Beyond Worst Case Analysis in Machine Learning

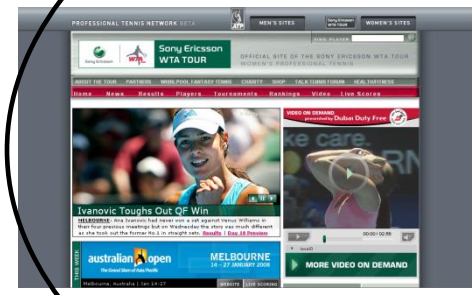
Maria-Florina Balcan

A PAC-style framework for Semi-Supervised Learning

Machine Learning

Image Classification

Document Categorization



Speech Recognition Protein Classification

Branch Prediction Fraud Detection Spam Detection

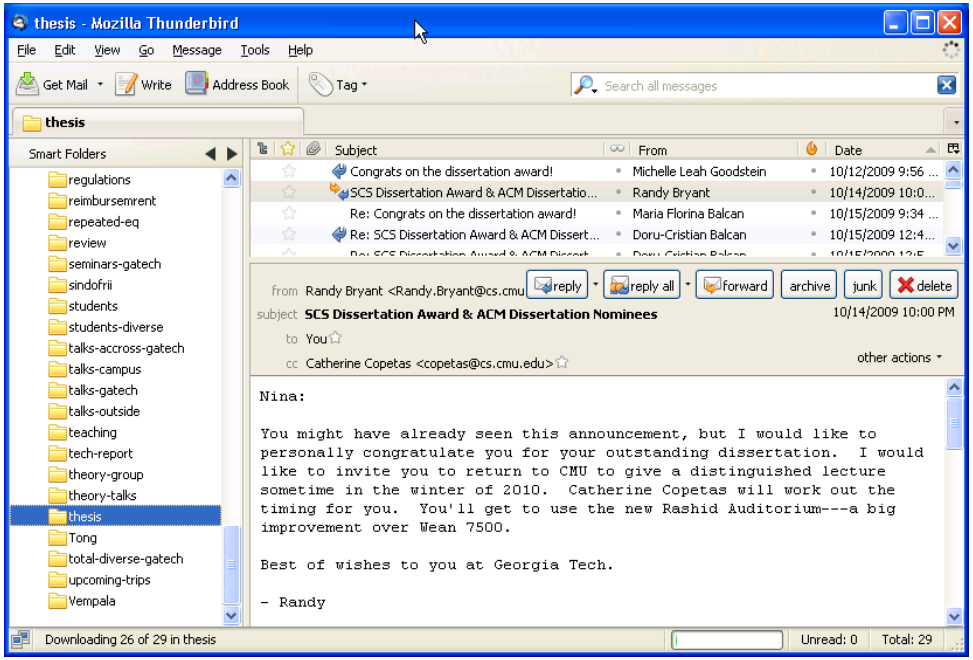
Supervised Classification

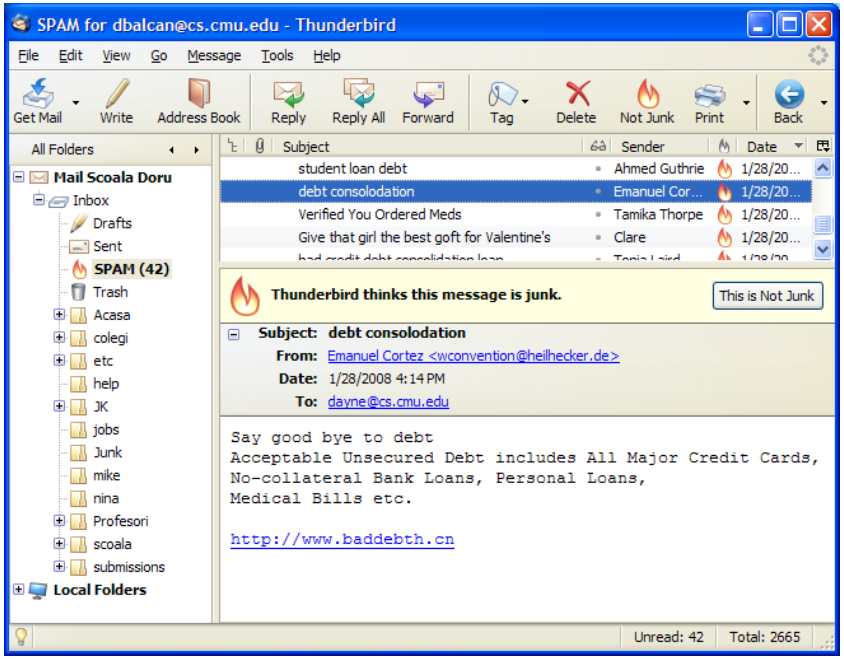
Example: Decide which emails are spam and which are important.

Supervised classification

Not spam

spam





Goal: use past emails to produce good rule for future data.

Example: Supervised Classification

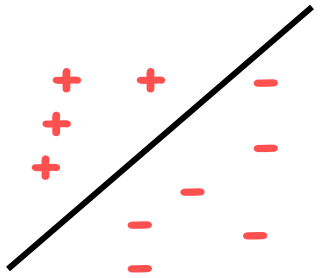
Represent each message by features. (e.g., keywords, spelling, etc.)

	"money"	"pills"	"Mr."	bad spelling	known-sender	spam?	
	Y	N	Y	Y	N	Y	
	N	N	N	Y	Y	N	
	N	Y	N	N	N	Y	
example	Y	N	N	N	Y	N	label
	N	N	Y	N	Y	N	
	Y	N	N	Y	N	Y	
	N	N	Y	N	N	N	

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

Predict SPAM if 2money + 3pills –5 known > 0



Linearly separable

Two Main Aspects in Machine Learning

Algorithm Design. How to optimize?

Automatically generate rules that do well on observed data.

Confidence Bounds, Generalization Guarantees

Confidence for rule effectiveness on future data.

Well understood for supervised learning.

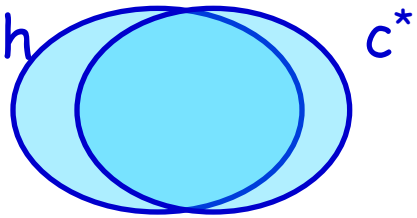
Supervised Learning Formalization

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

- 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

Sample Complexity Results

Theorem Finite C, realizable

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

labeled examples are sufficient s.t. with prob. at least $1-\delta$, all $h \in C$ with $\hat{err}(h) = 0$ have $err(h) \leq \varepsilon$.

Infinite C, realizable

Theorem

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

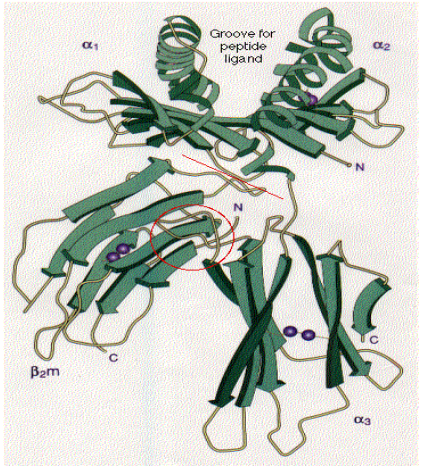
labeled examples are sufficient s.t. with prob. at least $1-\delta$, all $h \in C$ with $\hat{err}(h) = 0$ have $err(h) \leq \varepsilon$.

- Agnostic – replace ε with ε^2.
- In PAC, also talk about efficient algorithms.

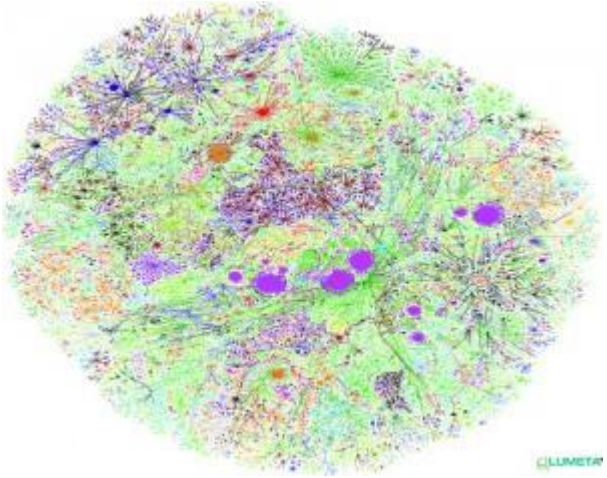
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



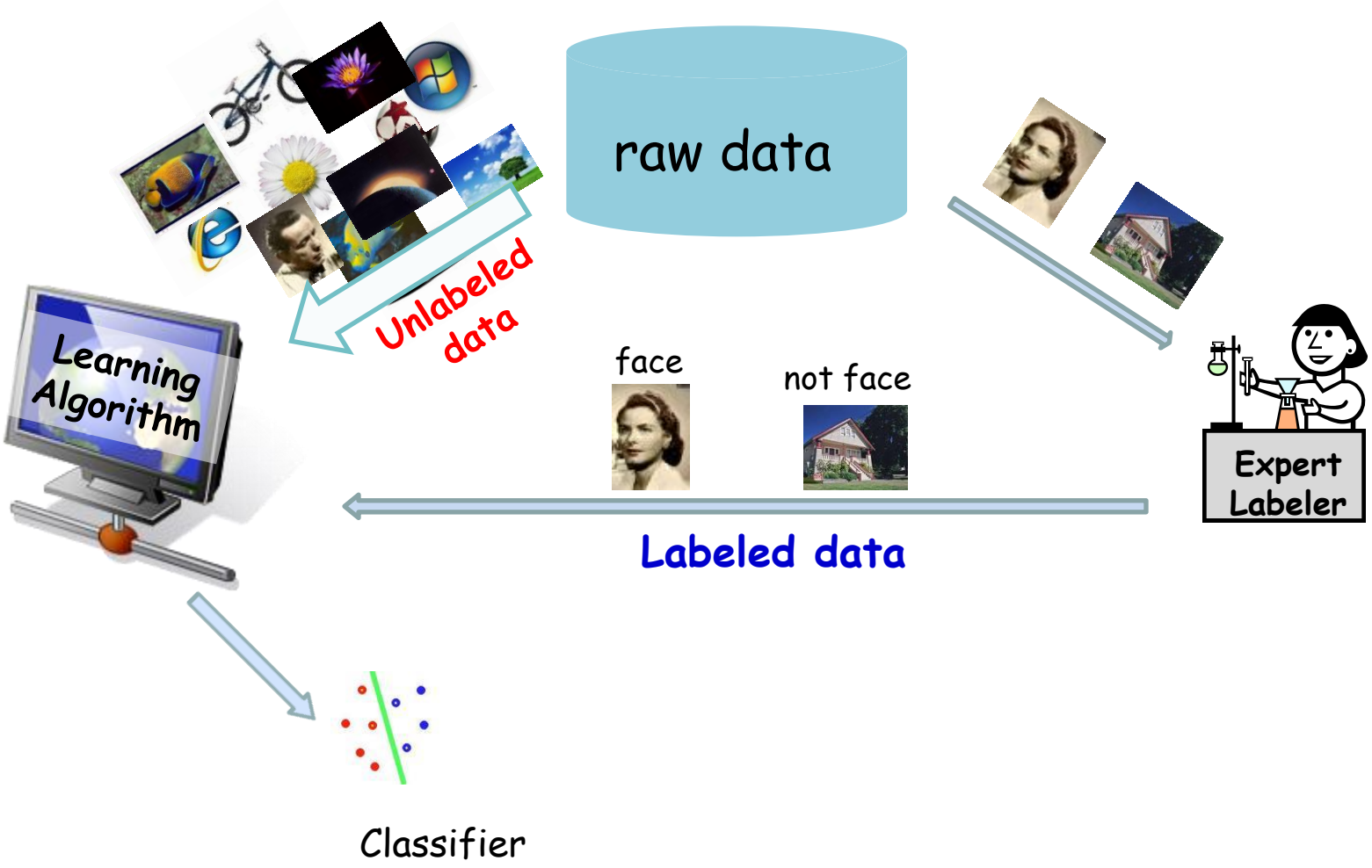
Protein sequences



Billions of webpages

Images

Semi-Supervised Learning



Semi-Supervised Learning

- 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]

- WC models not useful for SSL. Scattered and very specific theoretical results…

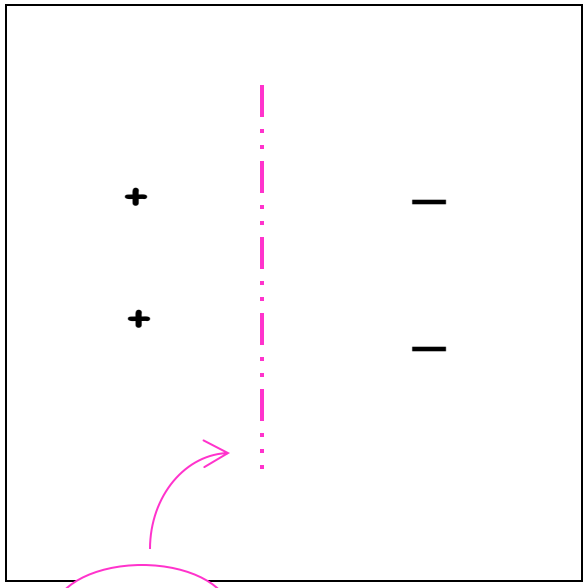
We provide: a general discriminative (PAC, SLT style) framework for SSL. [Balcan-Blum, COLT 2005; JACM 2010; book chapter, 2006]

Challenge: capture many of the assumptions typically used.

Different SSL algs based on very different assumptions.

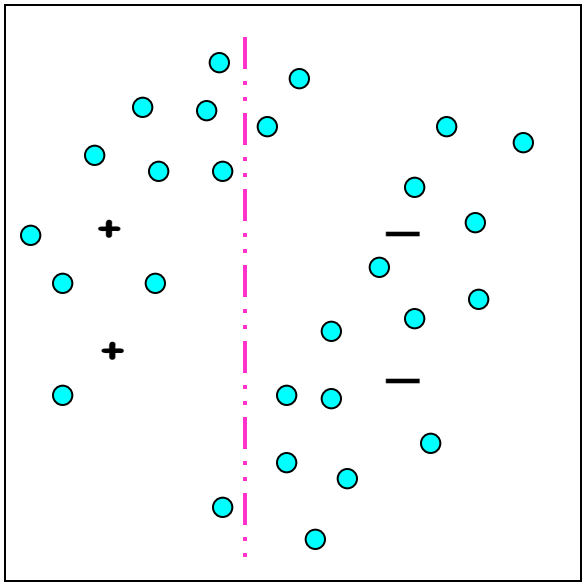
Example of "typical" assumption: Margins

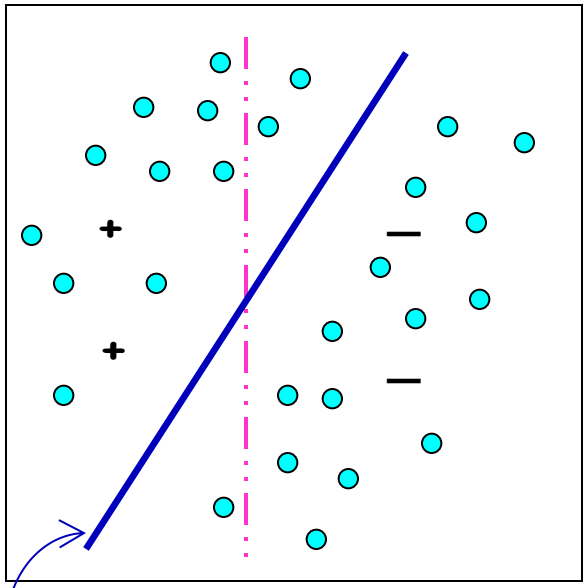
Belief: target goes through low density regions (large margin).



SVM

Labeled data only





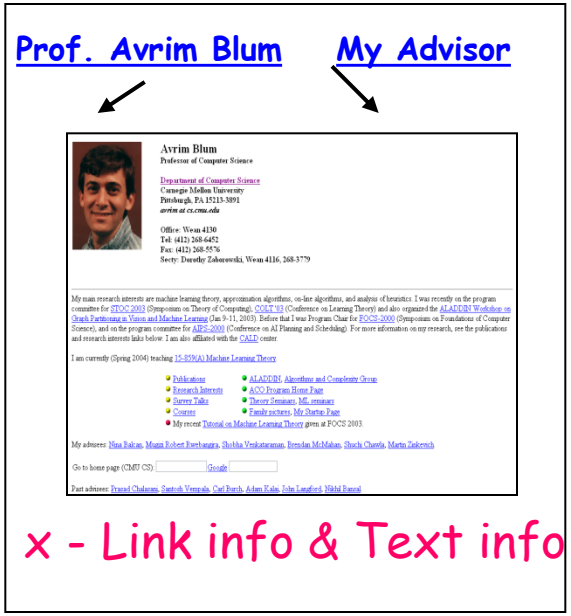
Transductive SVM

Another Example: Self-consistency

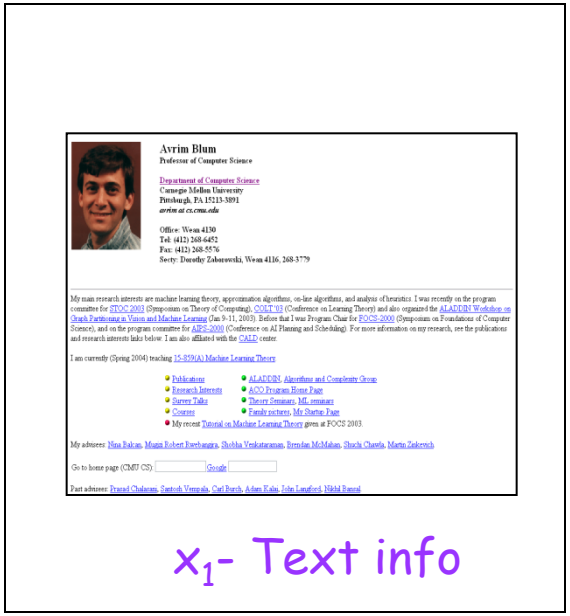
Agreement between two parts : co-training [Blum-Mitchell98].

- 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. $c_1(x_1)=c_2(x_2)=c^*(x)$

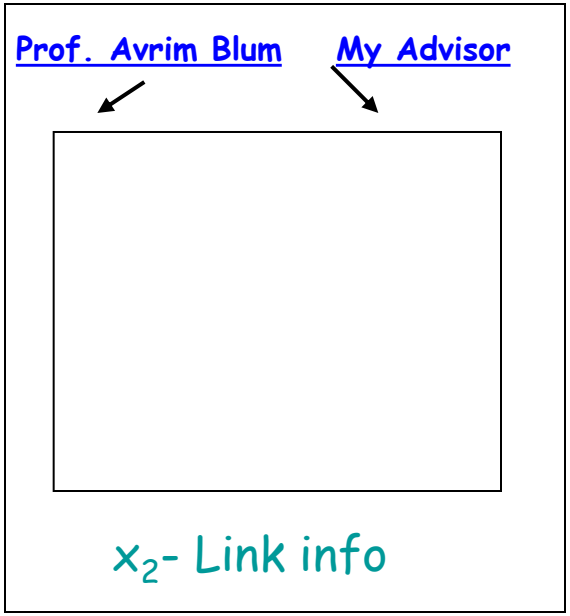
For example, if we want to classify web pages: $x = \langle x_1, x_2 \rangle$



x - Link info & Text info



x_1- Text info



x_2- Link info

New discriminative model for SSL

$S_u=\{x_i\}$ - 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.

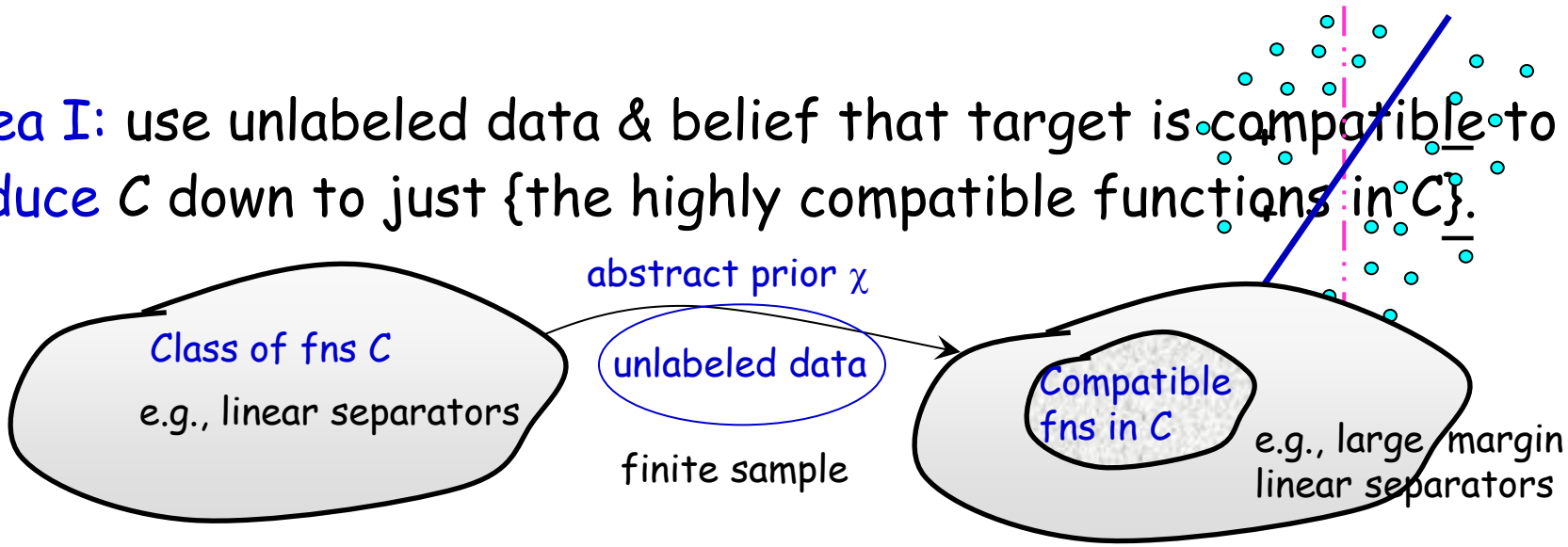
New model for SSL, Main Ideas

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)

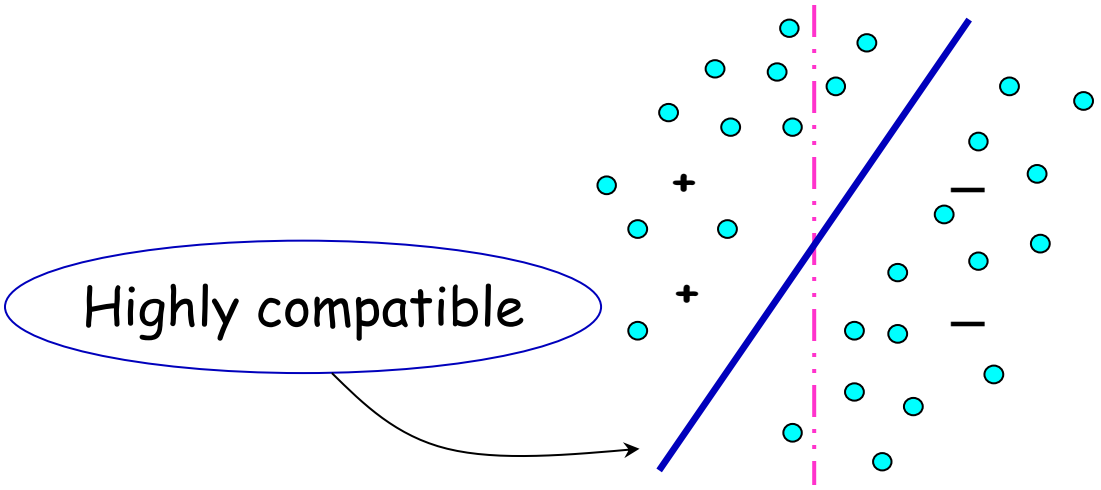
$\chi(h,D)=E_{x\in D}[\chi(h, x)]$ compatibility of h with D, $\chi(h,x)\in [0,1]$

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.

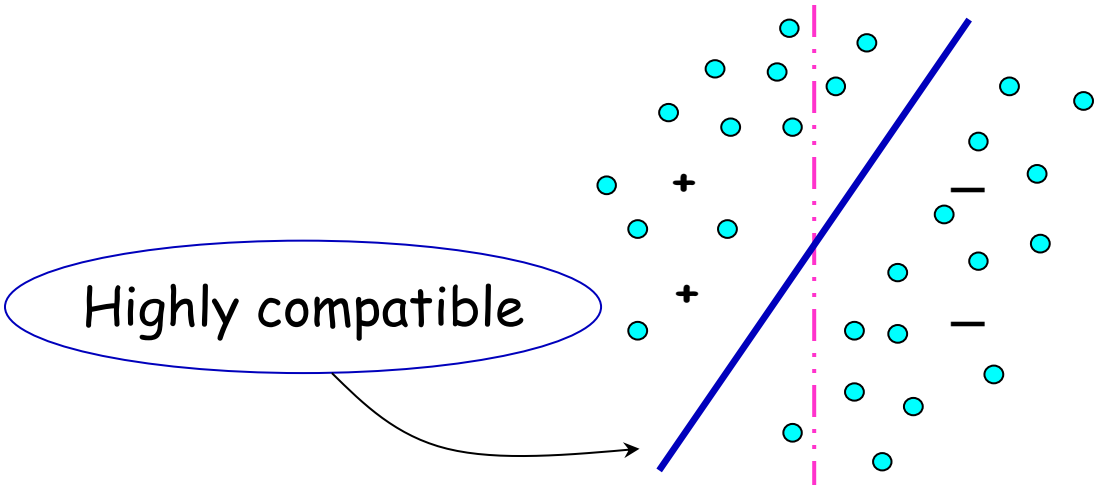
Can be written as an expectation over individual examples

$\chi(h,D)=E_{x \in D}[\chi(h,x)]$ 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 $\chi(h,x)$ 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 Our Model

As in PAC, can discuss algorithmic and sample complexity issues.

Can generically analyze sample complexity issues and instantiate them for specific cases.

- How much unlabeled data we need for learning.
- Ability of unlabeled data to reduce # of labeled examples needed for learning.

In PAC, either realizable or agnostic, now additional issue of unlabeled error rate.

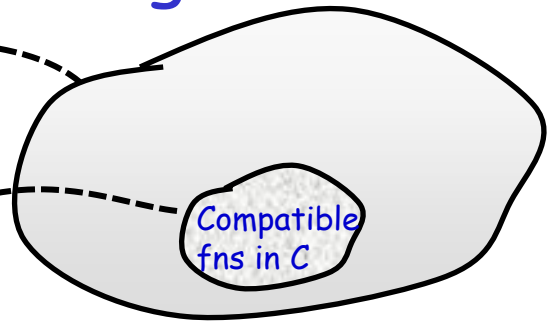
Sample Complexity, Uniform Convergence Bounds

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 $\hat{err}(h) = 0$ and $\hat{err}_{unl}(h) = 0$ (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.

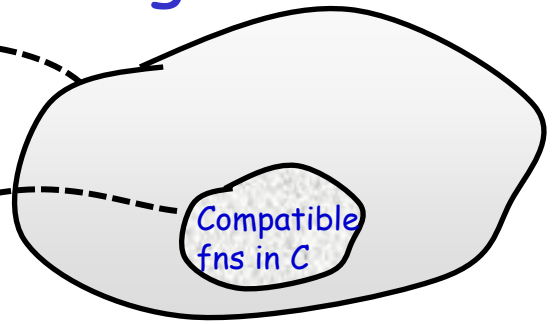
Sample Complexity, Uniform Convergence Bounds

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 $\hat{err}(h) = 0$ and $\hat{err}_{unl}(h) = 0$ (compatible with the sample) have $err(h) \leq \varepsilon$.

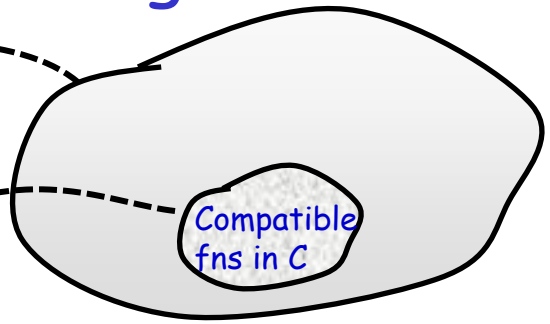
Bound # of labeled examples as a measure of the helpfulness of D wrt χ

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

Sample Complexity, Uniform Convergence Bounds

If we see

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

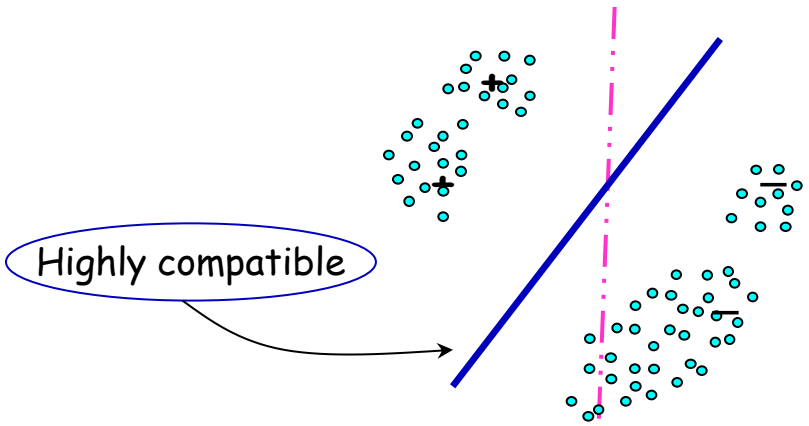


unlabeled examples and

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

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

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

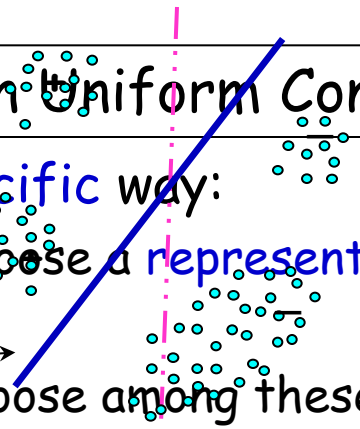
$$s = C_{D,\chi}(t + 2\varepsilon)[2m_l, D]$$

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



Summary

WC models (PAC, SLT) don't capture usefulness of unlabeled data in SSL.

We develop a new model for SSL.

Can analyze fundamental sample complexity aspects:

- How much unlabeled data is needed.
  - depends both on complexity of C and of compatibility notion.
- Ability of unlabeled data to reduce #of labeled examples.
  - compatibility of the target, helpfulness of the distribution
- Both uniform convergence bounds and epsilon-cover bounds.

Summary

WC models (PAC, SLT) don't capture usefulness of unlabeled data.

We develop a new model for SSL.

Can analyze fundamental sample complexity aspects.

Algorithmically, our analysis suggests better ways to do regularization based on unlabeled data.

Subsequent work using our framework

P. Bartlett, D. Rosenberg, AISTATS 2007; K. Sridharan, S. Kakade, COLT 2008

J. Shawe-Taylor et al., Neurocomputing 2007;

J. Zhu, survey in Encyclopedia of machine learning (2011)

Sample Complexity Subtleties

$f_t^* = \text{argmin}_{f \in C}[err(f) | err_{unl}(f) \leq \epsilon]$

Theorem

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

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 s. t. with probab. $1 - \delta$, the f in C that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(h) \leq t + \varepsilon$ has $err(f) \leq err(f_t^*) + \varepsilon + \sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*) + 2\varepsilon$

Theorem f in C that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(h) \leq t + \varepsilon$ has

$$err(f) \leq \widehat{err}(f) + \varepsilon_t + \sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*) + \varepsilon_t + \sqrt{\log(4/\delta)/(2m_l)}$$

$$\varepsilon_t = \sqrt{\frac{8}{m_l} \log(8C_{D,\chi}(t + 2\epsilon)[2m_l, D])/\delta}$$

The Agnostic Case

$f_t^* = \text{argmin}_{f \in C}[err(f) | err_{unl}(f) \leq \epsilon]$

Theorem f in C that optimizes $\widehat{err}(f)$ subject to $\widehat{err}_{unl}(h) \leq t + \varepsilon$ has

$$err(f) \leq \widehat{err}(f) + \varepsilon_t + \sqrt{\log(4/\delta)/(2m_l)} \leq err(f_t^*) + \varepsilon_t + \sqrt{\log(4/\delta)/(2m_l)}$$

Inherent tradeoff here between the quality of the comparison function f_t^* , which improves as t increases, and the estimation error ϵ_t, which gets worse as t increases. Ideally want:

$$\min_t [err(f_t^*) + \varepsilon_t] + \sqrt{\log(4/\delta)/(2m_l)}$$

Achieve it by using an estimate of ϵ_t

$$\hat{\varepsilon}_t = \sqrt{\frac{24}{m_l} \log(8C_{S,\chi}(t + 2\epsilon)[m_l, D])} \qquad f_t = \text{argmin}_{f' \in C}[err(f) | \widehat{err}_{unl}(f') \leq \epsilon]$$

Output $f = \text{argmin}_{f_t}[\widehat{err}(f_t) + \varepsilon_t]$