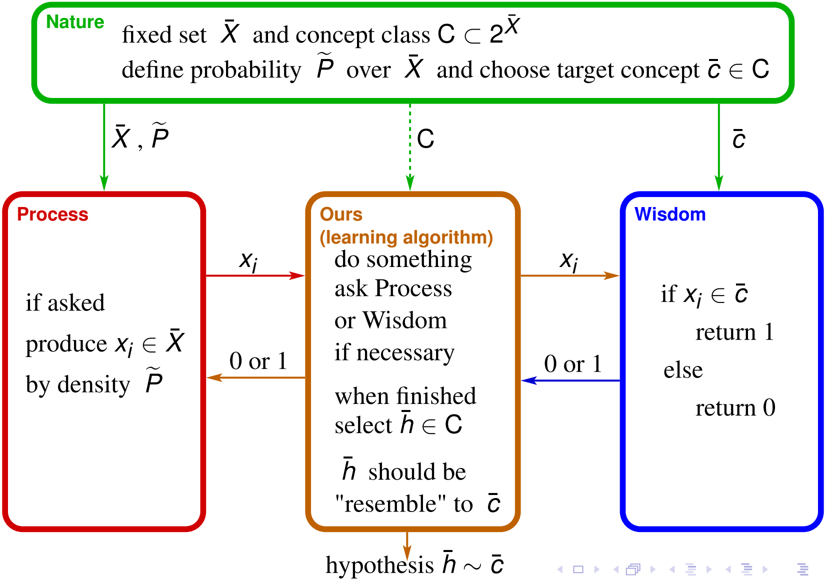
PAC learning model

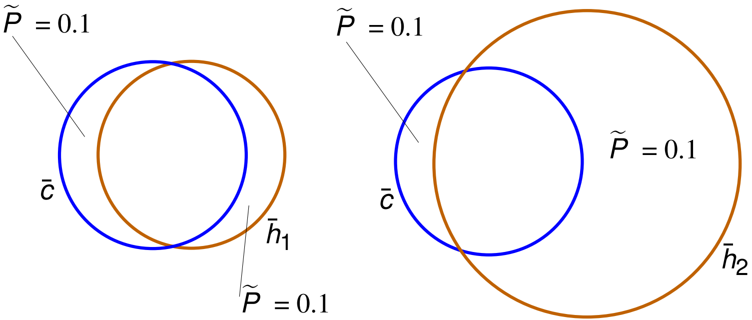
František Hakl

ICS AS CR
hakl@cs.cas.cz

Mar 2013


Nature
fixed set $\bar{X}$ and concept class $C \subset 2^{\bar{X}}$
define probability $\widetilde{P}$ over $\bar{X}$ and choose target concept $\bar{c} \in C$
$\bar{X}$, $\widetilde{P}$
C
$\bar{c}$
Process
if asked
produce $x_i \in \bar{X}$
by density $\widetilde{P}$
x_i
0 or 1
Ours
(learning algorithm)
do something
ask Process
or Wisdom
if necessary
when finished
select $\bar{h} \in C$
$\bar{h}$ should be
"resemble" to $\bar{c}$
x_i
0 or 1
Wisdom
if $x_i \in \bar{c}$
return 1
else
return 0
hypothesis $\bar{h} \sim \bar{c}$

" $\bar{h}$ should be resemble to $\bar{c}$ "



Definition

1. $e_{\widetilde{P}}(\bar{h}, \bar{c}) \stackrel{\text{def}}{=} \widetilde{P}(\bar{c} \triangle \bar{h}) \left(= \left(\bar{c} \dot{-} \bar{h}\right) \cup \left(\bar{h} \dot{-} \bar{c}\right)\right)$
2. $\bar{h}$ is consistent if and only if $\{x_i, \ldots, x_m\} \cap (\bar{c} \triangle \bar{h}) = \emptyset$

Definition (sample space)

Let $\breve{x} \stackrel{\text{def}}{=} \{x_1, \dots, x_m\}$, $x_i \in \bar{X}$, $i \in \{1, \dots, m\}$, $\vec{\boldsymbol{z}} \in \{-1, +1\}^m$ and let $\bar{c} \subset \bar{X}$. Then the ordered tuple

$$\left(\breve{x}, \vec{\boldsymbol{z}}\right)$$

is a m-SAMPLE OF CONCEPT $\bar{c}$ if and only if

$$(\forall i \in \{1, \dots, m\}) \left((x_i \in \bar{c}) \Leftrightarrow \left(\vec{\boldsymbol{z}}_i = 1\right)\right).$$

For concept class C define SAMPLE SPACE OF CONCEPT CLASS as

$$\bar{S}_{\mathrm{C}} \stackrel{\text{def}}{=} \bigcup_{m \geq 1} \left\{ \bigcup_{\bar{c} \in \mathrm{C}} \left\{ \left(\breve{x}, \vec{\boldsymbol{z}}\right) \middle| \left(\breve{x}, \vec{\boldsymbol{z}}\right) \text{ is a } m\text{-sample of concept } \bar{c} \right\} \right\}.$$

Definition ((ϵ, δ)-learning algorithm)

1. (ϵ, δ)-LEARNING ALGORITHM is each mapping $\widetilde{A^*} : \bar{S}_C \to C$ such that for all $\bar{c} \in C$, $\epsilon, \delta \in (0, 1)$ and $\widetilde{P}$ on $\bar{X}$, the probability of the set

$$\left\{ \breve{x} \,\middle|\, \left(\breve{x}, \vec{\boldsymbol{z}}\right) \text{ is } m\text{-sample of } \bar{c} \text{ and } e_{\widetilde{P}}\left(\bar{c}, \widetilde{A^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right)\right)\right) \geq \epsilon \right\}$$

is smaller than the number δ.

2. If such a learning algorithm exists we say that C IS UNIFORMLY LEARNABLE .

Theorem

Let us assume that C *is a concept class over finite set* $\bar{X}$ *and* $\mathrm{H} = \mathrm{C}$*. Then, for each learning algorithm* A^* *requiring*

$$\frac{1}{\epsilon}\ln\left(\frac{|\mathrm{C}|}{\delta}\right)$$

queries and producing for the given concept $\bar{c} \in \mathrm{C}$ *a consistent hypothesis it holds that*

$$Prob_{\widetilde{P}}\left(\mathrm{e}_{\widetilde{P}}\left(\bar{c}, \widetilde{\mathrm{A}^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right)\right)\right) \geq \epsilon\right) < \delta.$$

Definition

For any arbitrary $\mathrm{R} \subset 2^{\bar{X}}$ and for any arbitrary probability density $\widetilde{P}$ defined on $\bar{X}$ and for an arbitrary $\epsilon > 0$ let us define $\mathrm{R}_{\widetilde{P},\epsilon} \stackrel{\text{def}}{=} \left\{\bar{r} \in \mathrm{R} \,\middle|\, Prob_{\widetilde{P}}(\bar{r}) > \epsilon\right\}$. Then, we will call $\bar{T}_{\widetilde{P},\epsilon} \subset \bar{X}$ ϵ-TRANSVERSAL R just if

$$\left(\forall \bar{r} \in \mathrm{R}_{\widetilde{P},\epsilon}\right)\left(\bar{r} \cap \bar{T}_{\widetilde{P},\epsilon} \neq \emptyset\right)$$

Example

$\bar{X} = \langle 0, 1\rangle^n$, $\widetilde{P}$ uniform on $\bar{X}$, $\mathrm{R} = \left\{\bar{b} \subset \bar{X} \,\middle|\, \bar{b} \text{ is a ball}\right\}$, $\epsilon = \frac{1}{z}$.

Then $\bar{T} = \left\{k\epsilon \,\middle|\, k = 0, .., \frac{1}{\epsilon}\right\}^n$ is a $\left(\frac{\pi^{\frac{n}{2}}(\sqrt{n}\epsilon)^n}{\widetilde{\Gamma}\left(\frac{n}{2}+1\right)2^n}\right)$-transversal of R .

…if hypotheses $\bar{h}$ produced by an algoritm is consistent and has $\mathrm{e}_{\widetilde{P}}\left(\bar{h}, \bar{c}\right) > \epsilon$, then $\{x_i, \ldots, x_m\}$ can't be ϵ-transversal of the system $\mathrm{R} \stackrel{\text{def}}{=} \left\{\bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathrm{H}\right\}$ …

Definition

For each $m \geq 1$, $\epsilon > 0$ let

$$\bar{Q}_{m,\epsilon} \stackrel{\text{def}}{=} \left\{ \breve{x} \in \bar{X}^m \,\middle|\, \breve{x} \text{ do not form } \epsilon\text{-transversal of R} \right\}$$

and (assume that $\breve{x}, \breve{y} \in \bar{X}^m$)

$$\bar{J}_\epsilon^{2m} \stackrel{\text{def}}{=} \left\{ \breve{xy} \in \bar{X}^{2m} \,\middle|\, \left(\exists \bar{r} \in \mathrm{R}_{\widetilde{P},\epsilon}\right) \left(\breve{x} \cap \bar{r} = \emptyset \text{ and } \left|\breve{y} \cap \bar{r}\right| \geq \frac{\epsilon m}{2} \right) \right\}.$$

…the probability of the set $\bar{Q}_{m,\epsilon}$ is a probability of producing consistent hypothesis with error $\mathrm{e}_{\widetilde{P}}\left(\bar{h}, \bar{c}\right) > \epsilon$ …

Definition

1. The class H is well-behaved if the sets $\bar{Q}_{m,\epsilon}$ and $\bar{J}_{\epsilon}^{2m}$ are measurable for any probability $\widetilde{P}$, any $m \geq 1$, $\epsilon > 0$, and any system of sets $\mathrm{R} \stackrel{\text{def}}{=} \left\{ \bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathrm{H} \right\}$, where $\bar{c}$ is an arbitrary Borelian set.
2. The class $\mathrm{H} \subset 2^{\bar{X}}$ is universally separable, if there exists a countable subset T of the class H such that for all $\bar{h} \in \mathrm{H}$ there exists a sequence $\{\bar{h}_i\}_1^{\infty}$ of sets from T such that

$$\left(\forall x \in \bar{X}\right)\left(\exists n \geq 1\right)\left(\left(\forall i \geq n\right)\left(x \in \bar{h}_i \text{ if and only if } x \in \bar{h}\right)\right).$$

Theorem

If H *is universally separable, then* H *is well-behaved.*

Lemma

Let $\mathrm{R} \neq \emptyset$ *be a concept class and* $\widetilde{P}$ *be probability on* $\bar{X}$ *for which* $\bar{Q}_{m,\epsilon}$ *and* $\bar{J}_{\epsilon}^{2m}$ *are measurable for all* $m \geq 1$, $\epsilon > 0$. *Then for each* $\epsilon > 0$ *and* $m \geq \frac{2}{\epsilon}$

$$Prob_{\widetilde{P}^m}\left(\bar{Q}_{m,\epsilon}\right) < 2Prob_{\widetilde{P}^{2m}}\left(\bar{J}_{\epsilon}^{2m}\right) \leq 2\Pi_{\mathrm{R}}\left(2m\right)2^{-\frac{\epsilon m}{2}}.$$

Lemma

Let $m \geq 1$, $\bar{c} \subset \bar{X}$, $\mathrm{H} \subset 2^{\bar{X}}$ *and* $\mathrm{R} \stackrel{def}{=} \left\{\bar{h} \triangle \bar{c} \,\middle|\, \bar{h} \in \mathrm{H}\right\}$. *Then*

$$\Pi_{\mathrm{R}}\left(m\right) = \Pi_{\mathrm{H}}\left(m\right) .$$

Lemma

Let $d = \mathrm{VC}_{dim}\left(\mathrm{H}\right) \in N$, $\epsilon, \delta \in (0,1)$, $\Gamma \stackrel{def}{=} 2\Pi_{\mathrm{H}}\left(2m\right)2^{-\frac{\epsilon m}{2}}$. *Then*

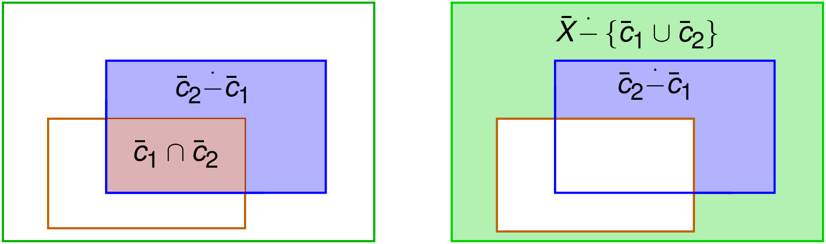
$$m \geq \max\left\{\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right\} \Rightarrow \Gamma \leq \delta .$$

Definition

Concept class C is called nontrivial iff

$$(\exists \bar{c}_1, \bar{c}_2 \in \mathrm{C}) \left(\bar{c}_1 \neq \bar{c}_2 \text{ and } \left(\bar{c}_1 \cap \bar{c}_2 \neq \emptyset \text{ or } \bar{c}_1 \cup \bar{c}_2 \neq \bar{X}\right)\right) .$$

Concept class is called trivial in other cases.



Two cases of minimal content of nontrivial concept class.
(colored sets are nonempty)

Theorem (main result of PAC theory)

Let C *be a nontrivial, well-behaved class. Then:*

1. *If* $\mathrm{VC}_{dim}(\mathrm{C}) = d < +\infty$. *Then*
    1. *for any* $0 < \epsilon < \frac{1}{2}$ *there is no* (ϵ, δ)*-learning algorithm with number of queries less than*

    $$\max\left(\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right), d\left(1 - 2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)\right) . \quad (1)$$

    2. *for arbitrary* $0 < \epsilon < 1$, *any learning algorithm using at least*

    $$\max\left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\epsilon}\log_2\left(\frac{12.611}{\epsilon}\right)\right) \quad (2)$$

    queries and returning a consistent hypothesis is an (ϵ, δ)*-learning algorithm.*
2. C *is uniformly learnable* **if and only if** $\mathrm{VC}_{dim}(\mathrm{C}) < +\infty$.

Sketch of the proof:

1. 1. - $\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right)$: (c&c) Any nontrivial concept class can be reduced to one of the cases discussed above. For uniform probability we get a contradiction.
      - $d\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)$: (c&c) Reduce $\bar{X}$ to d-element subset with uniform probability. Then use the "matrix" $\boldsymbol{Z}_{\bar{c},\bar{h}} \stackrel{\text{def}}{=} \mathrm{e}_{\widetilde{P}}\left(\bar{c},\bar{h}\right)$ to show, that $m > d\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)$ imply that $\left(\exists \bar{h}^*\right)$ contradicts (ϵ,δ)-property ... "broadly speaking".
   2. See previous slides.
2. - $\Leftarrow$ (construction) Use Zermelo's well-ordering theorem to well-order $\bar{H}$. Let algorithm get m-sample of $\bar{c}$ and return the first hypothesis consistent with $\bar{c}$. The statemet follows from 1)-2).
   - $\Rightarrow$ (by contradiction) For any $d \in N$ we carry out steps 1)-1)-(second term). Choose (ϵ,δ) such that $\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right) > 0$. Hence m can't be upper-bounded.

Definition (discrete delta rule)

Let $(\vec{\boldsymbol{x}}_1, y_1), \ldots, (\vec{\boldsymbol{x}}_t, y_t)$ be a given sequence of tuples in $\Re^n \times \{-1, +1\}$, $t \geq 1$. Further, let vector's sequence $\{\vec{\boldsymbol{w}}_i\}_1^\infty$ satisfy the following recursive formulas

1. put $\vec{\boldsymbol{w}}_1 \stackrel{\text{def}}{=} \vec{\boldsymbol{0}}$, $k = 1$
2. let $k = k + 1$ and $\bar{J} \stackrel{\text{def}}{=} \left\{ j \in \{1, \ldots, t\} \,\middle|\, \widetilde{sgn}\left(\langle \vec{\boldsymbol{w}}_k \,|\, \vec{\boldsymbol{x}}_j \rangle\right) \neq y_j \right\}$
   1. if $\bar{J} = \emptyset$ put $\vec{\boldsymbol{w}}_{k+1} = \vec{\boldsymbol{w}}_k$ and STOP,
   2. else let $j_k \in \bar{J}$ be arbitrary. Then put

$$\vec{\boldsymbol{w}}_{k+1} \stackrel{\text{def}}{=} \vec{\boldsymbol{w}}_k + y_{j_k} \vec{\boldsymbol{x}}_{j_k}$$

   and REPEAT step 2).

Then we say that $\{\vec{\boldsymbol{w}}_i\}_1^\infty$ is DELTA SEQUENCE of $(\vec{\boldsymbol{x}}_1, y_1), \ldots, (\vec{\boldsymbol{x}}_t, y_t)$.

Theorem (delta rule convergence)

Let $\{\vec{\boldsymbol{w}}_i\}_1^\infty$ *is a delta sequence and let there exists a vector* $\widehat{\vec{\boldsymbol{w}}}$ *such that for all indexes* $i \in \{1, \ldots, t\}$ *holds* $\widetilde{sgn}\left(\left\langle \widehat{\vec{\boldsymbol{w}}} \,\middle|\, \vec{\boldsymbol{x}}_i \right\rangle\right) = y_i$*. Further let*

$$\alpha \stackrel{def}{=} \max_{i \in \{1,\ldots,t\}} \left\{ \left\| \vec{\boldsymbol{x}}_i \right\|^2 \right\} \quad \text{and} \quad \beta \stackrel{def}{=} \min_{i \in \{1,\ldots,t\}} \left\{ \left| \left\langle \widehat{\vec{\boldsymbol{w}}} \,\middle|\, \vec{\boldsymbol{x}}_i \right\rangle \right| \right\} > 0 \,.$$

Then there exists an natural number $z > 0$ *satisfying* $\vec{\boldsymbol{w}}_{z+1} = \vec{\boldsymbol{w}}_z$ *and* z *can be estimated as*

$$z \leq \frac{\alpha \left\| \widehat{\vec{\boldsymbol{w}}} \right\|^2}{\beta^2} + 1 \,.$$

Theorem (delta rule complexity)

There exists a linearly separable dichotomy of the $\{-1, +1\}^n$ *such that any integer linear separator* $(\vec{\boldsymbol{w}}, t)$ *of this dichotomy satisfies estimation*

$$2^{\frac{n-2}{2}} \leq \sum_{k=1}^{n} \left|\vec{\boldsymbol{w}}_k\right| + |t|.$$

Definition (Mangasarian LP)

Let $\bar{A} \stackrel{\text{def}}{=} \{\vec{\boldsymbol{a}}_1, \ldots, \vec{\boldsymbol{a}}_i\}$ and $\bar{B} \stackrel{\text{def}}{=} \left\{\vec{\boldsymbol{b}}_1, \ldots, \vec{\boldsymbol{b}}_j\right\}$ be a finite subsets of the $\Re^n$. Then MANGASARIAN LINEAR PROBLEM is defined as the problem find vectors $\vec{\boldsymbol{y}} \in \Re^i$, $\vec{\boldsymbol{z}} \in \Re^j$, $\vec{\boldsymbol{w}} \in \Re^n$ and $t \in \Re$ that minimizes

$$\sum_{\alpha=1}^{i} \vec{\boldsymbol{y}}_\alpha + \sum_{\beta=1}^{j} \vec{\boldsymbol{z}}_\beta$$

subject to

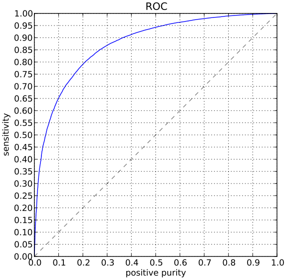
$$\begin{array}{rcll}
\vec{\boldsymbol{y}}_\alpha + \left\langle \vec{\boldsymbol{w}} \middle| \vec{\boldsymbol{a}}_\alpha \right\rangle - t & \geq & 1 & \text{for} \quad \alpha \in \{1, \ldots, i\} \\
\vec{\boldsymbol{z}}_\beta - \left\langle \vec{\boldsymbol{w}} \middle| \vec{\boldsymbol{b}}_\beta \right\rangle + t & \geq & 1 & \text{for} \quad \beta \in \{1, \ldots, j\} \\
\vec{\boldsymbol{y}}_\alpha & \geq & 0 & \text{for} \quad \alpha \in \{1, \ldots, i\} \\
\vec{\boldsymbol{z}}_\beta & \geq & 0 & \text{for} \quad \beta \in \{1, \ldots, j\} .
\end{array}$$

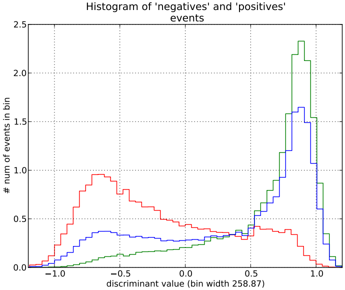
Theorem

Let $\bar{A} \stackrel{def}{=} \{\vec{\boldsymbol{a}}_1, \ldots, \vec{\boldsymbol{a}}_i\}$ *and* $\bar{B} \stackrel{def}{=} \left\{\vec{\boldsymbol{b}}_1, \ldots, \vec{\boldsymbol{b}}_j\right\}$ *be a finite subsets of the* $\Re^n$ *. Then*

1. *There exists a linear separator of the sets* $\bar{A}$ *and* $\bar{B}$ *if and only if the optimal value of the corresponding Mangasarian LP is zero.*
2. *If the optimal value of the corresponding Mangasarian LP is zero and* $\left(\vec{\boldsymbol{y}}^*, \vec{\boldsymbol{z}}^*, \vec{\boldsymbol{w}}^*, t^*\right)$ *is optimal solution, than* $\left(\vec{\boldsymbol{w}}^*, t^*\right)$ *is linear separator of the sets* $\bar{A}$ *and* $\bar{B}$ *.*

. . . a real case . . .





The necessary condition on consistency of hypotheses produced is unsatisfied, PAC model isn't applicable. We have to use

Probably Approximately Optimal (PAO) model

$\mathcal{PAC}$