Incorporating Unlabeled Data in the Learning Process

11/06/2013

Maria Florina Balcan

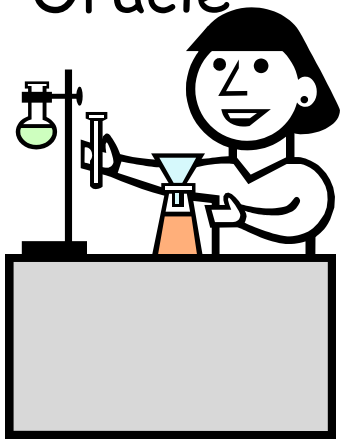
Supervised Passive Learning

Data Source

Learning Algorithm

Unlabeled examples

Expert / Oracle



Labeled examples

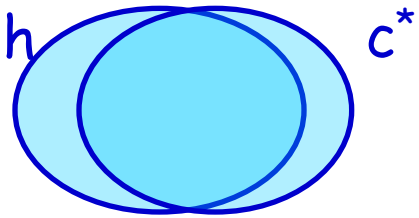
Algorithm outputs a classifier

Standard Passive Supervised Learning

- X – feature space
- S={(x, l)} - set of labeled examples
  - drawn i.i.d. from distr. D over X and labeled by target concept c^*

- Do optimization over S, find hypothesis $h \in C$.
- Goal: h has small error over D.

$$err(h)=Pr_{x \in D}(h(x) \neq c^*(x))$$



- c* in C, realizable case; else agnostic

Classic models: PAC (Valiant), SLT (Vapnik)

Standard Supervised Learning Setting

Sample Complexity well understood

Sample Complexity, Finite Hyp. Space, Realizable case

Theorem

$$m \geq \frac{1}{\varepsilon}\left[\ln(|C|) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in C$ with $err(h) \geq \varepsilon$ have $\hat{err}(h) > 0$.

Sample Complexity: Uniform Convergence Bounds

- Infinite Hypothesis Case

Theorem

$$m = O\left(\frac{1}{\varepsilon}\left[VCdim(C)\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in C$ with $err(h) \geq \varepsilon$ have $\hat{err}(h) > 0$.

E.g., if C - class of linear separators in R^d, then we need roughly $O(d/\varepsilon)$ examples to achieve generalization error ε.

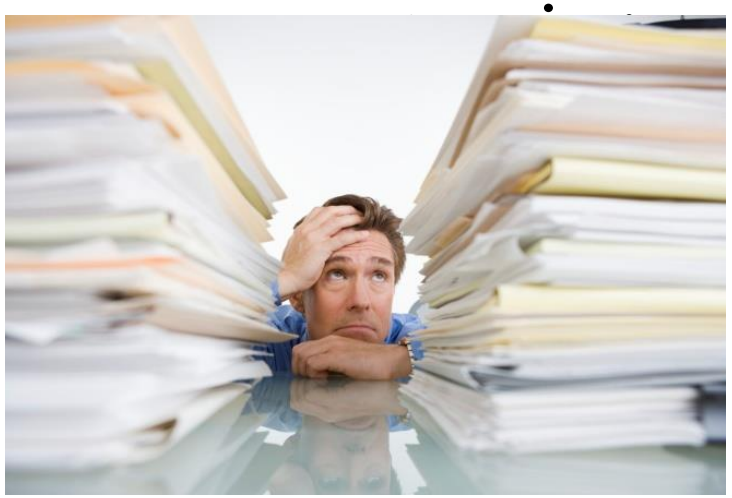
Non-realizable case – replace ε with ε^2.

- **In PAC, can also talk about efficient algorithms.**

Incorporating Unlabeled Data in the Learning process

Modern applications: lots of unlabeled data, labeled data is

- Web page, document classification
- OCR, Image classification
- Classification pbs in Computational Biology



Incorporating Unlabeled Data & Interaction

Areas of significant activity in modern ML.

- Semi-Supervised Learning

  Using cheap unlabeled data in addition to labeled data.

- Active Learning

  The algorithm interactively asks for labels of informative examples.

Foundations lacking a few years ago.

Does unlabeled data help?

Does interaction help?

Why and by how much?

Incorporating Unlabeled Data & Interaction

Areas of significant activity in modern ML.

- Semi-Supervised Learning

  Using cheap unlabeled data in addition to labeled data.

- Active Learning

  The algorithm interactively asks for labels of informative examples.

Foundations lacking a few years ago. 

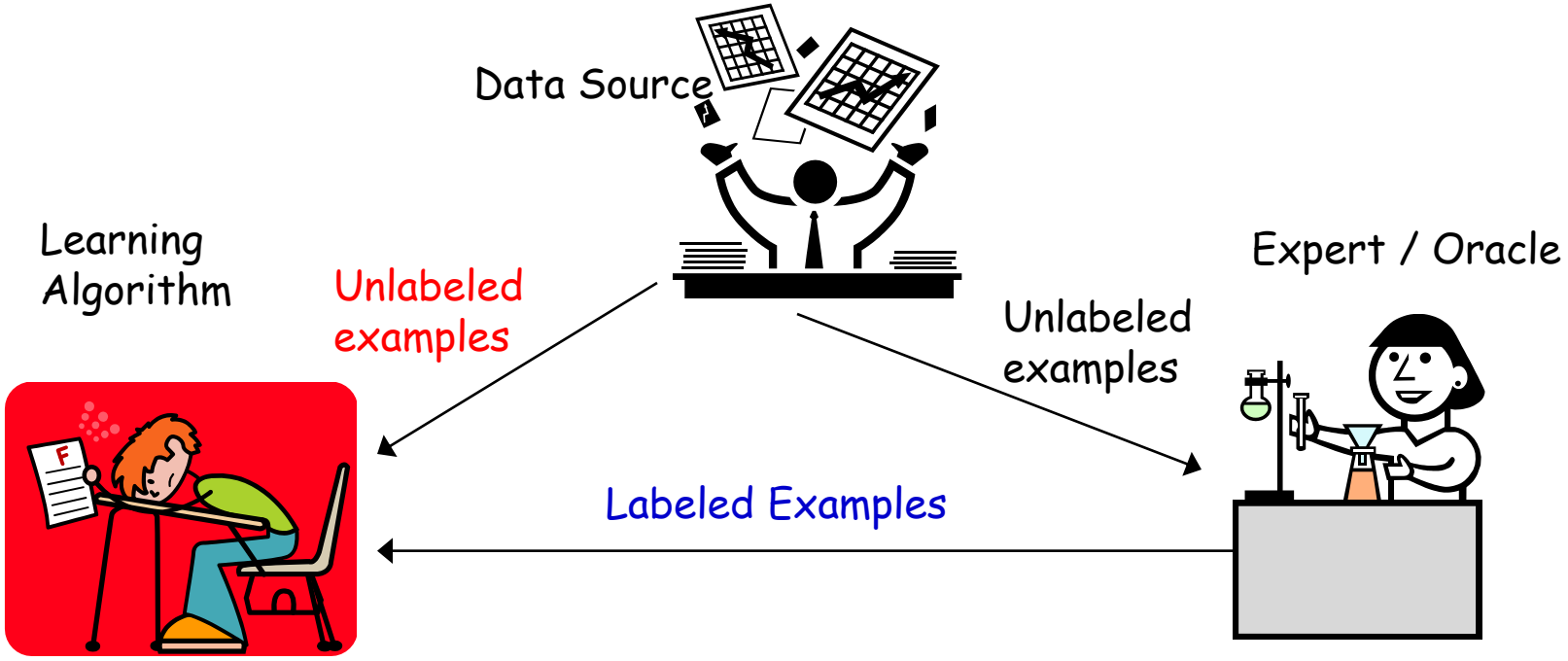
Significant progress recently. 

Mostly on understanding Sample Complexity.

Semi-Supervised Learning

$S_u=\{x_i\}$ - unlabeled examples i.i.d. from D

$S_l=\{(x_i, y_i)\}$ – labeled examples i.i.d. from D, labeled by target c*.



Algorithm outputs a classifier

Semi-Supervised Learning

- Variety of methods and experimental results. E.g.,:
  - Transductive SVM [Joachims '98]
  - Co-training [Blum & Mitchell '98]
  - Graph-based methods [Blum & Chawla01], [Zhu-Lafferty-Ghahramani'03]

- Scattered and very specific theoretical results (prior to 2005).

A general discriminative (PAC, SLT style) framework for SSL.

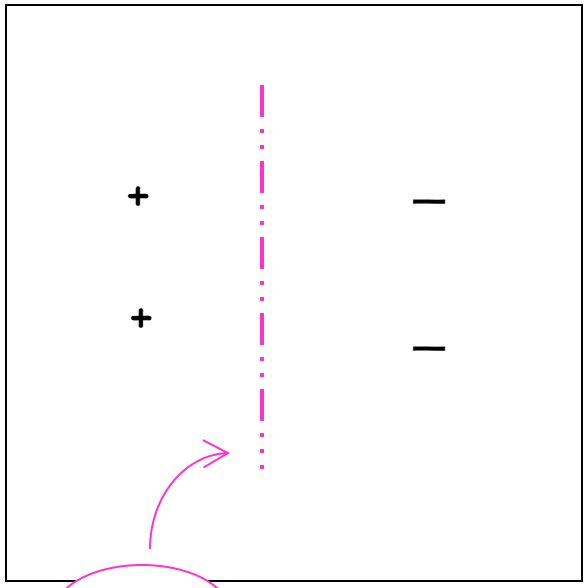
[Balcan-Blum, COLT 2005; JACM 2010; book chapter, 2006]

Challenge: capture many of the assumptions typically used.

Different SSL algs based on very different assumptions.

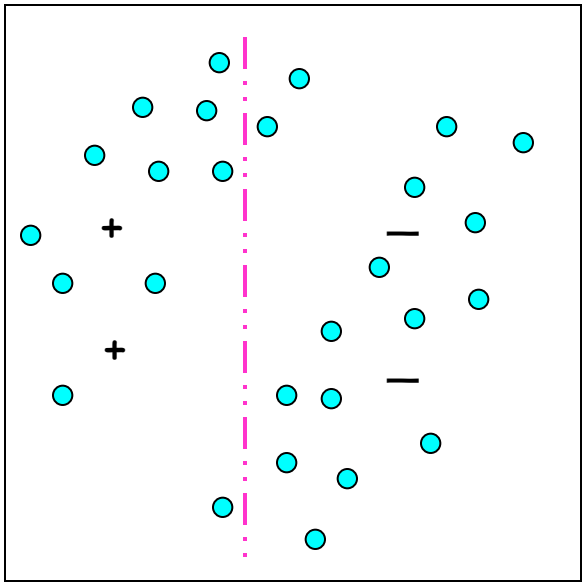
Example of "typical" assumption: Margins

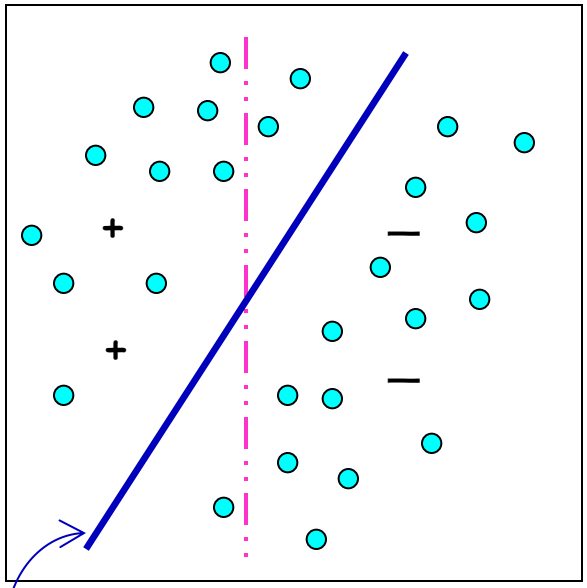
Belief: target goes through low density regions (large margin).



SVM

Labeled data only





Transductive SVM

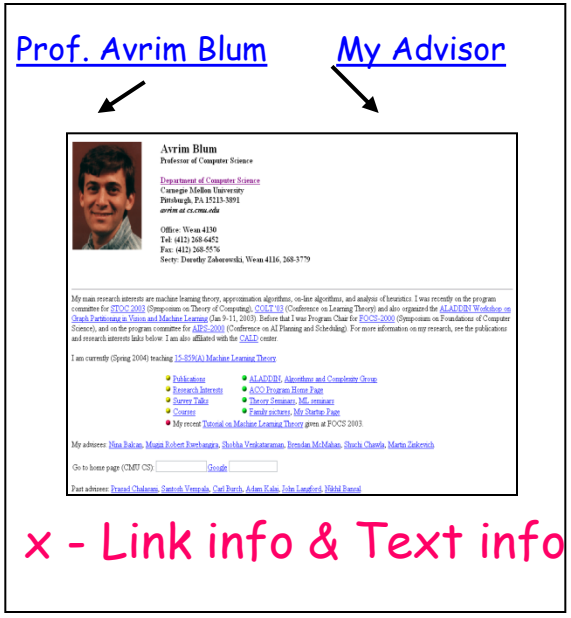
Due Joachims (see his talk tomorrow!!)

Another Example: Self-consistency

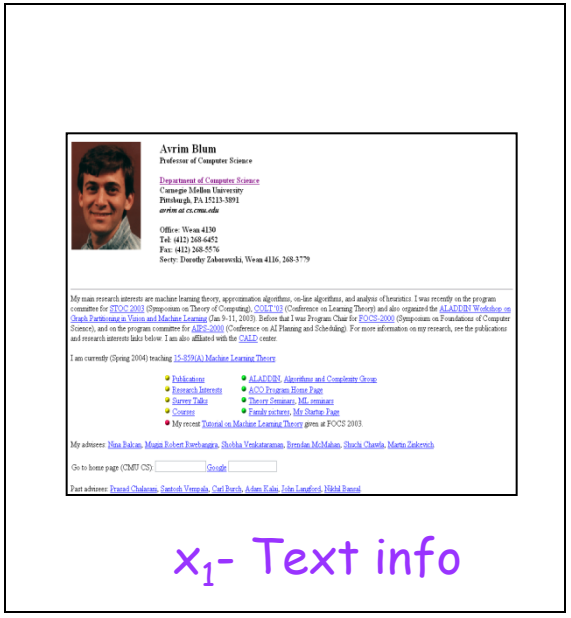
Agreement between two parts : co-training [Blum-Mitchell98].

- examples contain two sufficient sets of features, $x = \langle x_1, x_2 \rangle$
- belief: the parts are consistent, i.e. $\exists\ c_1, c_2$ s.t. $c_1(x_1)=c_2(x_2)=c^*(x)$

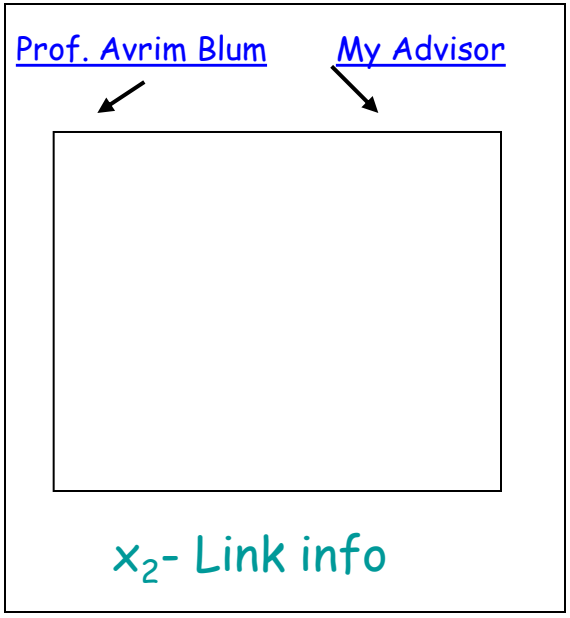
For example, if we want to classify web pages: $x = \langle x_1, x_2 \rangle$



x - Link info & Text info



x_1- Text info



x_2- Link info

New discriminative model for SSL

Problems with thinking about SSL in standard models

- PAC or SLT: learn a class C under (known or unknown) distribution D.
  - a complete disconnect between the target and D
- Unlabeled data doesn't give any info about which $c \in C$ is the target.

Key Insight

Unlabeled data useful if we have beliefs not only about the form of the target, but also about its relationship with the underlying distribution.

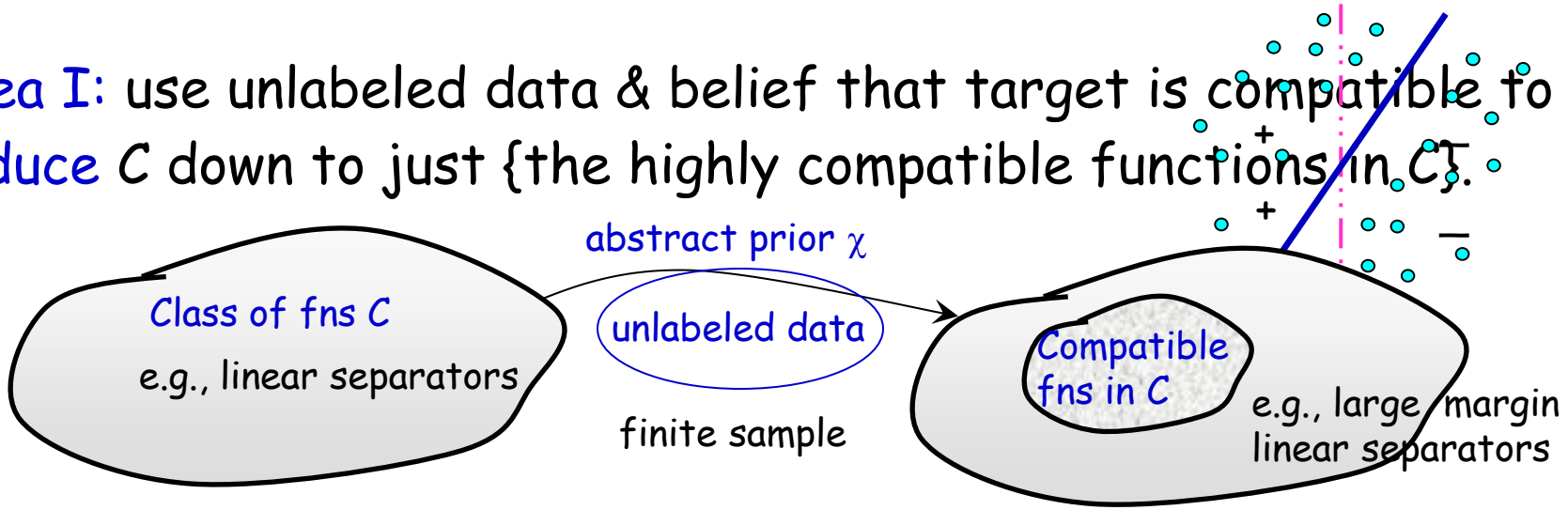
BB Model, Main Ideas

Augment the notion of a concept class C with a notion of compatibility χ between a concept and the data distribution.

"learn C" becomes "learn (C,χ)" (learn class C under χ)

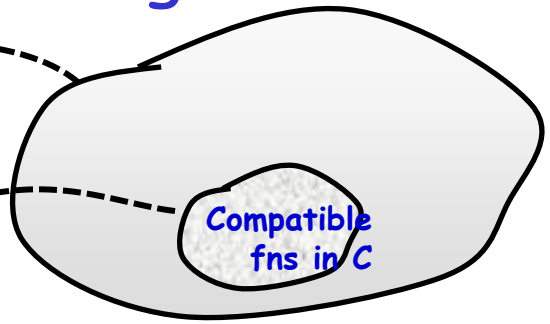
Express relationships that target and underlying distr. possess.

Idea I: use unlabeled data & belief that target is compatible to reduce C down to just {the highly compatible functions in C}.



Idea II: degree of compatibility estimated from a finite sample.

Sample Complexity, Uniform Convergence Bounds



If we see

$$m_u \geq \frac{1}{\varepsilon}\left[\ln|C| + \ln\frac{2}{\delta}\right]$$

unlabeled examples and

$$m_l \geq \frac{1}{\varepsilon}\left[\ln|C_{D,\chi}(\varepsilon)| + \ln\frac{2}{\delta}\right]$$

labeled examples, then with prob. $\geq 1-\delta$, all $h \in C$ with $\hat{err}(h) = 0$ and compatible with the sample have $err(h) \leq \varepsilon$.

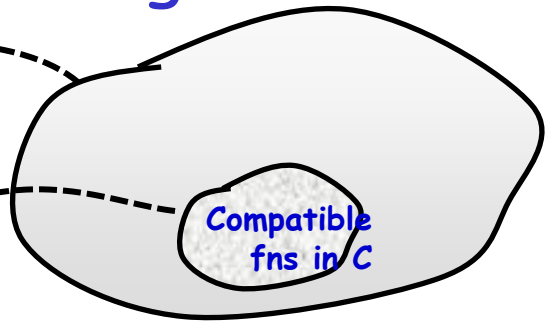
Bound # of labeled examples as a measure of the helpfulness of D wrt χ

- helpful D is one in which $C_{D,\chi}(\varepsilon)$ is small

Sample Complexity, Uniform Convergence Bounds

If we see

$$m_u \geq \frac{1}{\varepsilon}\left[\ln|C| + \ln\frac{2}{\delta}\right]$$

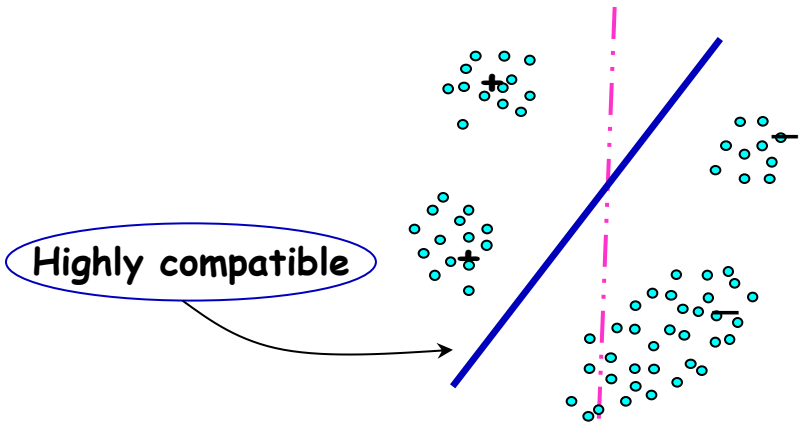


unlabeled examples and

$$m_l \geq \frac{1}{\varepsilon}\left[\ln|C_{D,\chi}(\varepsilon)| + \ln\frac{2}{\delta}\right]$$

labeled examples, then with prob. $\geq 1-\delta$, all $h \in C$ with $\hat{err}(h) = 0$ and compatible with the sample have $err(h) \leq \varepsilon$.

Helpful distribution



Non-helpful distribution

Key Aspects of the Model

Fundamental sample complexity aspects.

- How much unlabeled data is needed
  - depends both complexity of C and of the compatibility notion.
- Ability of unlabeled data to reduce # of labeled examples
  - compatibility of the target, helpfulness of the distribution

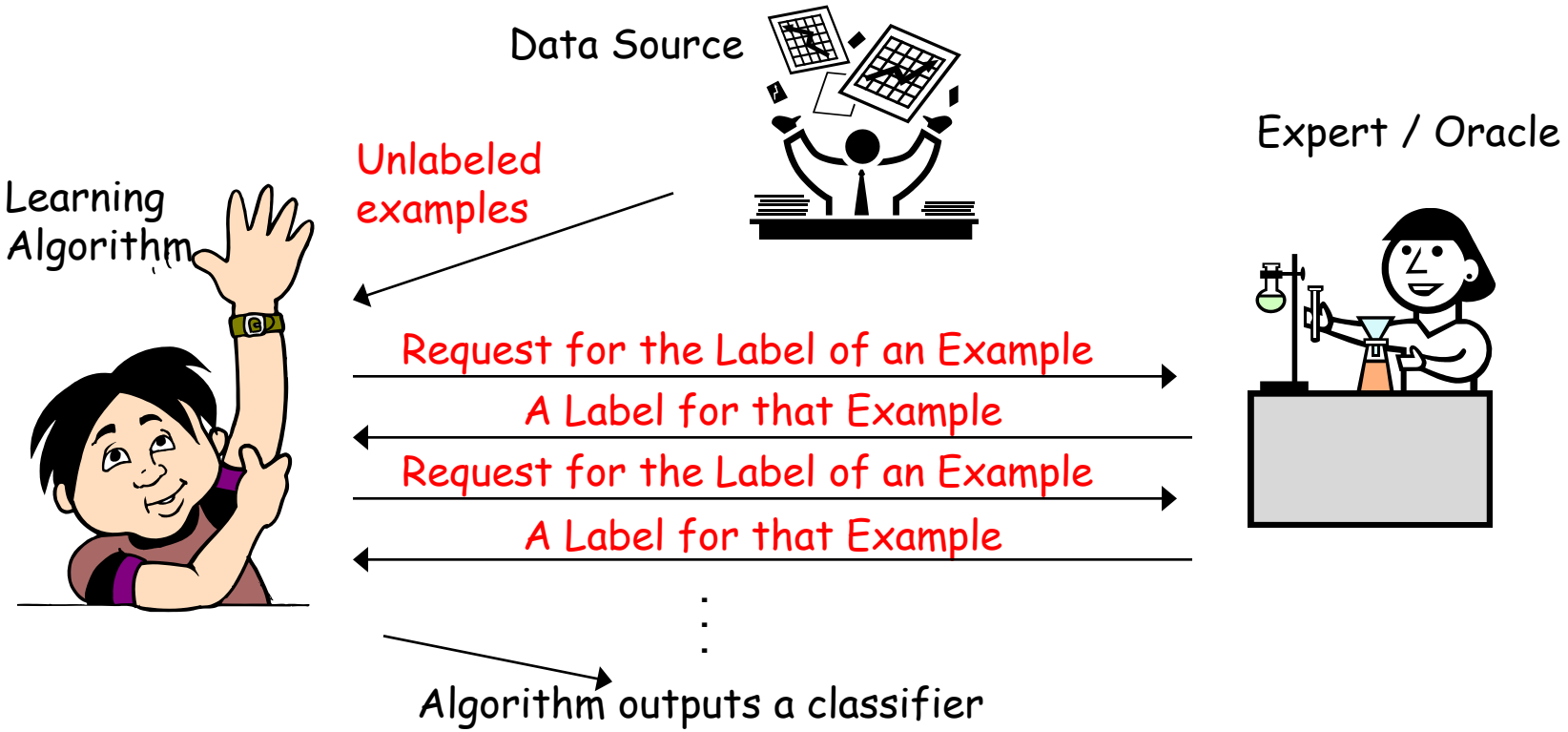
This analysis suggests better ways to do regularization based on unlabeled data.

Subsequent work using our framework

P. Bartlett, D. Rosenberg, AISTATS 2007; Kakade et al, COLT 2008

J. Shawe-Taylor et al., Neurocomputing 2007; Zhu, survey 2010

Active Learning



- The learner can choose specific examples to be labeled.
- He works harder, to use fewer labeled examples.

What Makes a Good Algorithm?

- Guaranteed to output a relatively good classifier for most learning problems.

- Doesn't make too many label requests.

Choose the label requests carefully, to get **informative** labels.

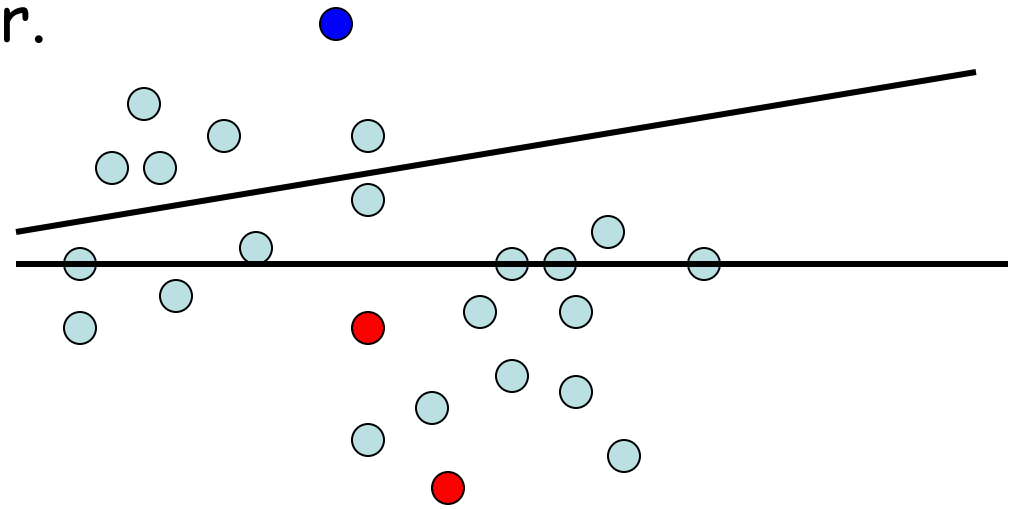
Can It Really Do Better Than Passive?

- YES! (sometimes)
- We often need far fewer labels for active learning than for passive.
- This is predicted by theory and has been observed in practice.

Active Learning in Practice

- Active SVM (Tong & Koller, ICML 2000) seems to be quite useful in practice.

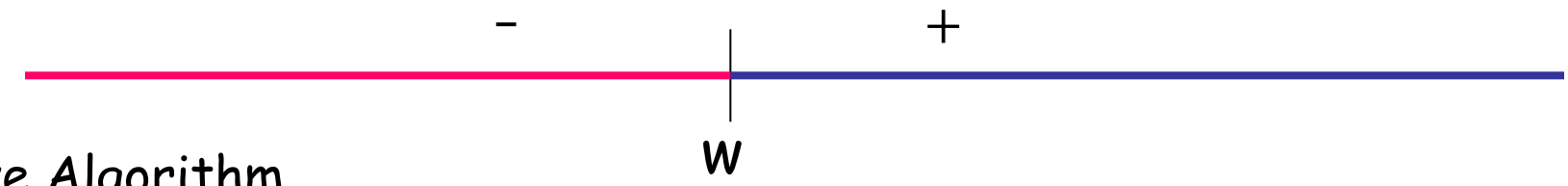
  At any time during the alg., we have a "current guess" of the separator: the max-margin separator of all labeled points so far.



E.g., strategy 1: request the label of the example closest to the current separator.

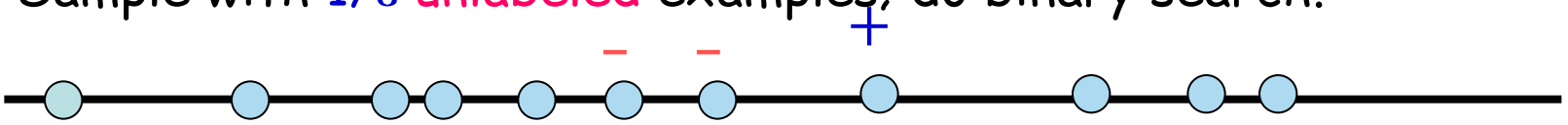
Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line: $h_w(x) = 1(x \geq w)$, $C = \{h_w : w \in R\}$



Active Algorithm

- Sample with $1/\varepsilon$ unlabeled examples; do binary search.



- Binary search – need just $O(\log 1/\varepsilon)$ labels.

Passive supervised: $\Omega(1/\varepsilon)$ labels to find an ε-accurate threshold.

Active: only $O(\log 1/\varepsilon)$ labels. Exponential improvement.

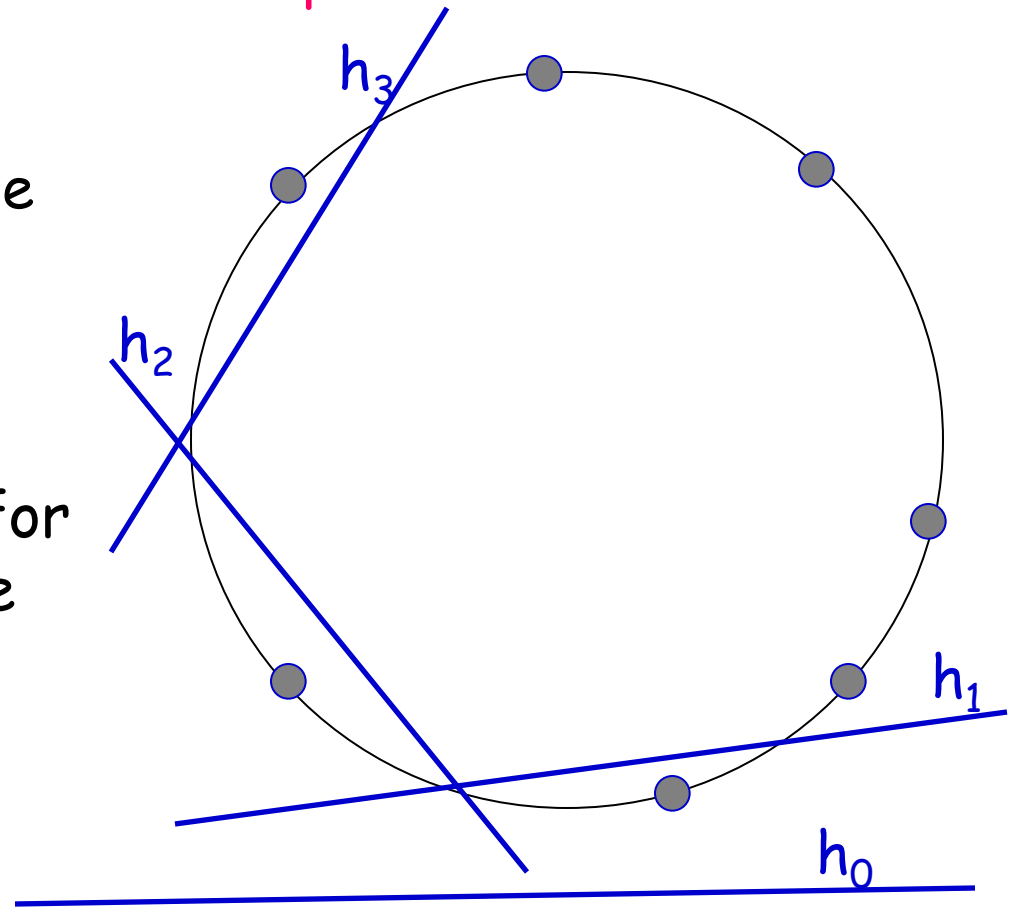
Other interesting results as well.

Active Learning might not help [Dasgupta04]

In general, number of queries needed depends on C and also on D.

C = {linear separators in R^1}:
active learning reduces sample complexity substantially.

C = {linear separators in R^2}:
there are some target hyp. for which no improvement can be achieved!
- no matter how benign the input distr.



In this case: learning to accuracy ε requires $1/\varepsilon$ labels…

Examples where Active Learning helps

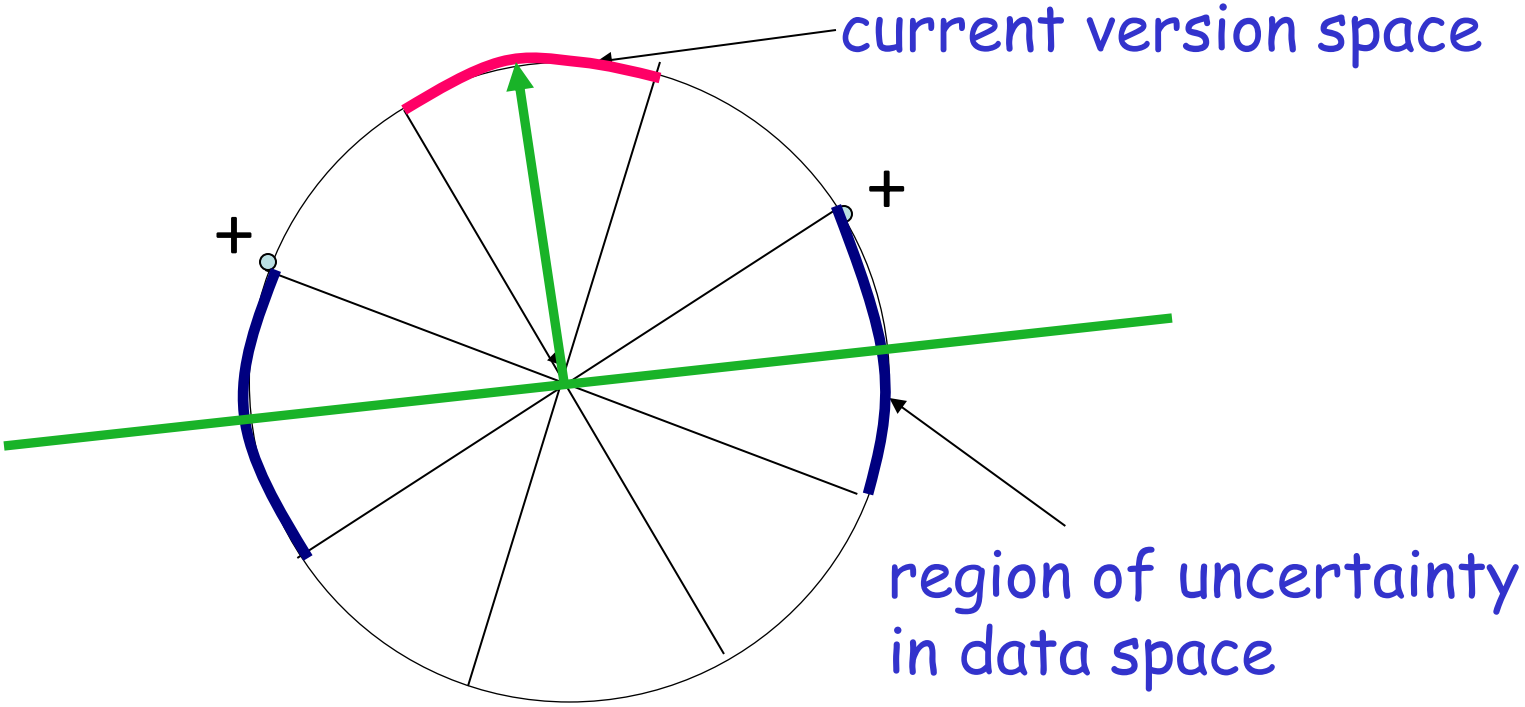
In general, number of queries needed depends on C and also on D.

- C = {linear separators in R^1}: active learning reduces sample complexity substantially no matter what is the input distribution.
- C - homogeneous linear separators in R^d, D - uniform distribution over unit sphere:
  - need only d log $1/\varepsilon$ labels to find a hypothesis with error rate < ε.
    - Dasgupta, Kalai, Monteleoni, COLT 2005
    - Freund et al., '97.
    - Balcan-Broder-Zhang, COLT 07

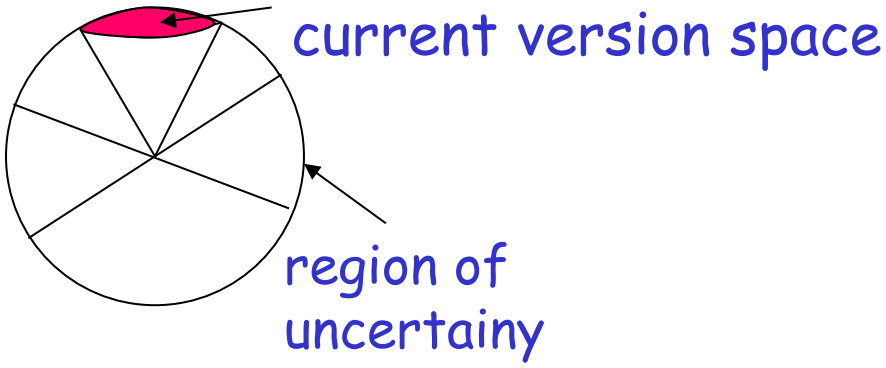
Region of uncertainty [CAL92]

- Current version space: part of C consistent with labels so far.
- "Region of uncertainty" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

- Example: data lies on circle in R^2 and hypotheses are homogeneous linear separators.



Region of uncertainty [CAL92]

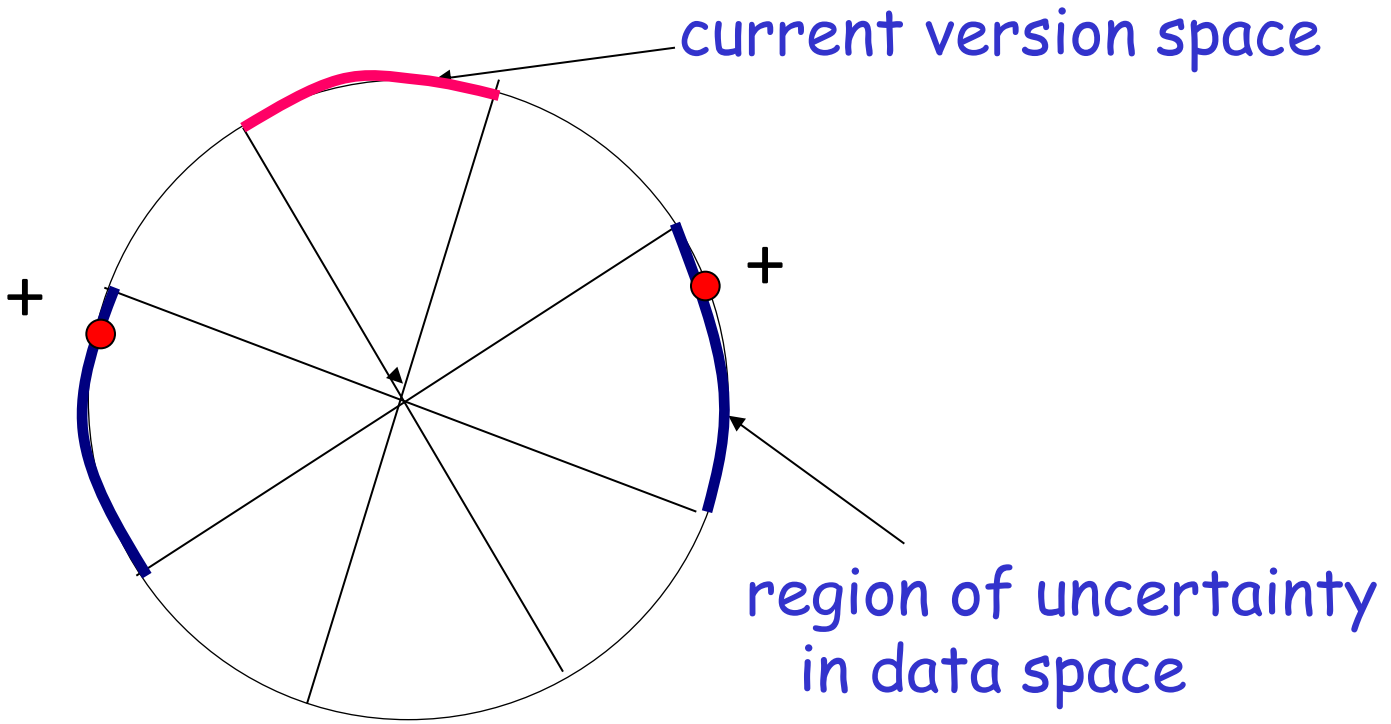


Algorithm:

Pick a few points at random from the current region of uncertainty and query their labels.

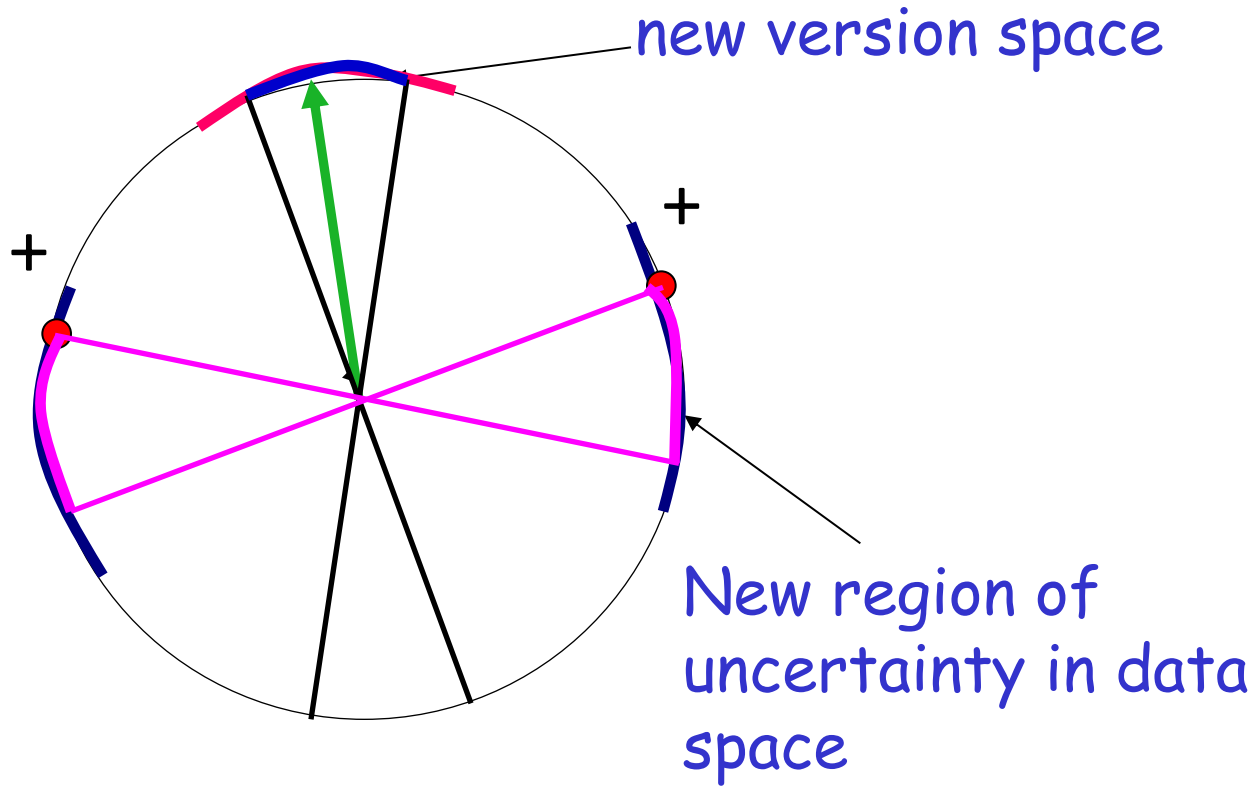
Region of uncertainty [CAL92]

- Current version space: part of C consistent with labels so far.
- "Region of uncertainty" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)



Region of uncertainty [CAL92]

- Current version space: part of C consistent with labels so far.
- "Region of uncertainty" = part of data space about which there is still some uncertainty (i.e. disagreement within version space)



Region of uncertainty [CAL92], Guarantees

Algorithm: Pick a few points at random from the current region of uncertainty and query their labels.

[Balcan, Beygelzimer, Langford, ICML'06]

Analyze a version of this alg. which is robust to noise.

- C- linear separators on the line, low noise, exponential improvement.
- C - homogeneous linear separators in R^d, D -uniform distribution over unit sphere.
  - low noise, need only $d^2 \log 1/\varepsilon$ labels to find a hypothesis with error rate < ε.
  - realizable case, $d^{3/2} \log 1/\varepsilon$ labels.
    - supervised -- d/ε labels.