RUHR-UNIVERSITÄT BOCHUM

RUB

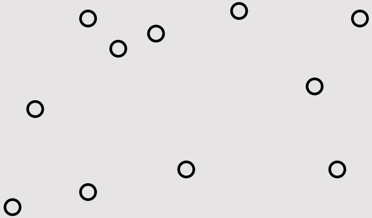
UNLABELED DATA DOES PROVABLY HELP

Malte Darnstädt, Hans Ulrich Simon
and Balázs Szörényi
Ruhr-University Bochum and University of Szeged

30TH SYMPOSIUM ON THEORETICAL ASPECTS OF COMPUTER SCIENCE, KIEL
February 28, 2013

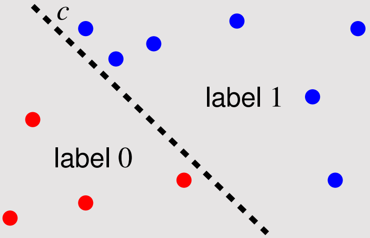
Outline

A theory of the learnable [Valiant 1984]



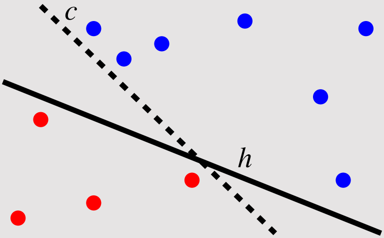
- Instance space X with distribution P
- m sample points drawn according to P

A theory of the learnable [Valiant 1984]



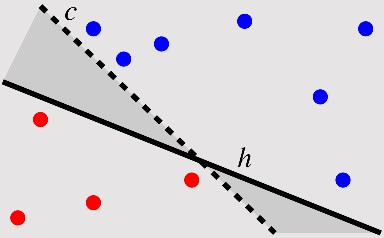
- Concept class: $C \subseteq \{c \mid c : X \to \{0, 1\}\}$
- Target concept: $c \in C$
- the sample is labeled by c

A theory of the learnable [Valiant 1984]



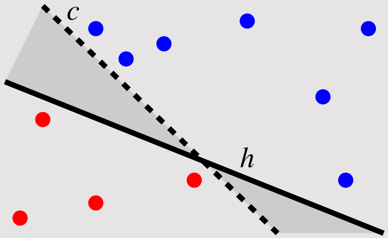
- Hypothesis class: $H \supseteq C$
- Learning algorithm designed for C
  Input: the labeled sample, Output: a hypothesis $h \in H$

A theory of the learnable [Valiant 1984]



- Hypothesis class: $H \supseteq C$
- Learning algorithm designed for C
  Input: the labeled sample, Output: a hypothesis $h \in H$

A theory of the learnable [Valiant 1984]



- The learner **PAC-learns** the class C if for all $\varepsilon, \delta > 0$ and large enough m:

$$P_{\text{sample}}\left(P_x(h(x) \neq c(x)) \leq \varepsilon\right) \geq 1 - \delta$$

A well known sample size bound

Theorem (Blumer, Ehrenfeucht, Haussler, Warmuth 1987)

Any finite class C is PAC-learnable from

$$m_{C,P}(\varepsilon,\delta) = O\Big(\frac{\log(|C|)+\log(1/\delta)}{\varepsilon}\Big)$$

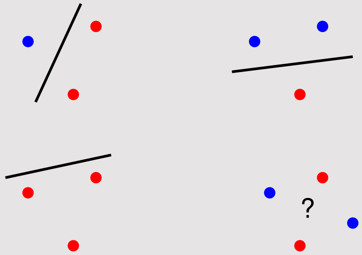
many labeled sample points under any distribution P.
The learner may output any concept from C which is consistent with the sample.

The VC dimension

- Measure of the **capacity** of a class C
- Shattering set: points from X that can be labeled arbitrarily by concepts from C
- The VC dimension of C is the size of the largest shattering set (or infinite if no such set exists)

The VC dimension – example

Class of half spaces: three points can be shattered



A second well known sample size bound

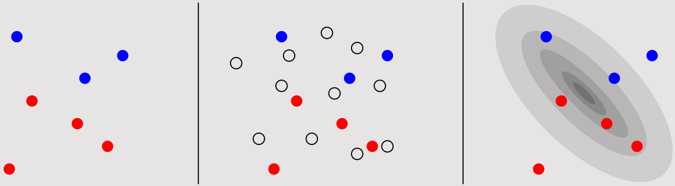
Theorem (Blumer, Ehrenfeucht, Haussler, Warmuth 1989)

A class C with finite VC-dimension d is PAC-learnable from

$$m_{C,P}(\varepsilon,\delta) = O\Big(\frac{d \cdot \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\Big)$$

many labeled sample points under any distribution P.
The learner may output any concept from C which is consistent with the sample.

Using unlabeled data



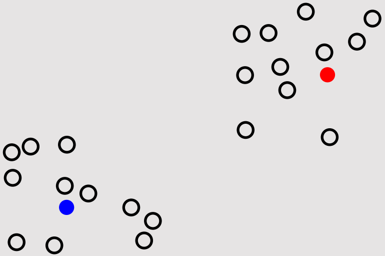
supervised learning

semi supervised learning

learning under
a fixed distribution

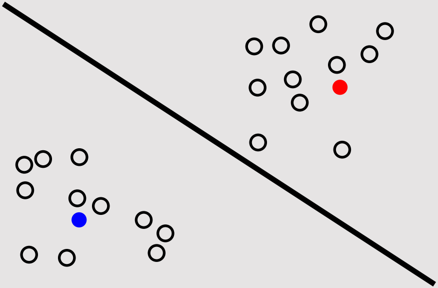
Learning becomes easier when additional unlabeled data is available. How much can we reduce the need for labels?

With an extra assumption



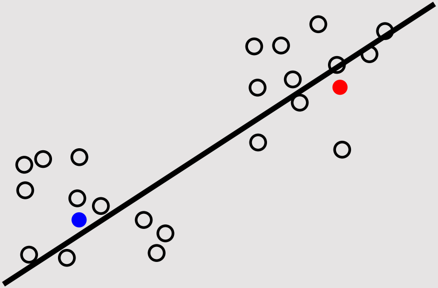
Usually used together with an **extra assumption**, that connects the data distribution with the target concept.
The number of needed labels can be reduced dramatically [e.g. Balcan, Blum 2010; Darnstädt, Simon, Szörényi 2011].

With an extra assumption



Usually used together with an **extra assumption**, that connects the data distribution with the target concept.
The number of needed labels can be reduced dramatically [e.g. Balcan, Blum 2010; Darnstädt, Simon, Szörényi 2011].

With an extra assumption



Usually used together with an **extra assumption**, that connects the data distribution with the target concept.
The number of needed labels can be reduced dramatically [e.g. Balcan, Blum 2010; Darnstädt, Simon, Szörényi 2011].

Without extra assumptions?

Conjecture by Ben-David, Lu and Pál 2008:
Learning under a fixed distribution without extra assumptions can only save a **constant factor** of labels.

- Proven for some special concept classes over the real line
- Similar result by Darnstädt and Simon 2011:
  Proven for finite classes and “most” pairs of target concepts and data distributions

Without extra assumptions?

Conjecture by Ben-David, Lu and Pál 2008:
Learning under a fixed distribution without extra assumptions can only save a **constant factor** of labels.

- Proven for some special concept classes over the real line
- Similar result by Darnstädt and Simon 2011:
  Proven for finite classes and "most" pairs of target concepts and data distributions

Today

The conjecture is false.

Where unlabeled data doesn't help

Theorem (1)

Let C contain the all-one and all-zero function.

1. *If C is finite, then for all P:*

$$\frac{\text{supervised} \quad m_{C,P}(2\varepsilon,\delta)}{\text{fixed distribution} \quad m_{C,P}(\varepsilon,\delta)} = O(\log(|C|))$$

2. *If the VC dimension d of C is finite, then for all P:*

$$\frac{\text{supervised} \quad m_{C,P}(2\varepsilon,\delta)}{\text{fixed distribution} \quad m_{C,P}(\varepsilon,\delta)} = O(d \cdot \log(1/\varepsilon))$$

Proof idea

Compare

- a lower bound on the number of needed labels for any fixed distribution learner
- with an upper bound for a specially designed supervised learner

Where unlabeled data does help

Theorem (2)

There exists a concept class C_* and a family of distributions $\mathscr{P}_*$ such that:

1. *There exists a semi supervised learner that learns successfully under any distribution $P \in \mathscr{P}_*$ with*

$$m_{C_*,P}(\varepsilon,\delta) = O\left(\frac{1}{\varepsilon^2} + \frac{\log(1/\delta)}{\varepsilon}\right)$$

2. *For any supervised learner and $\varepsilon < \frac{1}{2}$ it holds that*

$$\sup_{P \in \mathscr{P}_*} m_{C_*,P}(\varepsilon,\delta) = \infty$$

Where unlabeled data does help

Can we prove a statement like the first theorem with a universal constant not depending on the concept class?
Answer: No

Theorem (3)

There exists a sequence of concept classes $(C_n)_{n\geq 1}$ with arbitrary large VC dimensions and families of distributions $(\mathscr{P}_n)_{n\geq 1}$ such that the statement of Theorem 2 still holds (with an additional supremum over $n \geq 1$).

The classes C_* and C_n

Let $X_* = \{0,1\}^*$ and $X_n = \{0,1\}^n$.
C_* consist of all functions on words from X_* given by

$$c_i(x) = \begin{cases} 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0 \text{ or } i > |x| \end{cases}$$

for all $i \geq 0$. C_n is C_* restricted to functions on words of length n.

The classes C_* and C_n

Let $X_* = \{0,1\}^*$ and $X_n = \{0,1\}^n$.
C_* consist of all functions on words from X_* given by

$$c_i(x) = \begin{cases} 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0 \text{ or } i > |x| \end{cases}$$

for all $i \geq 0$. C_n is C_* restricted to functions on words of length n.

- The classes C_n are not exotic or artificially constructed
- All classes of infinite VC dimension contain subclasses that are significantly easier to learn in the semi-supervised setting
- Classes and distributions similar to C_* and $\mathscr{P}_*$ were first published by Dudley, Kulkarni, Richardson and Zeitouni in 1994

The distribution families $\mathscr{P}_*$ and $\mathscr{P}_n$

For $i \geq 1$ let

$$p_i = \frac{1}{\log_2(3+i)}$$

Let σ be any permutation on $\{1,\ldots,n\}$. Then the distribution P_σ on X_n is defined by setting

$$P_\sigma(x_{\sigma(i)} = 1) = p_i$$

independently for each $1 \leq i \leq n$.
Now define

$$\mathscr{P}_n = \{P_\sigma\}_\sigma, \quad \mathscr{P}_* = \bigcup_{n \geq 1} \mathscr{P}_n$$

Proof sketch – 1st part

C_* is semi supervised PAC learnable under any $P_\sigma \in \mathscr{P}_*$

- Use the following upper bound of the sample complexity for learning under a fixed distribution [Benedek, Itai 1991]:

$$O\left(\frac{\log(|\frac{\varepsilon}{2}\text{-cover of } C|) + \log(1/\delta)}{\varepsilon}\right)$$

- The p_i fall fast enough that an $\varepsilon/2$-cover of C_* of size $2^{2/\varepsilon}$ exists
- Without knowledge of P_σ:
  use unlabeled data to find a cover with high probability
  (apply Chernoff bounds)

Proof sketch – 2nd part

C_* under $\mathscr{P}_*$ is hard to learn for a supervised learner

- A game between the learner and an adversary, who plays a probabilistic strategy
- The adversary picks a large enough n, chooses a random permutation σ and selects $c_{\sigma(1)}$ as the target

$x_{\sigma(1)}$	$x_{\sigma(2)}$	$x_{\sigma(3)}$	$x_{\sigma(4)}$	$x_{\sigma(5)}$	$\cdots$	$x_{\sigma(n)}$	label
0	1	0	1	0	$\cdots$	0	
1	0	1	0	1	$\cdots$	0	
0	0	0	0	0	$\cdots$	0	
1	1	1	0	1	$\cdots$	0	
0	0	0	0	0	$\cdots$	0	
0	1	0	0	0	$\cdots$	0	
1	0	1	0	1	$\cdots$	0	
1	1	0	0	1	$\cdots$	1	

$x_{\sigma(1)}$	$x_{\sigma(2)}$	$x_{\sigma(3)}$	$x_{\sigma(4)}$	$x_{\sigma(5)}$	$\cdots$	$x_{\sigma(n)}$	label
0	1	0	1	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
1	1	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
0	1	0	0	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
1	1	0	0	1	$\cdots$	1	?

$x_{\sigma(1)}$	$x_{\sigma(2)}$	$x_{\sigma(3)}$	$x_{\sigma(4)}$	$x_{\sigma(5)}$	$\cdots$	$x_{\sigma(n)}$	label
0	1	0	1	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
1	1	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
0	1	0	0	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
1	1	0	0	1	$\cdots$	1	?

$x_{\sigma(1)}$	$x_{\sigma(2)}$	$x_{\sigma(3)}$	$x_{\sigma(4)}$	$x_{\sigma(5)}$	$\cdots$	$x_{\sigma(n)}$	label
0	1	0	1	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
1	1	1	0	1	$\cdots$	0	1
0	0	0	0	0	$\cdots$	0	0
0	1	0	0	0	$\cdots$	0	0
1	0	1	0	1	$\cdots$	0	1
1	1	0	0	1	$\cdots$	1	?

- p_i falls slowly enough that the first column is repeated arbitrarily often for large n with high probability (Borel-Cantelli lemma)
- Let J be the set of indices of these columns

Proof sketch – 2nd part

Bayes strategy to label a test instance x

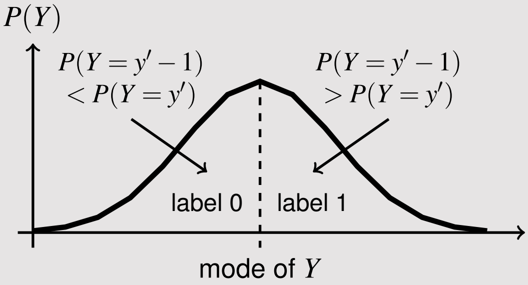
- By symmetry, the decision only depends on the value y' of the random variable Y':

$$Y' = \sum_{j \in J} x_j$$

- Choose label 1 iff $P(Y' = y' \,|\, x_{\sigma(1)} = 1) > P(Y' = y' \,|\, x_{\sigma(1)} = 0)$
  iff $P(Y = y' - 1) > P(Y = y')$

  where $Y = Y' - x_{\sigma(1)}$

Proof sketch – 2nd part



- Depends on the true label only if Y hits its mode!
- The Bayes error is at least $\frac{1}{2}\left(1 - P(Y = \text{mode}(Y))\right)$
- This value is large by the Lindeberg-Feller Central Limit theorem

Open questions

- Are the bounds of Theorem 1 tight?
- For which other classes and distributions do Theorem 2 or 3 hold?
- Can we extend the results to the agnostic setting?

Thank you!

Do you have any questions?