The PAC Learning Framework -I

Prof. Dan A. Simovici

UMB

Notations:

- $\mathcal{X}$ is the set of possible *examples* or *instances* (also, the *input space*);
- $\mathcal{Y}$ is the set of all possible labels; initially $\mathcal{Y} = \{0, 1\}$;
- a *concept* is a mapping $c : \mathcal{X} \longrightarrow \mathcal{Y}$.

Concepts can be viewed as

- mappings $c : \mathcal{X} \longrightarrow \{0, 1\}$;
- subsets of $\mathcal{C}$ determined as $\{x \in \mathcal{X} \mid c(x) = 1\}$.

The Learning Problem

Basic assumption: examples in $\mathcal{X}$ are independently and identically distributed (iid) random variables according to a probabilistic distribution $\mathcal{D}$. If X is a random variable having the distribution $\mathcal{D}$ (e.g. binomial, normal, Poission, etc.) we write $X \sim \mathcal{D}$.

- the learner considers a set of possible concepts $\mathcal{H}$ referred to as hypotheses which may or not coincide with $\mathcal{C}$;
- learner receives a sample $S = \{x_1, \ldots, x_m\}$ drawn iid from $\mathcal{X}$ as well as their labels $c(x_1), \ldots, c(x_m)$ which are based on the concept c to be learned;
- the task of the learner is to select a hypothesis $h_S \in \mathcal{H}$ such that

$$P(\{x \in \mathcal{X}, x \sim \mathcal{D} \mid h_S(x) \neq c(x)\})$$

is small.

Samples

An $\mathcal{X}$-sample of size m is a sequence of random variables $S = (x_1, \ldots, x_m)$; m is also the volume of the sample.

- variables $x_1, \ldots, x_m$ are iid according to the distribution $\mathcal{D}$;
- $h(x_1), \ldots, h(x_m)$ are iid random variables ranging over $\{0, 1\}$.

Generalization and Empirical Errors

The number

$$R(h) = P(\{x \in \mathcal{X}, x \sim \mathcal{D} \mid h_S(x) \neq c(x)\})$$

is the generalization error of the hypothesis h. This is the expected error over $\mathcal{D}$ and it is not computable by the learner since $\mathcal{D}$ and c are unknown. The empirical error of h is the random variable

$$\begin{aligned} \hat{R}(h) &= \frac{1}{m} \sum_{i=1}^{m} |\{x_i \mid h(x_i) \neq c(x_i)\}|, \\ &= \frac{1}{m} \sum_{i=1}^{m} I_{h(x_i) \neq c(x_i)}, \end{aligned}$$

which is the average error over the sample S.

The Relationship between the generalization and the empirical error

Note that the empirical error is a random variable because x_i in the definition of $\hat{R}$ are random variables.
By the linearity of expectations we have

$$\begin{aligned} E[\hat{R}(h)] &= \frac{1}{m}\sum_{i=1}^{m} E[I_{h(x_i)\neq c(x_i)}], \\ &= \frac{1}{m}\sum_{i=1}^{m} E_{x\sim\mathcal{D}}[I_{h(x)\neq c(x)}], \\ &\quad \text{(because all variables } x_i \text{ has the same distribution } \mathcal{D} \text{ as } x) \\ &= R(h). \end{aligned}$$

Probably Approximately Correct Learning

- size(c): the maximal cost of the representation of a concept $c \in \mathcal{C}$;
- $O(n)$ an upper bound on the cost of a representation of an example $x \in \mathcal{X}$ (e.g. if $x \in \mathbb{R}^n$ the cost of representing x is $O(n)$);
- $\mathcal{D}$ distribution on $\mathcal{X}$.

PAC Learning Definition

Definition

A concept class $\mathcal{C}$ is PAC-learnable if there is an algorithm $\mathcal{A}$ such that

- for all probability distributions $\mathcal{D}$ on $\mathcal{X}$,
- for any target concept $c \in \mathcal{C}$,
- for any $\epsilon > 0$ and $\delta > 0$,
- for representation cost n of an example in $\mathcal{X}$,

there is a polynomial p such that if the size m of the sample S is such that $m \geqslant p\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c)\right)$, then $\mathcal{A}$ produces a hypothesis h_S such that

$$P(R(h_S) \leqslant \epsilon) \geqslant 1 - \delta.$$

If q is a polynominal such that $\mathcal{A}$ runs in polynomial time $q\left(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \text{size}(c)\right)$, then $\mathcal{C}$ is said to be efficiently PAC-learnable and $\mathcal{A}$ is referred as a PAC-learning algorithm.

Comments on the PAC Definition

- $\mathcal{C}$ is PAC-learnable if the hypothesis returned by $\mathcal{A}$ after observing a number of examples polynomial in $1/\epsilon$ and $1/\delta$ is approximately correct (with generalization error less than ϵ) with high probability;
- δ is used to define the confidence $1-\delta$;
- ϵ gives the accuracy $1-\epsilon$;

If the running time is polynomial, then the sample size m must also be polynomial in $1/\epsilon$ and $1/\delta$.

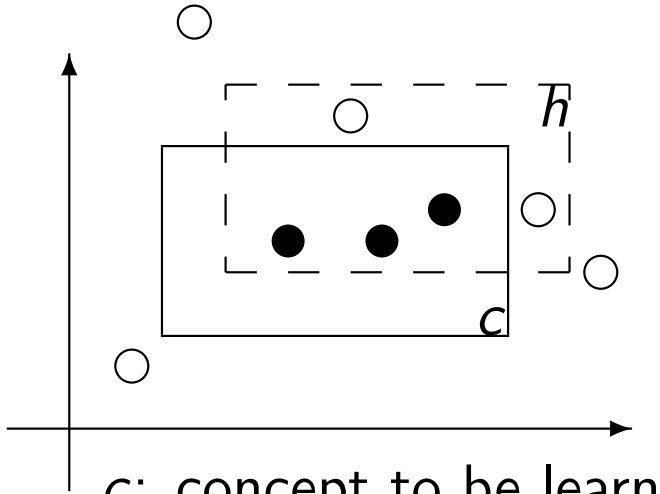
Other Features of the PAC Defintion

- PAC-learning is distribution-free;
- the training examples and the test sample used to define the error are drawn according to the same distribution;
- PAC deals with learnability for a concept class $\mathcal{C}$, not a particular example.

The parameters n and size(c) will be typically omitted.

Example

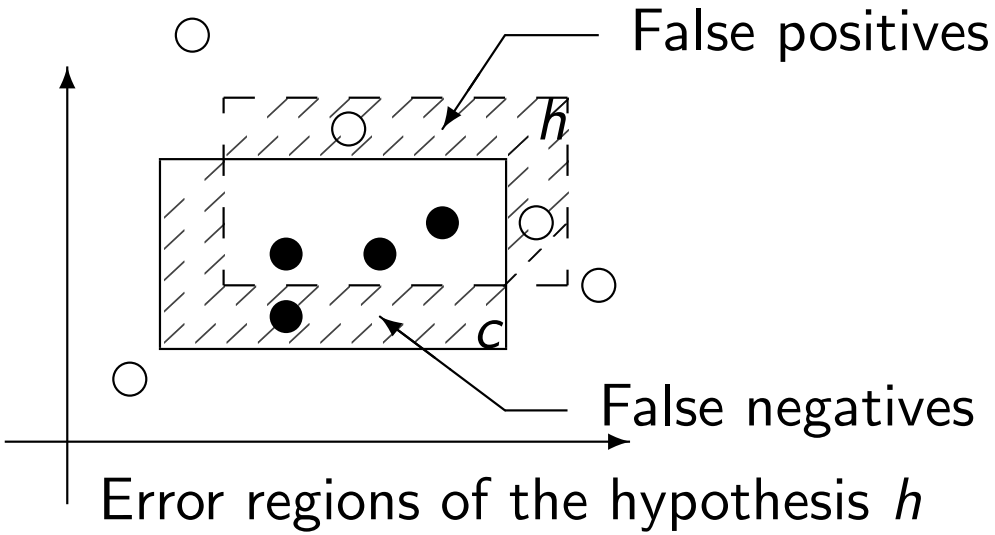
Let $\mathcal{X} = \mathbb{R}^2$ and let $\mathcal{R}$ be the set of all rectangles in $\mathbb{R}^2$.



c: concept to be learned; h: hypothesis

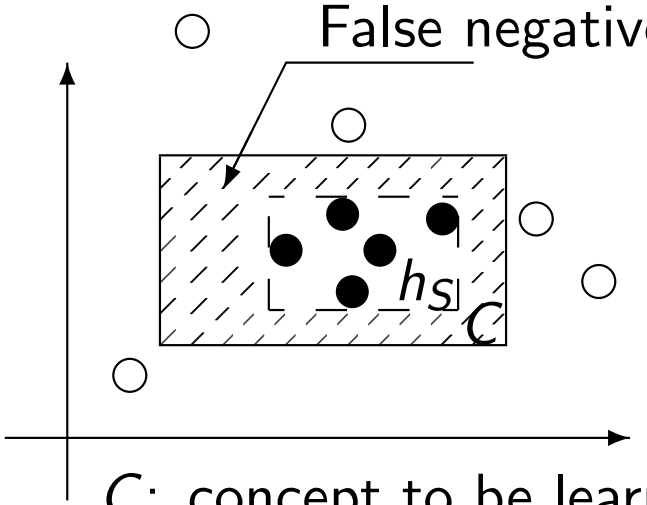
A concept is a particular rectangle.

Example (cont'd)



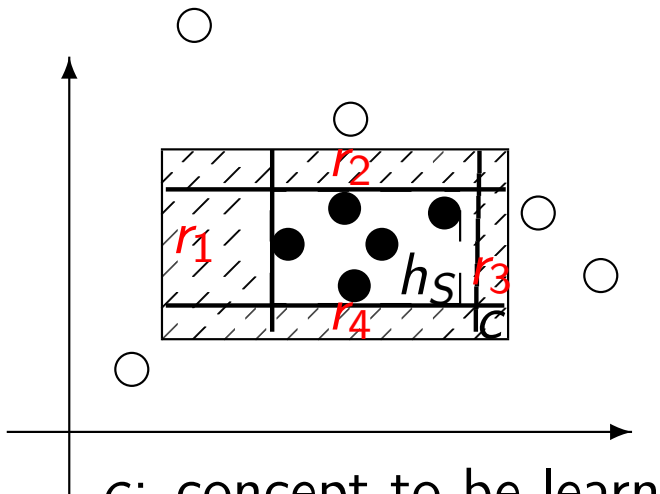
Example (cont'd)

Sample size is m, $S = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$.
Algorithm $\mathcal{A}$ consists in returning the tightest rectangle h_S that contains all positive examples.



C: concept to be learned; H: hypothesis

We are estimating the probability that the generalization error of h_S is greater than ϵ, that is, that the area between c and h_S has probability greater than ϵ. That area is covered by four rectangles r_1, r_2, r_3, r_4 each with probabilities $\frac{\epsilon}{4}$. Each of these rectangles can be obtained by starting with an empty rectangle along a side and increasing the rectangle until the probability equals $\frac{\epsilon}{4}$.



c: concept to be learned; h_S: hypothesis

To ensure that the probability of the hypothesis h_S is at least $1-\epsilon$ is equivalent to saying that the probability that the generalization error of h_S is less than ϵ.
If probability of the hypothesis h_S is at least $1-\epsilon$, the rectangle that corresponds to h_S must intersect at least one of the rectangles r_1, r_2, r_3, or r_4. This allows us to write

$$\begin{aligned} P(R(h_S) > \epsilon) &\leqslant P_{S\sim\mathcal{D}}\left(\bigcup_{i=1}^{4}[(h_S \cap r_i)] = \emptyset\right) \\ &\leqslant \sum_{i=1}^{4} P\left((h_S \cap r_i) = \emptyset\right) \\ &\leqslant 4\left(1-\frac{\epsilon}{4}\right)^m \leqslant 4e^{-\frac{m\epsilon}{4}}, \end{aligned}$$

by the inequality $1-x \leqslant e^{-x}$ for all $x \in \mathbb{R}$.

To ensure $P(R(h_S) > \epsilon) \leqslant \delta$ we impose $4e^{-\frac{m\epsilon}{4}} \leqslant \delta$, which implies

$$m \geqslant \frac{4}{\epsilon} \log \frac{4}{\delta}.$$

Conclusions

If $u = e^{-x}$, the inequality $1 - x \leqslant e^{-x}$ is equivalent to $\log u \leqslant u - 1$. Therefore, if

$$m \geqslant \frac{4}{\epsilon}\left(\frac{4}{\delta} - 1\right),$$

it follows that $m \geqslant \frac{4}{\epsilon}\log\frac{4}{\delta}$. The role of the polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ is played by

$$p\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right) = \frac{4}{\epsilon}\left(\frac{4}{\delta} - 1\right),$$

which shows that the class of axis-aligned rectangles in PAC learnable.

Further Conclusions

- For any $\epsilon > 0$ and $\delta > 0$ if the sample size m is larger than $\frac{4}{\epsilon} \log \frac{4}{\delta}$, then $P(R(h_S) > \epsilon) \leqslant \delta$.
- The class of axis-aligned rectangles in PAC learnable.
- There is no error on the sample S for the hypothesis h_S (we say that h_S is consistent).

Finite Hypothesis Space; the Consistent Case

Theorem

Let H be a finite set of functions $H = \mathcal{Y}^{\mathcal{X}}$ and let $\mathcal{A}$ be an algorithm that returns a consistent hypothesis h_S for any target concept $c \in H$ and iid sample S.
For any $\epsilon, \delta > 0$ the inequality

$$P_{S \sim \mathcal{D}^m}(R(h_S) \leqslant \epsilon) \geqslant 1 - \delta$$

holds if

$$m \geqslant \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right).$$

Proof

Note that the consistency condition that h_S satisfies means that $\hat{R}(h_S) = 0$.

Fix $\epsilon > 0$. $\mathcal{A}$ can select any of the consistent hypotheses h_S of H. We need to upper bound the probability that any of the consistent hypotheses of H ($\hat{R}(h) = 0$) will have a generalization error more than ϵ ($R(h) > \epsilon$), that is,

$$P\left((\exists h \in H) \mid \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\right).$$

Proof (cont'd)

We have:

$$
\begin{aligned}
& P\left((\exists h \in H) \mid \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\right) \\
= \quad & P\left(\bigcup_{i=1}^{|H|}(h_i \in H, \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon)\right) \\
\leqslant \quad & \sum_{i=1}^{|H|} P(h_i \in H, \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon) \\
\leqslant \quad & \sum_{i=1}^{|H|} P(h_i \in H, \hat{R}(h_i) = 0 | R(h_i) > \epsilon).
\end{aligned}
$$

because

$$P(h_i \in H, \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon) \leqslant P(h_i \in H, \hat{R}(h_i) = 0 | R(h_i) > \epsilon),$$

from the definition of conditional probability.

Proof (cont'd)

If $h \in H$ is a consistent hypothesis ($\hat{R}(h) = 0$) with $R(h) > \epsilon$, then

$$P\left(\hat{R}(h) = 0 \mid R(h) > \epsilon\right) \leqslant (1-\epsilon)^m,$$

because, h is consistent with S and the probability of this happening when the generalization error rate of h is at least ϵ is smaller than $(1-\epsilon)^m$ when S has size m.
Therefore,

$$P\left((\exists h \in H) \mid \hat{R}(h) = 0 \text{ and } R(h) > \epsilon\right) \leqslant |H|(1-\epsilon)^m,$$

If we require

$$|H|(1-\epsilon)^m \leqslant \delta$$

we have

$$\log|H| + m\log(1-\epsilon) \leqslant \log\delta,$$

and taking into account that $1-\epsilon < e^{-\epsilon}$, it suffices to require

$$\log|H| - \epsilon m \leqslant \log\delta,$$

which yields

$$m \geqslant \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right).$$

Comments on the Theorem

- when the hypothesis H is finite a consistent algorithm $\mathcal{A}$ is a PAC-learning algorithm;
- learning algorithms benefit from larger sample sizes;
- growth in the sample size is only logarithmic in the size of H.

Conjunctions of Boolean Literals

Example

$\mathcal{C}_n$ the concept class consists of conjunctions of Boolean literals $x_1, \ldots, x_n, \overline{x}_1, \ldots, \overline{x}_n$. There are 3^n such conjunctions.
For $n = 4$, an example is $x_1 \wedge \overline{x}_3 \wedge x_4$.

- A positive example for this concept is $(1, 0, 0, 1)$; a negative example is $(0, 1, 0, 1)$.
- The existence of a positive example like $(1, 0, 0, 1)$ for a concept c implies that c may not contain $\overline{x}_1$ or $\overline{x}_4$.
- A negative example is less informative because we do not know which of its bits is incorrect.

Conjunctions of Boolean Literals - the Algorithm

An algorithm for finding a consistent hypothesis:

- algorithm is based on positive examples;
- for each positive example $\boldsymbol{b} = (b_1, \ldots, b_n)$ and $1 \leqslant i \leqslant n$, if $b_i = 1$, then $\overline{x}_i$ is excluded; if $b_i = 0$, then x_i is ruled out;
- the conjunction of all literals not ruled out is a hypothesis consitent with the target.

Example of Algorithm Application

	b_1	b_2	b_3	b_4	b_5	b_6	
1.	0	1	1	0	1	1	+
2.	0	1	1	1	1	1	+
3.	0	0	1	1	0	1	−
4.	0	1	1	1	1	1	+
5.	1	0	0	1	1	0	−
6.	0	1	0	0	1	1	+

1. exclude $x_1, \overline{x}_2, \overline{x}_3, x_4, \overline{x}_5, \overline{x}_6$
2. exclude $x_1, \overline{x}_2, \overline{x}_3, \overline{x}_4, \overline{x}_5, \overline{x}_6$
4. exclude $x_1, \overline{x}_2, \overline{x}_3, \overline{x}_4, \overline{x}_5, \overline{x}_6$
6. exclude $x_1, \overline{x}_2, x_3, x_4, \overline{x}_5, \overline{x}_6$.

Hypothesis: $\overline{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$.

Since $|H| = |\mathcal{C}_n| = 3^n$ we have

$$m \geqslant \frac{1}{\epsilon}\left(n \log 3 + \log \frac{1}{\delta}\right).$$

The class of conjuctions of at most n Boolean literals is PAC-learnable. For $\delta = 0.02$, $\epsilon = 0.1$ and $n = 10$ the bound is $m \geqslant 148.98$. Thus, if $m \geqslant 149$ with a probability of at least 98% the accuracy is 90%.