Semi-Supervised Learning

Maria-Florina Balcan

11/04/2013

Supervised Learning: Formalization (PAC)

- X - instance space
- $S_l=\{(x_i, y_i)\}$ - labeled examples drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in \{-1,1\}$ - binary classification

- Algorithm A PAC-learns concept class C if for any target c^* in C, any distrib. D over X, any $\varepsilon, \delta > 0$:

  - A uses at most poly(n,$1/\varepsilon$,$1/\delta$,size(c^*)) examples and running time.
  - With probab. $1-\delta$, A produces h in C of error at $\leq \varepsilon$.

Supervised Learning, Big Questions

- Algorithm Design
  - How might we automatically generate rules that do well on observed data?

- Sample Complexity/Confidence Bound
  - What kind of confidence do we have that they will do well in the future?

Sample Complexity: Uniform Convergence

Finite Hypothesis Spaces

Realizable Case

Theorem After

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

examples, with probab. $1-\delta$, all $h \in C$ with $err(h) \geq \varepsilon$ have $e\hat{r}r(h) > 0$.

Agnostic Case

- What if there is no perfect h?

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in C$ have $|err(h) - e\hat{r}r(h)| < \varepsilon$, for

$$m_l \geq \frac{2}{\varepsilon^2}\left[\ln(|C|) + \ln\left(\frac{2}{\delta}\right)\right]$$

Sample Complexity: Uniform Convergence

Infinite Hypothesis Spaces

- C[S] – the set of splittings of dataset S using concepts from C.
- C[m] - maximum number of ways to split m points using concepts in C; i.e. $C[m] = \max_{|S|=m} |C[S]|$

- C[m,D] - expected number of splits of m points from D with concepts in C.

- Fact #1: previous results still hold if we replace |C| with C[2m].
- Fact #2: can even replace with C[2m,D].

Sample Complexity: Uniform Convergence
Infinite Hypothesis Spaces

For instance:

Theorem For any class C, distrib. D, if the number of labeled examples seen m_l satisfies

$$m_l \geq \frac{2}{\varepsilon}\left[\log_2(2C[2m]) + \log_2\left(\frac{1}{\delta}\right)\right]$$

then with probab. $1-\delta$, all $h \in C$ with $err(h) \geq \varepsilon$ have $\hat{err}(h) > 0$.

Sauer's Lemma, C[m]=O($m^{VC\text{-}dim(C)}$) implies:

Theorem

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

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

Sample Complexity: ε-Cover Bounds

- C_ε is an ε-cover for C w.r.t. D if for every $h \in C$ there is a $h' \in C_\varepsilon$ which is ε-close to h.

- To learn, it's enough to find an ε-cover and then do empirical risk minimization w.r.t. the functions in this cover.

- In principle, in the realizable case, the number of labeled examples we need is

$$O\left(\frac{1}{\varepsilon}\left[\ln(|C_{\epsilon/4}|) + \ln\left(\frac{1}{\delta}\right)\right]\right)$$

Usually, for fixed distributions.

Sample Complexity: ε-Cover Bounds

Can be much better than Uniform-Convergence bounds!

Simple Example (Realizable case)

- $X=\{1, 2, \ldots, n\}$, $C = C_1 \cup C_2$, D= uniform over X.
- C_1 - the class of all functions that predict positive on at most $\epsilon \cdot n/4$ examples.
- C_2 - the class of all functions that predict negative on at most $\epsilon \cdot n/4$ examples.

If the number of labeled examples $m_l < \epsilon \cdot n/4$, don't have uniform convergence yet.

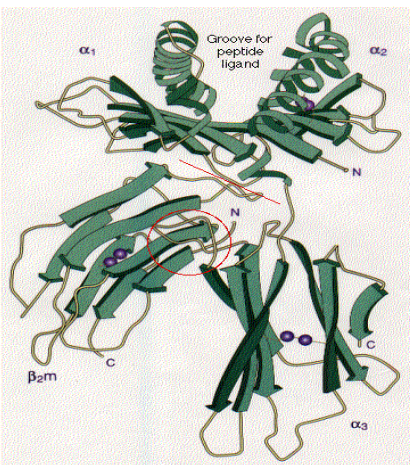
The size of the smallest $\epsilon/4$-cover is 2, so we can learn with only $O(1/\epsilon)$ labeled examples.

In fact, since the elements of this cover are far apart, much fewer examples are sufficient.

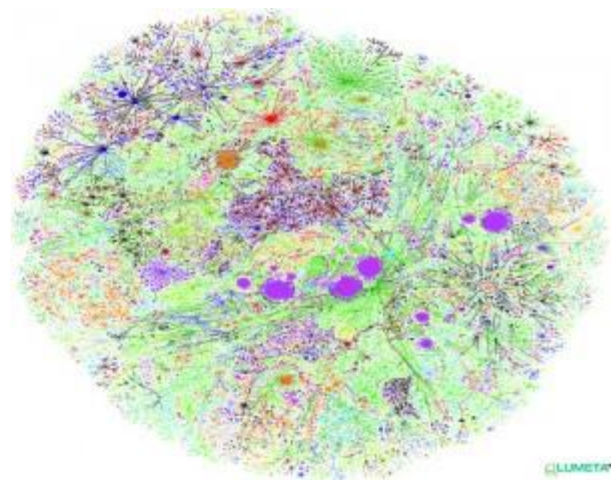
Classic Paradigm Insufficient Nowadays

Modern applications: **massive amounts** of raw data.

Only a tiny fraction can be annotated by human experts.



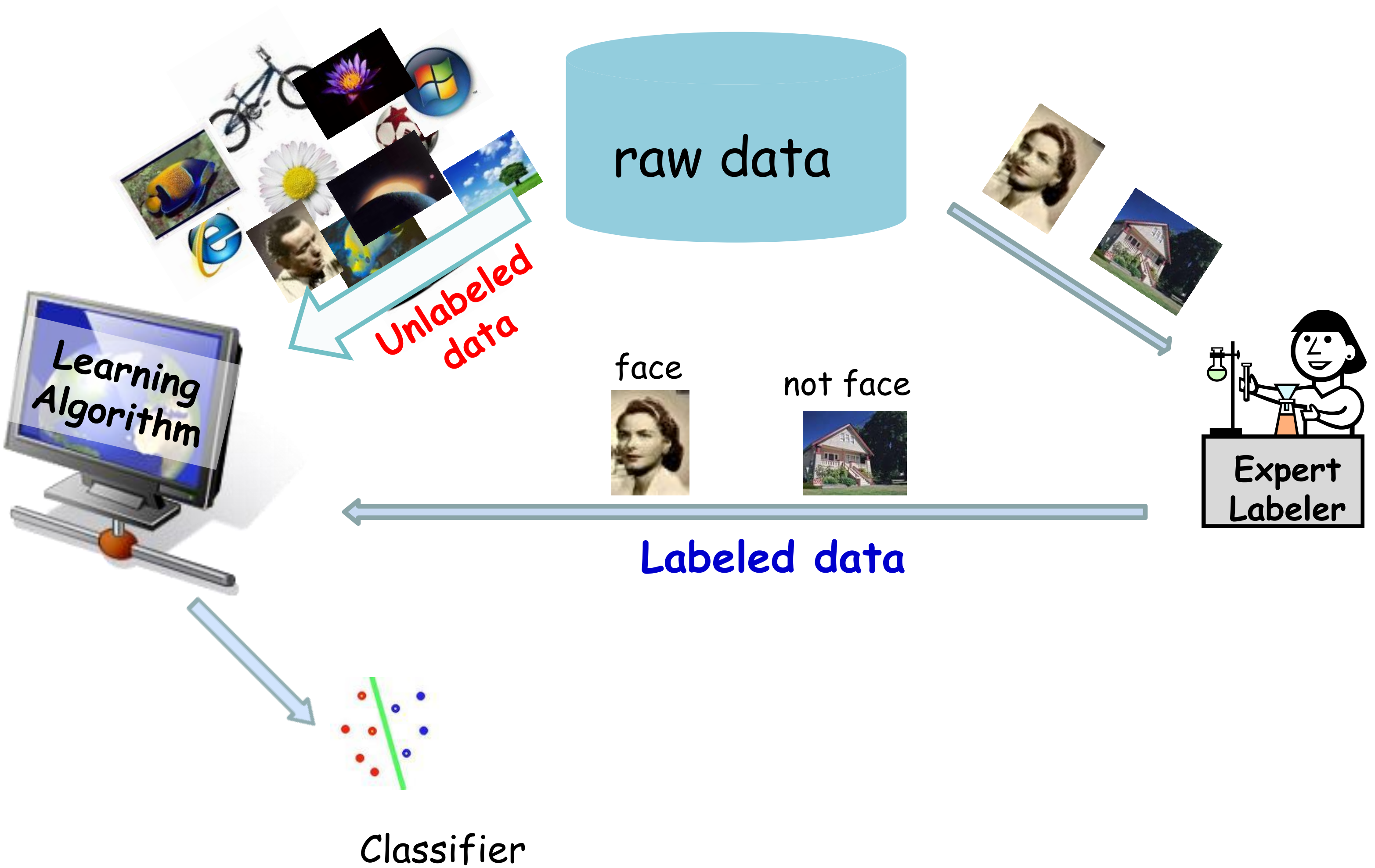
Protein sequences



Billions of webpages

Images

Semi-Supervised Learning



Semi-Supervised Learning

Hot topic in recent years in Machine Learning.

- Many applications have lots of unlabeled data, but labeled data is rare or expensive:
  - Web page, document classification
  - OCR, Image classification

Workshops [ICML '03, ICML' 05]

Books: Semi-Supervised Learning, MIT 2006

O. Chapelle, B. Scholkopf and A. Zien (eds)

Combining Labeled and Unlabeled Data

- Several methods have been developed to try to use unlabeled data to improve performance, e.g.:
  - Transductive SVM [Joachims '99]
  - Co-training [Blum & Mitchell '98], [BBY04]
  - Graph-based methods [Blum & Chawla01], [ZGL03]

- Augmented PAC model for SSL [Balcan & Blum '05]

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

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

Different model: the learner gets to pick the examples to Labeled – Active Learning.

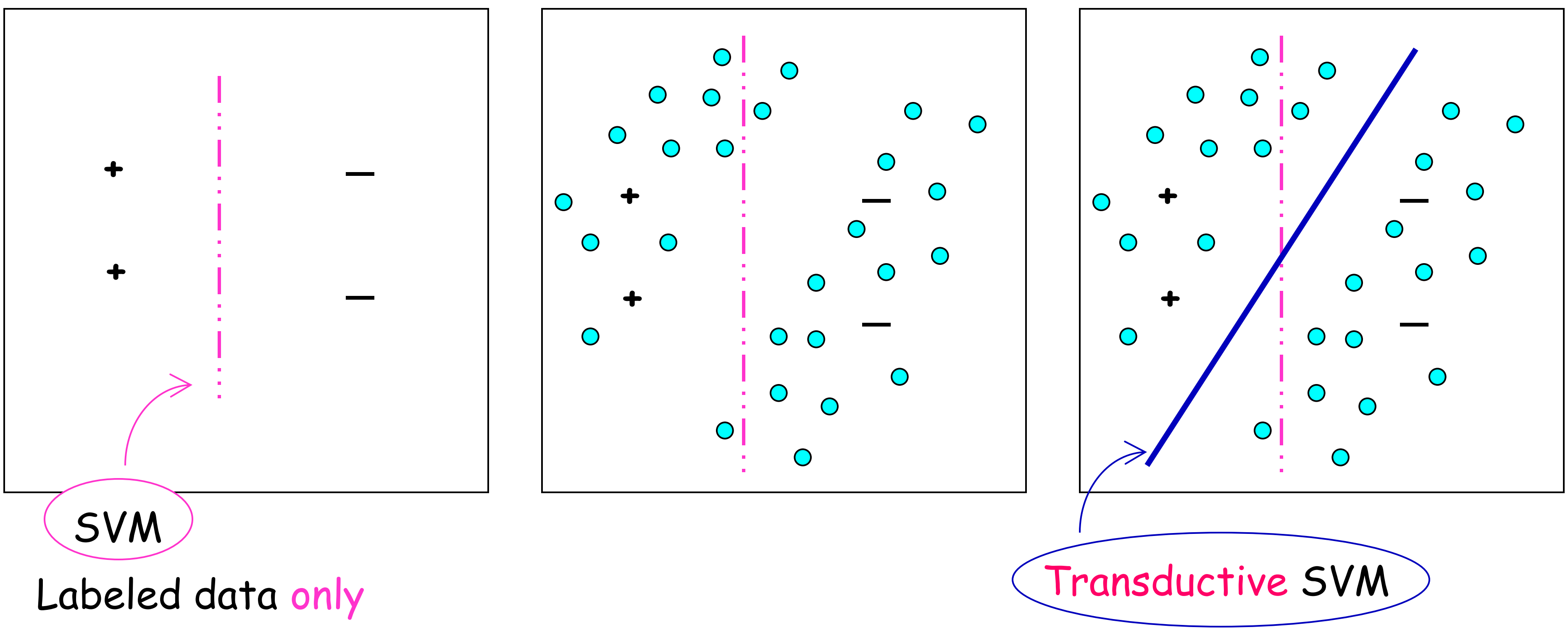
Can we extend the PAC/SLT models to deal with Unlabeled Data?

- PAC/SLT models – nice/standard models for learning from labeled data.

- Goal – extend them naturally to the case of learning from both labeled and unlabeled data.
  - Different algorithms are based on different assumptions about how data should behave.
  - Question – how to capture many of the assumptions typically used?

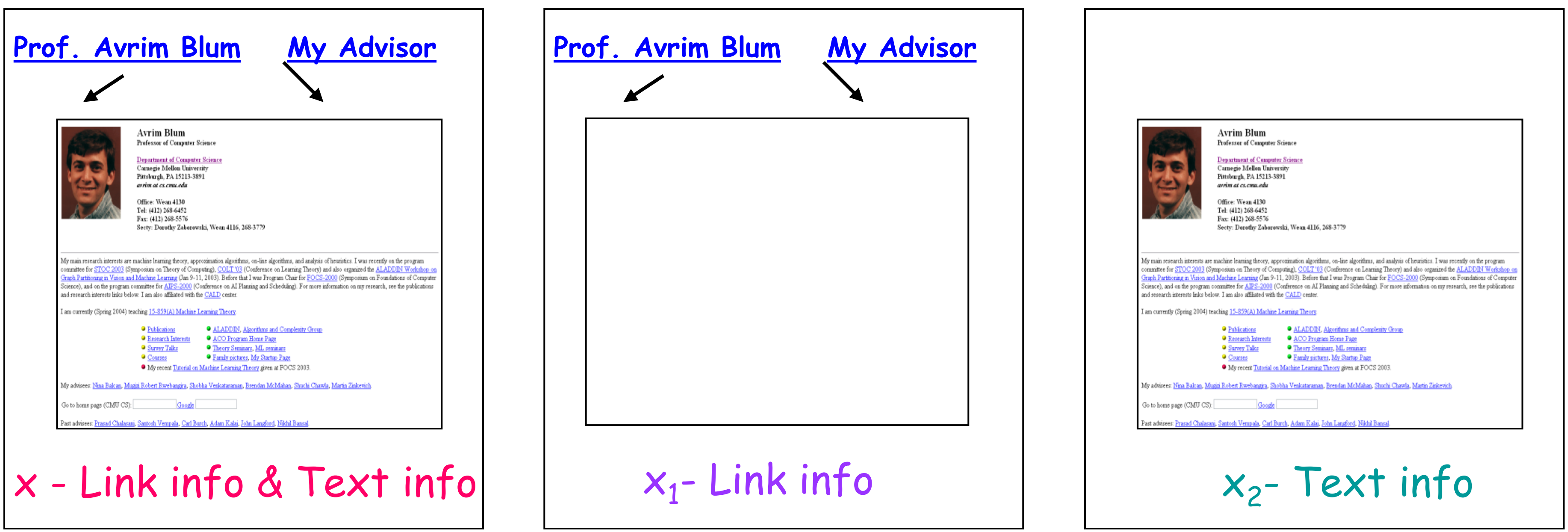
Example of "typical" assumption: Margins

- The separator goes through low density regions of the space/large margin.
  - assume we are looking for linear separator
  - belief: should exist one with large separation

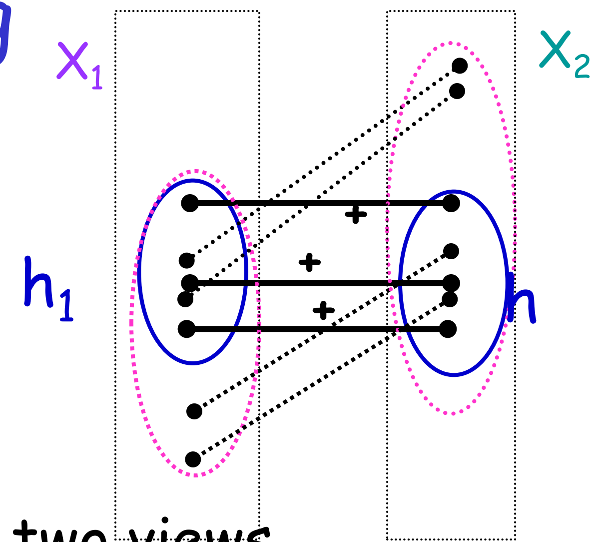


Another Example: Self-consistency

- Agreement between two parts : co-training.
  - examples contain two sufficient sets of features, i.e. an example is $x = \langle x_1, x_2 \rangle$ and the belief is that the two parts of the example are consistent, i.e. $\exists\, c_1, c_2$ such that $c_1(x_1)=c_2(x_2)=c^*(x)$
  - for example, if we want to classify web pages: $x = \langle x_1, x_2 \rangle$



Iterative Co-Training



Works by using unlabeled data to propagate learned information.

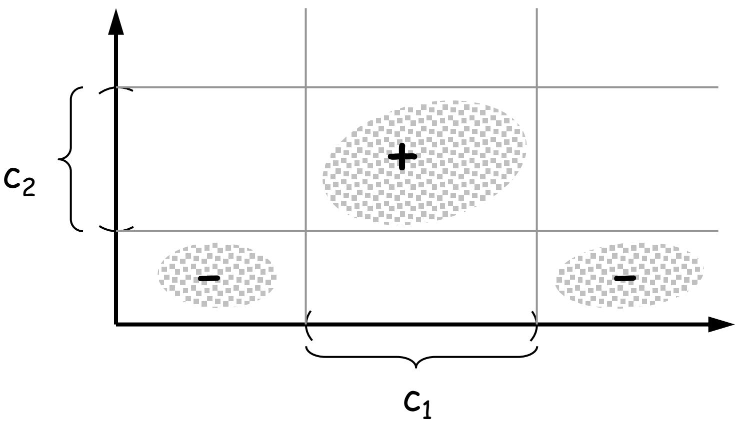
- Have learning algos A_1, A_2 on each of the two views.
- Use labeled data to learn two initial hyp. h_1, h_2.

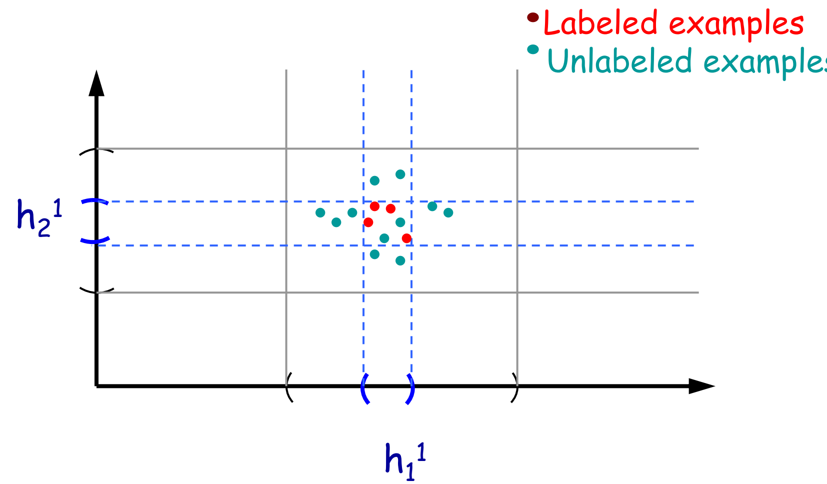
Repeat

- Look through unlabeled data to find examples where one of h_i is confident but other is not.
- Have the confident h_i label it for algorithm A_{3-i}.

Iterative Co-Training

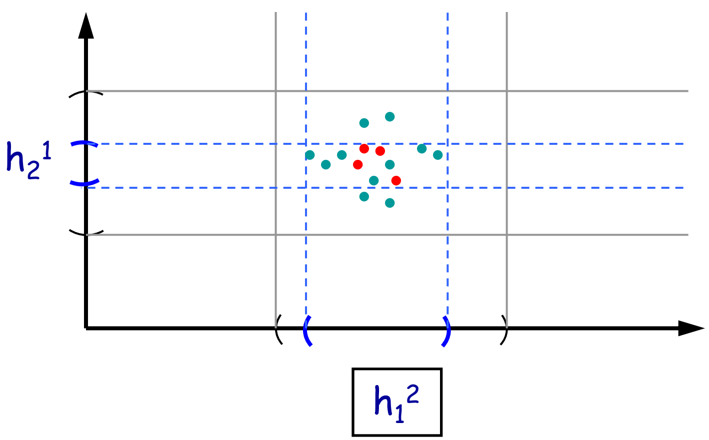
A Simple Example: Learning Intervals

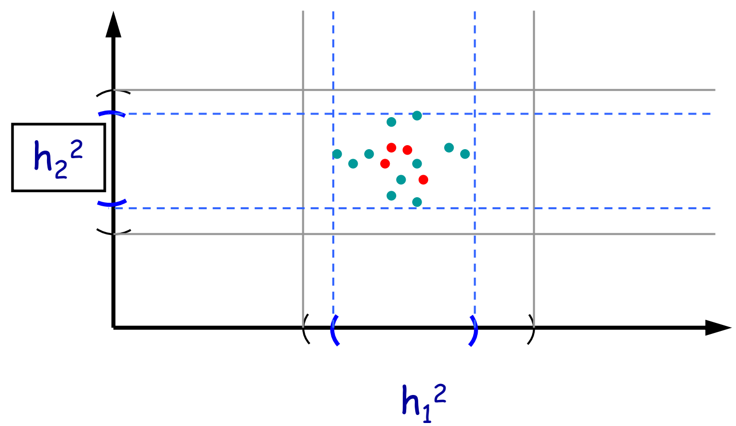




Use labeled data to learn h_1^1 and h_2^1

Use unlabeled data to bootstrap





Co-training: Theoretical Guarantees

- What properties do we need for co-training to work well?
- We need assumptions about:
  1. the underlying data distribution
  2. the learning algorithms on the two sides

[Blum & Mitchell, COLT '98]

1. Independence given the label
2. Alg. for learning from random noise.

[Balcan, Blum, Yang, NIPS 2004]

1. Distributional expansion.
2. Alg. for learning from positve data only.

Problems thinking about SSL in the PAC model

- PAC model talks of learning a class C under (known or unknown) distribution D.
  - Not clear what unlabeled data can do for you.
  - Doesn't give you any info about which $c \in C$ is the target function.

- Can we extend the PAC model to capture these (and more) uses of unlabeled data?

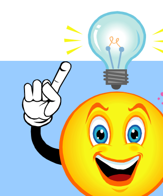
  - Give a unified framework for understanding when and why unlabeled data can help.

New discriminative model for SSL

$S_u=\{x_i\}$ - x_i i.i.d. from D and $S_l=\{(x_i, y_i)\}$ -x_i i.i.d. from D, $y_i = c^*(x_i)$.

Problems with thinking about SSL in standard WC 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.

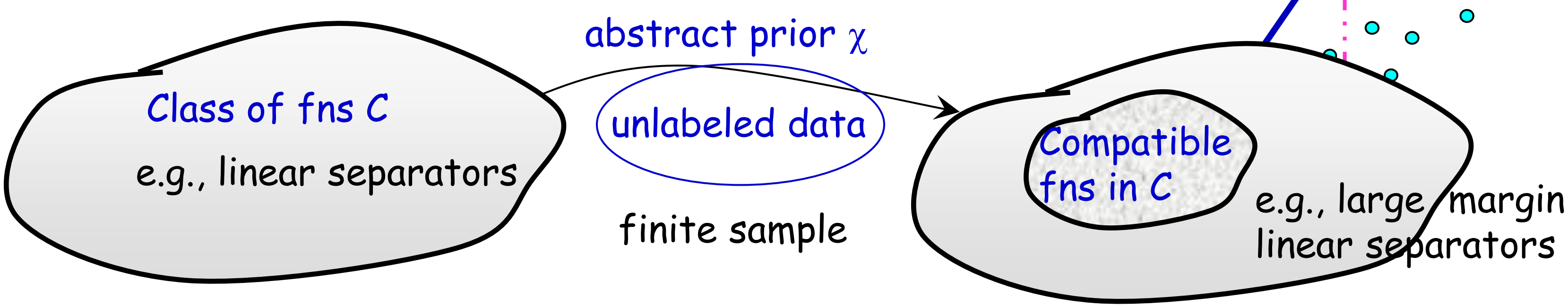
New model for SSL, Main Ideas

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.

Formally

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

Require compatibility $\chi(h,D)$ to be expectation over individual examples. (don't need to be so strict but this is cleanest)

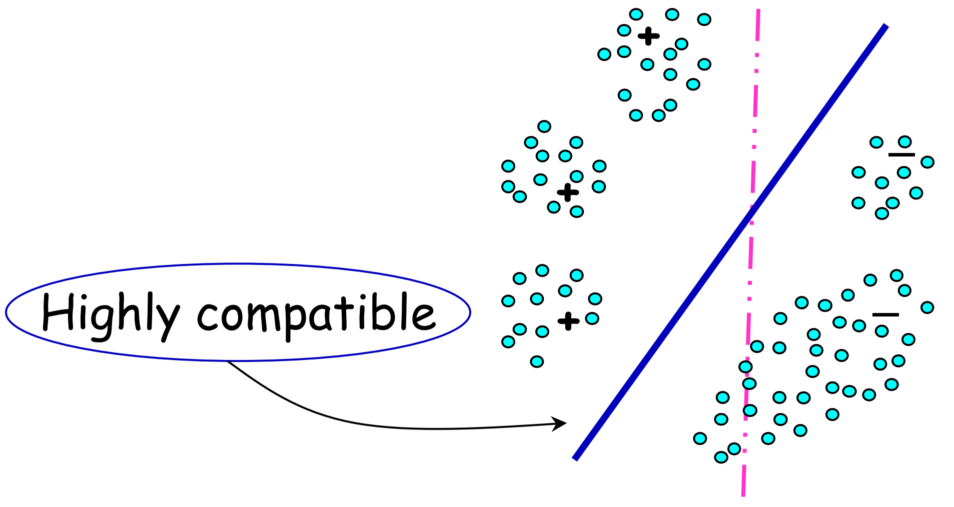
$\chi(h,D)=E_{x\in D}[\chi(h, x)]$ compatibility of h with D, $\chi(h,x)\in [0,1]$

View *in*compatibility as unlabeled error rate

$err_{unl}(h)=1-\chi(h, D)$ incompatibility of h with D

Margins, Compatibility

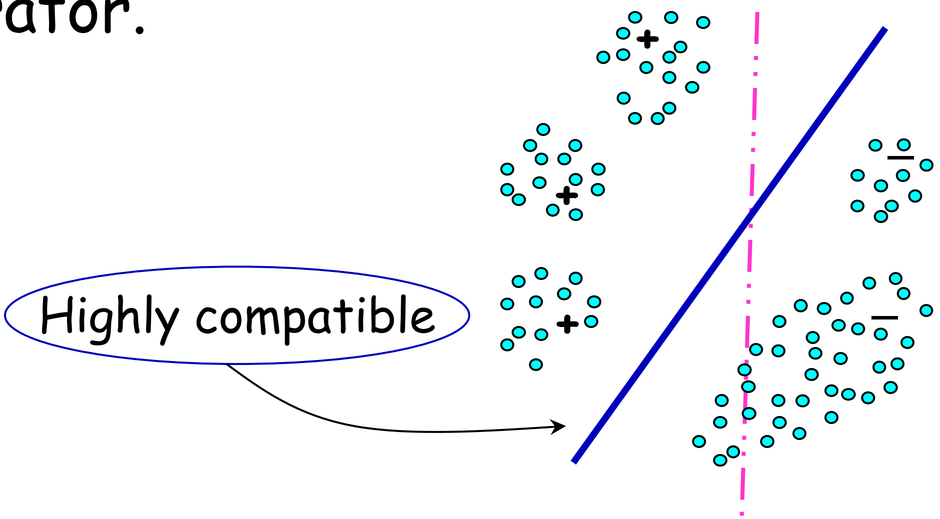
- Margins: belief is that should exist a large margin separator.



- Incompatibility of h and D (unlabeled error rate of h) – the probability mass within distance γ of h.
- Can be written as an expectation over individual examples $\chi(h,D)=E_{x \in D}[\chi(h,x)]$ where:
    - $\chi(h,x)=0$ if $dist(x,h) \leq \gamma$
    - $\chi(h,x)=1$ if $dist(x,h) \geq \gamma$

Margins, Compatibility

- Margins: belief is that should exist a large margin separator.



- If do not want to commit to γ in advance, define χ(h,x) to be a smooth function of dist(x,h), e.g.:

$$\chi(h,x) = 1 - e^{\left[-\frac{dist(x,h)}{2\sigma^2}\right]}$$

- Illegal notion of compatibility: the largest γ s.t. D has probability mass exactly zero within distance γ of h.

Co-Training, Compatibility

- Co-training: examples come as pairs $\langle x_1, x_2 \rangle$ and the goal is to learn a pair of functions $\langle h_1, h_2 \rangle$
- Hope is that the two parts of the example are consistent.

- Legal (and natural) notion of compatibility:
  - the compatibility of $\langle h_1, h_2 \rangle$ and D:

$$\Pr_{\langle x_1, x_2 \rangle \in D}[h_1(x_1) = h_2(x_2)]$$

  - can be written as an expectation over examples:

$$\chi(\langle h_1, h_2 \rangle, \langle x_1, x_2 \rangle) = 1 \text{ if } h_1(x_1) = h_2(x_2)$$

$$\chi(\langle h_1, h_2 \rangle, \langle x_1, x_2 \rangle) = 0 \text{ if } h_1(x_1) \neq h_2(x_2)$$

Types of Results in the [BB05] Model

- As in the usual PAC model, can discuss algorithmic and sample complexity issues.

Sample Complexity issues that we can address:

- How much unlabeled data we need:
  - depends both on the complexity of C and the complexity of our notion of compatibility.
- Ability of unlabeled data to reduce number of labeled examples needed:
  - compatibility of the target
  - (various measures of) the helpfulness of the distribution

- Give both uniform convergence bounds and epsilon-cover based bounds.

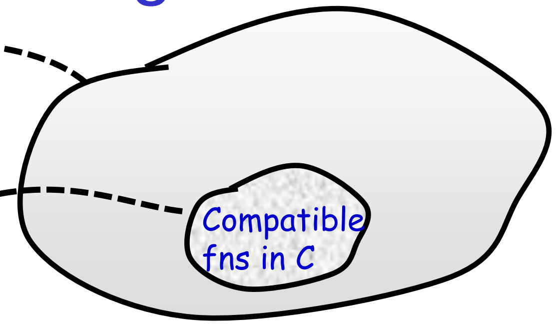
Sample Complexity, Uniform Convergence Bounds

If we see

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

unlabeled examples and

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



$C_{D,\chi}(\varepsilon) = \{h \in C : err_{unl}(h) \leq \varepsilon\}$

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

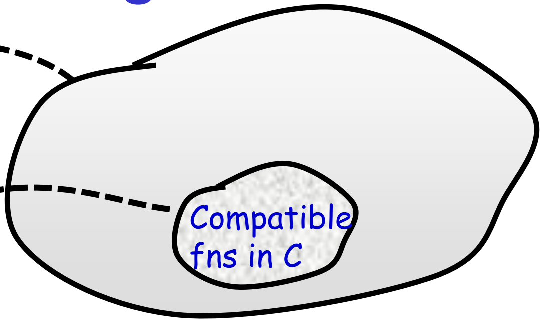
Proof

Probability that h with $err_{unl}(h) > \epsilon$ is compatible with S_u is $(1-\epsilon)^{m_u} \leq \delta/(2|C|)$

By union bound, prob. $1-\delta/2$ only hyp in $C_{D,\chi}(\varepsilon)$ are compatible with S_u

m_l large enough to ensure that none of fns in $C_{D,\chi}(\varepsilon)$ with $err(h) \geq \epsilon$ have an empirical error rate of 0.

Sample Complexity, Uniform Convergence Bounds



$C_{D,\chi}(\varepsilon) = \{h \in C : err_{unl}(h) \leq \varepsilon\}$

If we see

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

unlabeled examples and

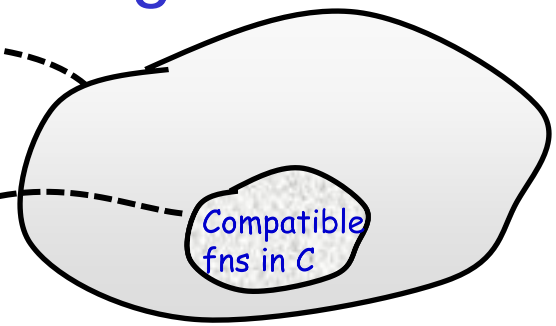
$$m_l \geq \frac{1}{\varepsilon}\left[\ln|C_{D,\chi}(\varepsilon)| + \ln\frac{2}{\delta}\right]$$

labeled examples, then with prob. $\geq 1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and $\widehat{err}_{unl}(h) = 0$ (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

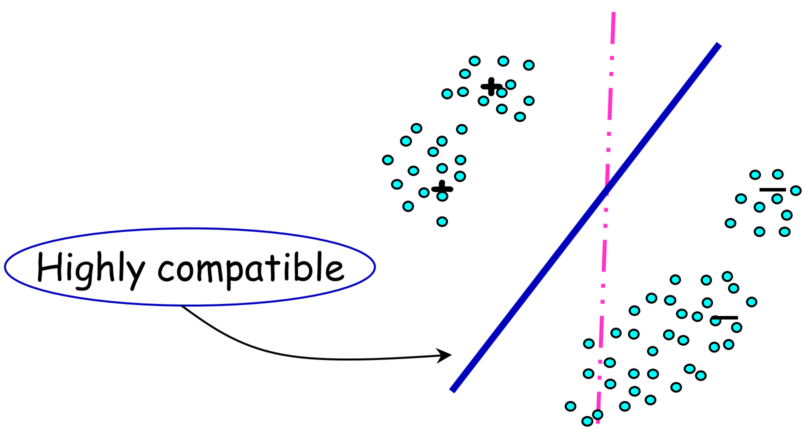
$$m_u \geq \frac{1}{\varepsilon}\left[\ln|C| + \ln\frac{2}{\delta}\right]$$

unlabeled examples and

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

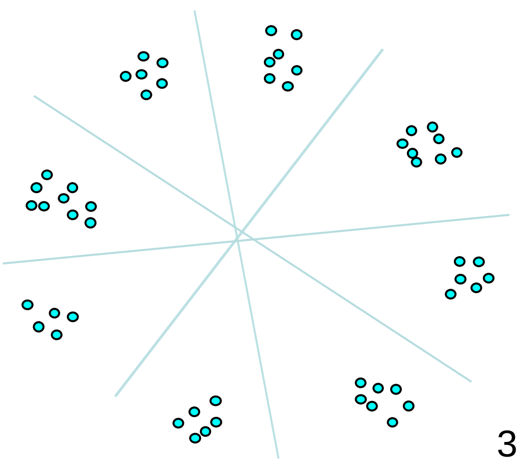
labeled examples, then with prob. $\geq 1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and compatible with the sample have $err(h) \leq \varepsilon$.

Helpful distribution



Non-helpful distribution

$1/\gamma^2$ clusters, all partitions separable by large margin



Examples of results: Sample Complexity - Uniform convergence bounds

Finite Hypothesis Spaces – c* not fully compatible:

Theorem

Given $t \in [0,1]$, if we see

$$m_u \geq \frac{2}{\varepsilon^2}\left[\ln|C| + \ln\frac{4}{\delta}\right]$$

unlabeled examples and

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

labeled examples, then with prob. $\geq 1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and $\widehat{err}_{unl}(h) \leq t+\varepsilon$ have $err(h) \leq \varepsilon$; furthermore all $h \in C$ with $err_{unl}(h) \leq t$ have $\widehat{err}_{unl}(h) \leq t+\varepsilon$.

Implication If $err_{unl}(c^*) \leq t$ and $err(c^*) = 0$ then with probability $\geq 1-\delta$ the $h \in C$ that optimizes $\widehat{err}(h)$ and $\widehat{err}_{unl}(h)$ has $err(h) \leq \epsilon$.

Examples of results: Sample Complexity - Uniform convergence bounds

Infinite Hypothesis Spaces

Assume $\chi(h,x) \in \{0,1\}$ and $\chi(C) = \{\chi_h : h \in C\}$ where $\chi_h(x) = \chi(h,x)$.

C[m,D] - expected # of splits of m points from D with concepts in C.

Theorem

$$m_u = O\left(\frac{VCdim\left(\chi(C)\right)}{\varepsilon^2}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{2}{\delta}\right)$$

unlabeled examples and

$$m_l > \frac{2}{\varepsilon}\left[\log(2s) + \log\frac{2}{\delta}\right]$$

labeled examples, where

$$s = C_{D,\chi}(t + 2\varepsilon)[2m_l, D]$$

are sufficient so that with probability at least $1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and $\widehat{err}_{unl}(h) \leq t + \varepsilon$ have $err(h) \leq \varepsilon$; furthermore all $h \in C$ have

$$|err_{unl}(h) - \widehat{err}_{unl}(h)| \leq \varepsilon$$

Implication: If $err_{unl}(c^*) \leq t$, then with probab. $\geq 1-\delta$, the $h \in C$ that optimizes both $\widehat{err}(h)$ and $\widehat{err}_{unl}(h)$ has $err(h) \leq \varepsilon$.

Examples of results: Sample Complexity - Uniform convergence bounds

- For $S \subseteq X$, denote by U_S the uniform distribution over S, and by $C[m, U_S]$ the expected number of splits of m points from U_S with concepts in C.
- Assume $err(c^*)=0$ and $err_{unl}(c^*)=0$.
- Theorem

An unlabeled sample S of size

$$O\left(\frac{\max[VCdim(C), VCdim(\chi(C))]}{\epsilon^2} log\frac{1}{\epsilon} + \frac{1}{\epsilon^2} log\frac{2}{\delta}\right)$$

is sufficient so that if we label m_l examples drawn uniformly at random from S, where

$$m_l > \frac{4}{\epsilon}\left[\log(2s) + \log\frac{2}{\delta}\right] \quad \text{and} \quad s = C_{S,\chi}(0)[2m_l, U_S]$$

then with probability $\geq 1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and $\widehat{err}_{unl}(h) = 0$ have $err(h) \leq \epsilon$.

- The number of labeled examples depends on the unlabeled sample.
- Useful since can imagine the learning alg. performing some calculations over the unlabeled data and then deciding how many labeled examples to purchase.

Sample Complexity Subtleties

Uniform Convergence Bounds

Depends both on the complexity of C and on the complexity of χ

Theorem

$$m_u = O\left(\frac{VCdim\left(\chi(C)\right)}{\varepsilon^2}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{2}{\delta}\right)$$

unlabeled examples and

Distr. dependent measure of complexity

$$m_l > \frac{2}{\varepsilon}\left[\log(2s) + \log\frac{2}{\delta}\right]$$

labeled examples, where

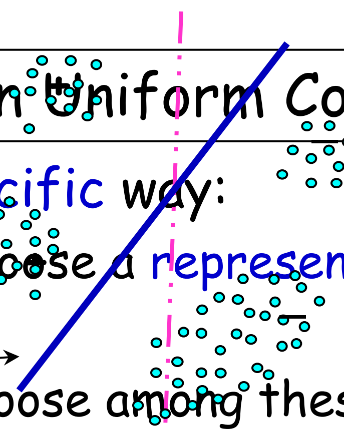
$$s = C_{D,\chi}(t + 2\varepsilon)[2m_l, D]$$

are sufficient s. t. with probab. $1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ and $\widehat{err}_{unl}(h) \leq t + \varepsilon$ have $err(h) \leq \varepsilon$.

ε-Cover bounds much better than Uniform Convergence bounds.

For algorithms that behave in a specific way:

- first use the un choose a representative set of compatible hypotheses
- then use the labeled sample to choose among these



Examples of results: Sample Complexity, ε-Cover-based bounds

- For algorithms that behave in a specific way:
  - first use the unlabeled data to choose a representative set of compatible hypotheses
  - then use the labeled sample to choose among these

Theorem

If t is an upper bound for $err_{unl}(c^*)$ and p is the size of a minimum ε − cover for $C_{D,\chi}(t+4\varepsilon)$, then using

$$m_u = O\left(\frac{VCdim\left(\chi(C)\right)}{\varepsilon^2} log\frac{1}{\varepsilon} + \frac{1}{\varepsilon^2} log\frac{2}{\delta}\right)$$

unlabeled examples and

$$m_l = O\left(\frac{1}{\varepsilon}\ln\frac{p}{\delta}\right)$$

labeled examples, we can with probab. $1-\delta$ identify a hyp. which is 10ϵ close to c^*.

- Can result in much better bound than uniform convergence!

Implications of the [BB05] analysis

Ways in which unlabeled data can help

- If c^* is highly compatible with D and have enough unlabeled data to estimate χ over all $h \in C$, then can reduce the search space (from C down to just those $h \in C$ whose estimated unlabeled error rate is low).

- By providing an estimate of D, unlabeled data can allow a more refined distribution-specific notion of hypothesis space size (e.g., Annealed VC-entropy or the size of the smallest ε-cover).

- If D is nice so that the set of compatible $h \in C$ has a small ε-cover and the elements of the cover are far apart, then can learn from even fewer labeled examples than the $1/\varepsilon$ needed just to verify a good hypothesis.