathcal{X}, \mathcal{H}, A, h) \leq \text{EXT-TD}(\mathcal{H}) \lceil \log_2 |\mathcal{H}| \rceil$$

Moreover, the Halving algorithm achieves the upper bound.

Ex. $\text{EXT-TD}(\mathcal{H}_{\text{thresholds}}) = 2$.

If $b \in \mathcal{H}$, just need to give example to left + right of threshold value.

If $b \notin \mathcal{H}$ then $\exists$ a "0" before a "1" and so choose any pair

or $b = 0^n$ and just return $S = \{1\}$. since $h(1) = 1$ $\forall h \in \mathcal{H}$.

Ex. $\text{EXT-TD}(\mathcal{H}_{\text{haystack}}) = |\mathcal{H}| - 1$.

Consider $b = 0^n$