Generalization Bounds for Passive and Active Learning

Maria-Florina (Nina) Balcan

Carnegie Mellon University

Plan for the talk

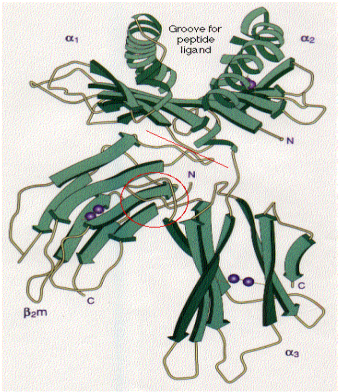
- Briefly recap classic distributional learning model and generalization bounds for supervised machine learning, discuss data-dependent bounds for deep nets.
- Active learning: learning algo takes a much more active role than in classic supervised learning in order to minimize the need for expert intervention.

Plan for the talk

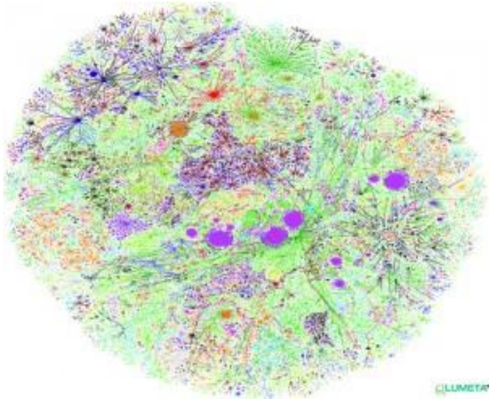
Active learning: learning algo takes a much more active role than in classic supervised learning in order to minimize the need for expert intervention.

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

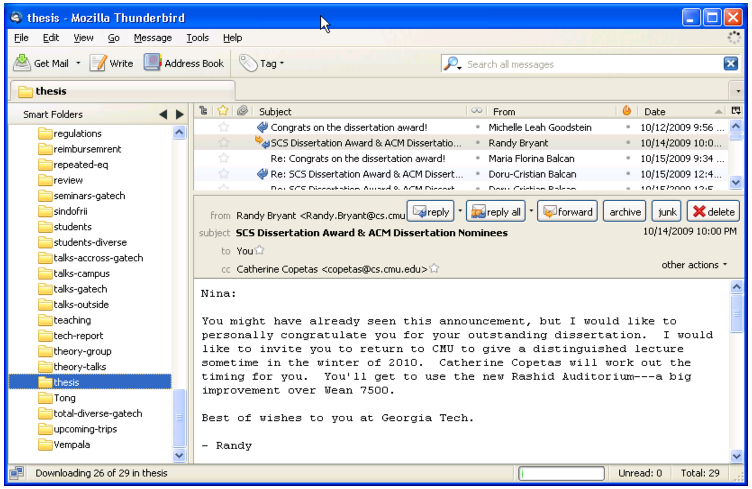
Images

Passive Supervised Learning

Supervised Learning

- E.g., which emails are spam and which are important.

Not spam



Subject	From	Date
Congrats on the dissertation award!	Michelle Leah Goodstein	10/12/2009 9:56 ...
SCS Dissertation Award & ACM Dissertatio...	Randy Bryant	10/14/2009 10:0...
Re: Congrats on the dissertation award!	Maria Florina Balcan	10/15/2009 9:34 ...
Re: SCS Dissertation Award & ACM Dissert...	Doru-Cristian Balcan	10/15/2009 12:4...

from Randy Bryant <Randy.Bryant@cs.cmu

subject **SCS Dissertation Award & ACM Dissertation Nominees** 10/14/2009 10:00 PM

to You

cc Catherine Copetas <copetas@cs.cmu.edu>

Nina:

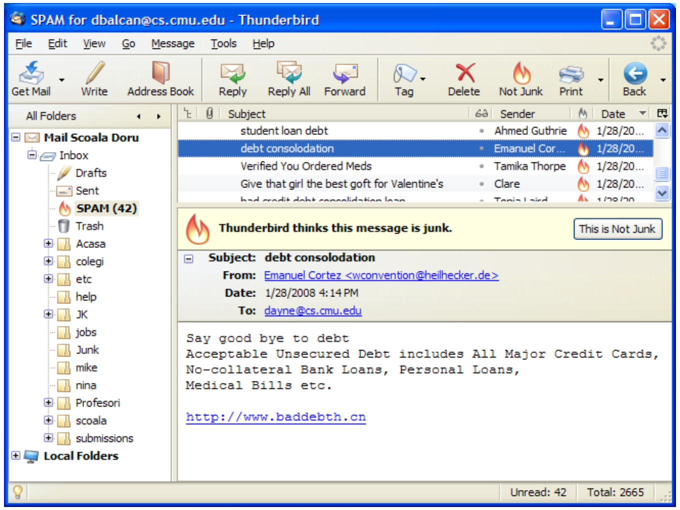
You might have already seen this announcement, but I would like to personally congratulate you for your outstanding dissertation. I would like to invite you to return to CMU to give a distinguished lecture sometime in the winter of 2010. Catherine Copetas will work out the timing for you. You'll get to use the new Rashid Auditorium---a big improvement over Wean 7500.

Best of wishes to you at Georgia Tech.

- Randy

Downloading 26 of 29 in thesis Unread: 0 Total: 29

spam



Subject	Sender	Date
student loan debt	Ahmed Guthrie	1/28/20...
debt consolodation	Emanuel Cor...	1/28/20...
Verified You Ordered Meds	Tamika Thorpe	1/28/20...
Give that girl the best goft for Valentine's	Clare	1/28/20...

Thunderbird thinks this message is junk.

Subject: debt consolodation

From: Emanuel Cortez <wconvention@heilhecker.de>

Date: 1/28/2008 4:14 PM

To: dayne@cs.cmu.edu

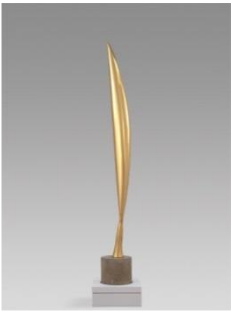
Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards,
No-collateral Bank Loans, Personal Loans,
Medical Bills etc.

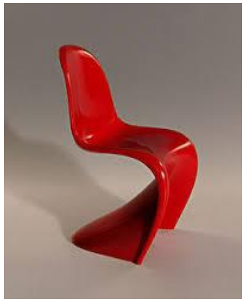
http://www.baddebth.cn

Unread: 42 Total: 2665

- E.g., classify objects as chairs vs non chairs.

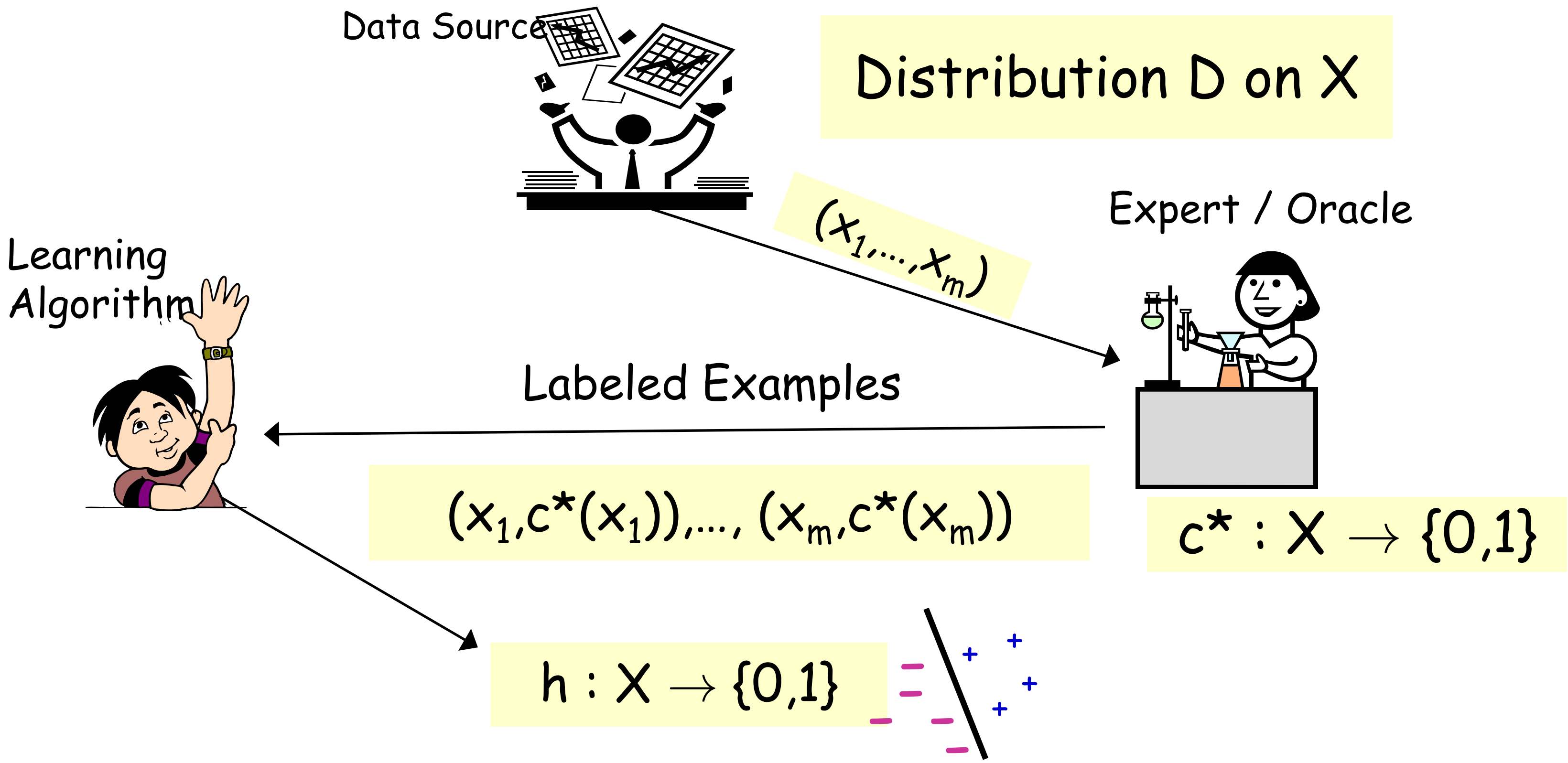
Not chair





chair

Statistical / PAC learning model



Data Source

Distribution D on X

$(x_1,\dots,x_m)$

Expert / Oracle

Learning Algorithm

Labeled Examples

$(x_1,c^*(x_1)),\dots,(x_m,c^*(x_m))$

$c^* : X \to \{0,1\}$

$h : X \to \{0,1\}$

- Algo sees $(x_1,c^*(x_1)),\dots,(x_m,c^*(x_m))$, x_i i.i.d. from D
  - Does optimization over S, finds hypothesis $h \in H$.
  - Goal: h has small error, $err(h)=Pr_{x \in D}(h(x) \neq c^*(x))$
- c^* in H, realizable case; else agnostic

Two Main Aspects in Classic Machine Learning

Algorithm Design. How to optimize? [Luigi]

Automatically generate rules that do well on observed data.

Runing time: $\mathrm{poly}\left(\mathrm{d}, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$

Generalization Guarantees, Sample Complexity [Guido]

Confidence for rule effectiveness on future data.

Sample Complexity: $\mathrm{O}\left(\frac{1}{\epsilon^2}\left(\mathrm{VCdim}(\mathrm{H}) + \log\left(\frac{1}{\delta}\right)\right)\right)$

Sample Complexity for Supervised Learning
Realizable Case

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Prob. over different samples of m training examples

Linear in 1/ϵ

Theorem

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

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

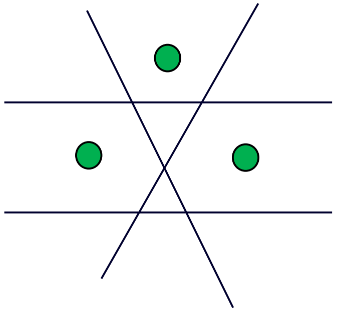
VC-dimension [Vapnik-Chervonenkis, 1971]

VC-dimension of a function class H is the cardinality of the largest set S that can be labeled in all possible ways $2^{|S|}$ by H.

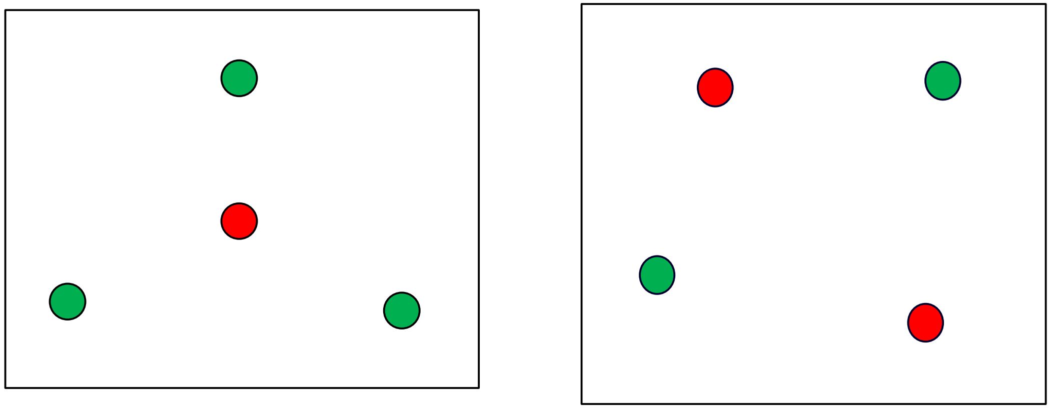
[If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(H) = \infty$]

E.g., H= linear separators in $\mathbf{R}^2$ $\mathrm{VCdim}(H) = 3$

$\mathrm{VCdim}(H) \geq 3$



$\mathrm{VCdim}(H) < 4$



E.g., H= linear separators in $\mathbf{R}^d$ $\mathrm{VCdim}(H) = d + 1$

Sample Complexity: Infinite Hypothesis Spaces

Realizable Case

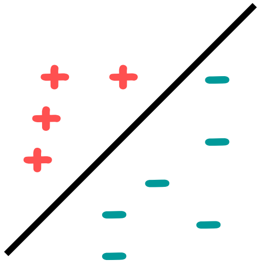
Theorem

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

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

E.g., H= linear separators in R^d

VCdim(H)= d+1



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

Sample complexity linear in d

So, if double the number of features, then only need roughly twice the number of samples to do well.

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

labeled examples are sufficient s.t. with probab. $\geq 1 - \delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$.

$1/\epsilon^2$ dependence [as opposed to $1/\epsilon$ for realizable], but get for something stronger.

2) Statistical Learning Theory style:

With prob. at least $1 - \delta$, for all $\mathrm{h} \in \mathrm{H}$:

$\sqrt{\frac{1}{m}}$ as opposed to $\frac{1}{m}$ for realizable

$$\mathrm{err}_{\mathrm{D}}(\mathrm{h}) \leq \mathrm{err}_{\mathrm{S}}(\mathrm{h}) + \sqrt{\frac{1}{2\mathrm{m}}\left(\ln\left(2|\mathrm{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right)}.$$

Sample Complexity: Infinite Hypothesis Spaces
Agnostic Case

1) How many examples suffice to get UC whp (so success for ERM).

Theorem

$$m = O\left(\frac{1}{\varepsilon^2}\left[VCdim(H) + \log\left(\frac{1}{\delta}\right)\right]\right)$$

labeled examples are sufficient so that with probab. $1-\delta$, all $h \in H$ with $|err_D(h) - err_S(h)| \leq \epsilon$.

2) Statistical Learning Theory style:

With prob. at least $1-\delta$, for all $\mathrm{h} \in \mathrm{H}$:

$$\mathrm{err_D(h)} \leq \mathrm{err_S(h)} + O\left(\sqrt{\frac{1}{2m}\left(\mathrm{VCdim(H)} + \ln\left(\frac{1}{\delta}\right)\right)}\right).$$

Tight bounds in the worst case.

VC-Dimension of Neural Networks

Theorem: H class of neural networks with L layers, W weights.

- Piecewise constant (linear threshold units): $\mathrm{VCdim(H)} = \tilde{\mathrm{O}}(\mathrm{W})$.
  [Baum-Haussler, 1989]

- Piecewise linear (ReLUs): $\mathrm{VCdim(H)} = \tilde{\mathrm{O}}(\mathrm{WL})$.
  [Bartlett-Harvey-Liaw-Mehrabian, 2017]

- Piecewise polynomial: $\mathrm{VCdim(H)} = \tilde{\mathrm{O}}(\mathrm{WL}^2)$.
  [Bartlett-Maiorov-Meir, 1998]

(Note: all final output values thresholded to $\{-1,1\}$)

Nearly tight bounds.

Classic VCdim bounds have a strong explicit dependence on # of parameters in the network.

Trivial if # of parameters exceeds the number of examples.

Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds
- Algorithm Does Implicit Regularization (finds local optima with special properties)
- Transfer Learning

Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds
- Algorithm Does Implicit Regularization (finds local optima with special properties)
- Transfer Learning

Data Dependent Generalization Bounds

- Distribution/data dependent. Tighter for nice distributions.
- Apply to general classes of real valued functions & can be used to recover the VC-bounds for supervised classification.
- Prominent technique for generalization bounds since 2000.

Covering Numbers Generalization Bounds

See Anthony-Bartlett, “Neural Network Learning: Theoretical Foundations”, 1999.

Rademacher Complexity Generalization Bounds

See Bousquet-Boucheron-Lugosi, “Introduction to Statistical Learning Theory”, 2014.

Rademacher Complexity

Problem Setup

- A space Z and a distr. $D_{|Z}$
- F be a class of functions from Z to $[0,1]$
- $S = \{z_1, \dots, z_m\}$ be i.i.d. from $D_{|Z}$

Want a high prob. uniform convergence bound, all $f \in F$ satisfy:

$$E_D[f(z)] \leq E_S[f(z)] + \boxed{\text{term}(\text{complexity of F, niceness of D/S})}$$

What measure of complexity?

General discrete Y

E.g., $Z = X \times Y$, $Y = \{-1,1\}$, $H = \{h\!: X \rightarrow Y\}$ hyp. space (e.g., lin. sep)

$F = L(H) = \{l_h\!: X \times Y \rightarrow [0,1]\}$, where $l_h\big(z = (x,y)\big) = 1_{\{h(x) \neq y\}}$

[Loss fnc induced by h and 0/1 loss]

Then $E_{z \sim D}[l_h(z)] = \text{err}_D(h)$ and $E_S[l_h(z)] = \text{err}_S(h)$.

$$\text{err}_D[h] \leq \text{err}_S[h] + \text{term}(\text{complexity of H, niceness of D/S})$$

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

sup measures for any given set S and Rademacher vector σ, the max correlation between $f(z_i)$ and σ_i for all $f \in F$

So, taking the expectation over σ this measures the ability of class F to fit random noise.

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\widehat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\widehat{R}_m(F)]$

Theorem: Whp all $f \in F$ satisfy:

Useful if it decays with m.

$$E_D[f(z)] \le E_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$E_D[f(z)] \le E_S[f(z)] + 2\,\widehat{R}_m(F) + 3\sqrt{\frac{\ln(1/\delta)}{m}}$$

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\widehat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\widehat{R}_m(F)]$

E.g.,:

1) F={f}, then $\widehat{R}_m(F) = 0$

[Linearity of expectation: each $\sigma_i f(z_i)$ individually has expectation 0.]

2) F={all 0/1 fnc}, then $\widehat{R}_m(F) = 1/2$

[To maximize set $f(z_i) = 1$ when $\sigma_i = 1$ and $f(z_i) = 0$ when $\sigma_i = -1$. Then quantity inside expectation is $\#1's \in \sigma$, which is m/2 by linearity of expectation.]

Rademacher Complexity

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \dots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\widehat{R}_m(F) = E_{\sigma_1,\dots,\sigma_m}\left[\sup_{f\in F} \frac{1}{m}\sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables **chosen** uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\widehat{R}_m(F)]$

E.g.,:

1) F={f}, then $\widehat{R}_m(F) = 0$

2) F={all 0/1 fnc}, then $\widehat{R}_m(F) = 1/2$

3) F=L(H), H=binary classifiers then: $R_S(F) \le \sqrt{\frac{\ln(2|H[S]|)}{m}}$

H finite: $R_S(F) \le \sqrt{\frac{\ln(2|H|)}{m}}$

Rademacher Complexity Bounds

Space Z and a distr. $D_{|Z}$; F be a class of functions from Z to [0,1]

Let $S = \{z_1, \ldots, z_m\}$ be i.i.d from $D_{|Z}$.

The empirical Rademacher complexity of F is:

$$\hat{R}_m(F) = E_{\sigma_1,\ldots,\sigma_m}\left[\sup_{f \in F} \frac{1}{m} \sum_i \sigma_i f(z_i)\right]$$

where σ_i are i.i.d. Rademacher variables chosen uniformly from $\{-1,1\}$.

The Rademacher complexity of F is: $R_m(F) = E_S[\hat{R}_m(F)]$

Theorem: Whp all $f \in F$ satisfy:

Data dependent bound!

$$E_D[f(z)] \leq E_S[f(z)] + 2R_m(F) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

$$E_D[f(z)] \leq E_S[f(z)] + 2\,\hat{R}_m(F) + 3\sqrt{\frac{\ln(1/\delta)}{m}}$$

Bound expectation of each f in terms of its empirical average & the RC of F

Proof uses Symmetrization and Ghost Sample Tricks! (same as for VC bound)

Rademacher Complex: Binary classification

Fact: $H = \{h: X \to Y\}$ hyp. space (e.g., lin. sep) F= L(H), d=VCdim(H):

$$R_S(F) \leq \sqrt{\frac{\ln(2|H[S]|)}{m}}$$

So, by Sauer's lemma, $R_S(F) \leq \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}}$

Theorem: For any H, any distr. D, w.h.p. $\geq 1-\delta$ all $h \in H$ satisfy:

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + R_m(H) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

$$\mathrm{err}_D(h) \leq \mathrm{err}_S(h) + \sqrt{\frac{2d\ln\left(\frac{em}{d}\right)}{m}} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

generalization bound

Many more uses!!! Margin bounds for SVM, boosting, regression bounds, margin based bounds for deep nets, etc.

Data-Dependent Bounds for Deep Networks

E.g., very recent papers:

- Via covering numbers: “Spectrally-normalized margin bounds for neural networks”. [Bartlett-Foster-Telgarsky, NIPS 2017]

- Via Rademacher complexity: “Size-independent sample complexity of neural networks”. [Golowich-Rakhlin-Shamir, COLT 2018]

Data-Dependent Bounds for Deep Networks

- Spectrally-normalized margin bounds for neural networks. [Bartlett-Foster-Telgarsky, NIPS 2017]

Theorem: With high probability, every $\mathrm{f_W}$ with $\mathrm{R_W} \leq \mathrm{r}$ satisfies

$$\Pr(\mathrm{M}(\mathrm{f_W}(\mathrm{X}), \mathrm{Y}) \leq 0) \leq \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{1}[\mathrm{M}(\mathrm{f_W}(\mathrm{X_i}), \mathrm{Y_i}) \leq \gamma] + \tilde{\mathrm{O}}\left(\frac{\mathrm{RL}}{\gamma\sqrt{\mathrm{n}}}\right)$$

- Network with L layers, parameters $\mathrm{W_1}, \ldots, \mathrm{W_L}$:

$$\mathrm{f_W}(\mathrm{x}) \coloneqq \sigma(\mathrm{W_L}\sigma_{\mathrm{L}-1}(\mathrm{W}_{\mathrm{L}-1} \ldots \sigma_1(\mathrm{W_1 x}) \ldots))$$

$$R_W \coloneqq \prod_{i=1}^{L} \left\| W_i \right\|_* \left(\sum_{i=1}^{L} \frac{\left\| W_i \right\|_{2,1}^{2/3}}{\left\| W \right\|_*^{2/3}} \right)^{3/2}$$

spectral norm

[Golowich-Rakhlin-Shamir, COLT 2018] provide a related bound via a Rademacher complexity argument

Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Stronger Data-Dependent Bounds
- Algorithm Does Implicit Regularization (finds local optima with special properties)

"Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations". [Li-Ma-Zhang. COLT 2018]

- Transfer Learning

"Risk Bounds for Transferring Representations With and Without Fine-Tuning". [McNamara-Balcan. ICML 2017]

Generalization in Deep Nets

How can we explain successful training of very deep networks?

- Str
- Alg
  with

  “A
  N
- Tr

  “F
  T

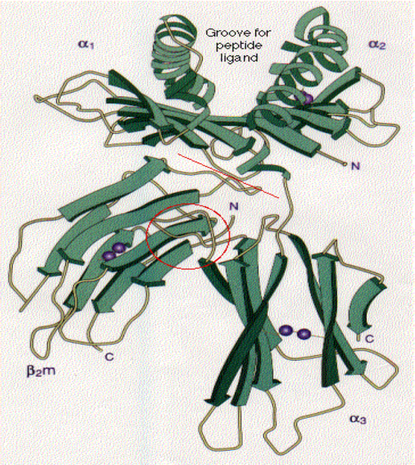
Lots of open questions.

Active Learning

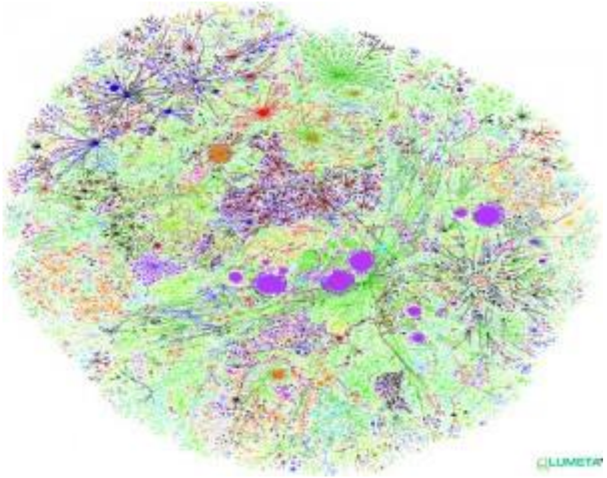
Classic Fully Supervised Learning Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

Images

Modern ML: New Learning Approaches

Modern applications: massive amounts of raw data.

Techniques that best utilize data, minimizing need for expert/human intervention.

Paradigms where there has been great progress.

- Semi-supervised Learning, (Inter)active Learning.

Active Learning

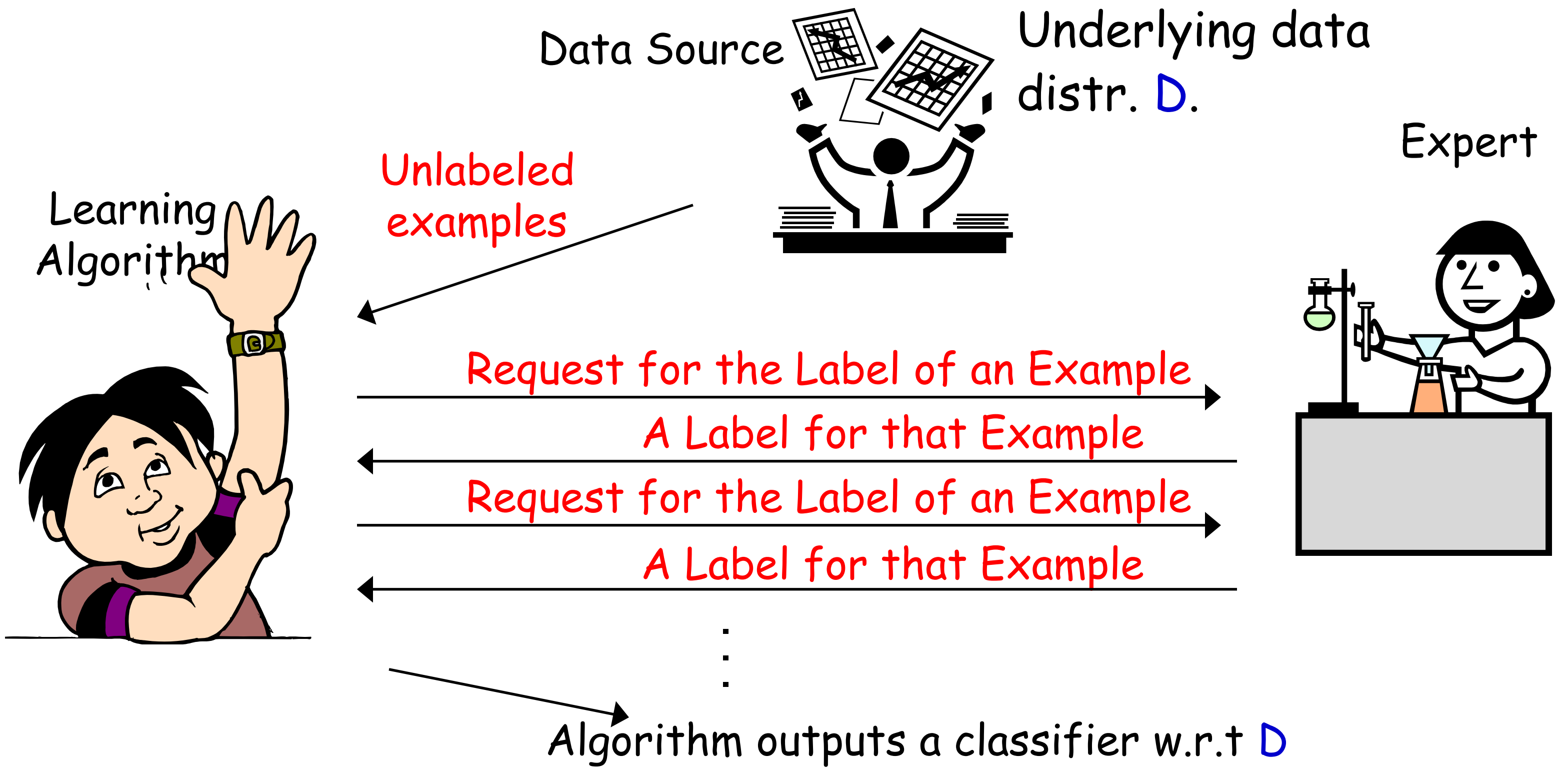
Nice resources:

- Two faces of active learning. Sanjoy Dasgupta. 2011.
- Active Learning. Bur Settles. 2012.
- Active Learning. Balcan-Urner. Encyclopedia of Algorithms. 2015

- Interactive Learning Workshop, Foundations of Machine Learning Semester, Simons Theory of Computing:

https://simons.berkeley.edu/workshops/machinelearning2017-1

Batch Active Learning



Data Source

Underlying data distr. D.

Expert

Learning Algorithm

Unlabeled examples

Request for the Label of an Example

A Label for that Example

Request for the Label of an Example

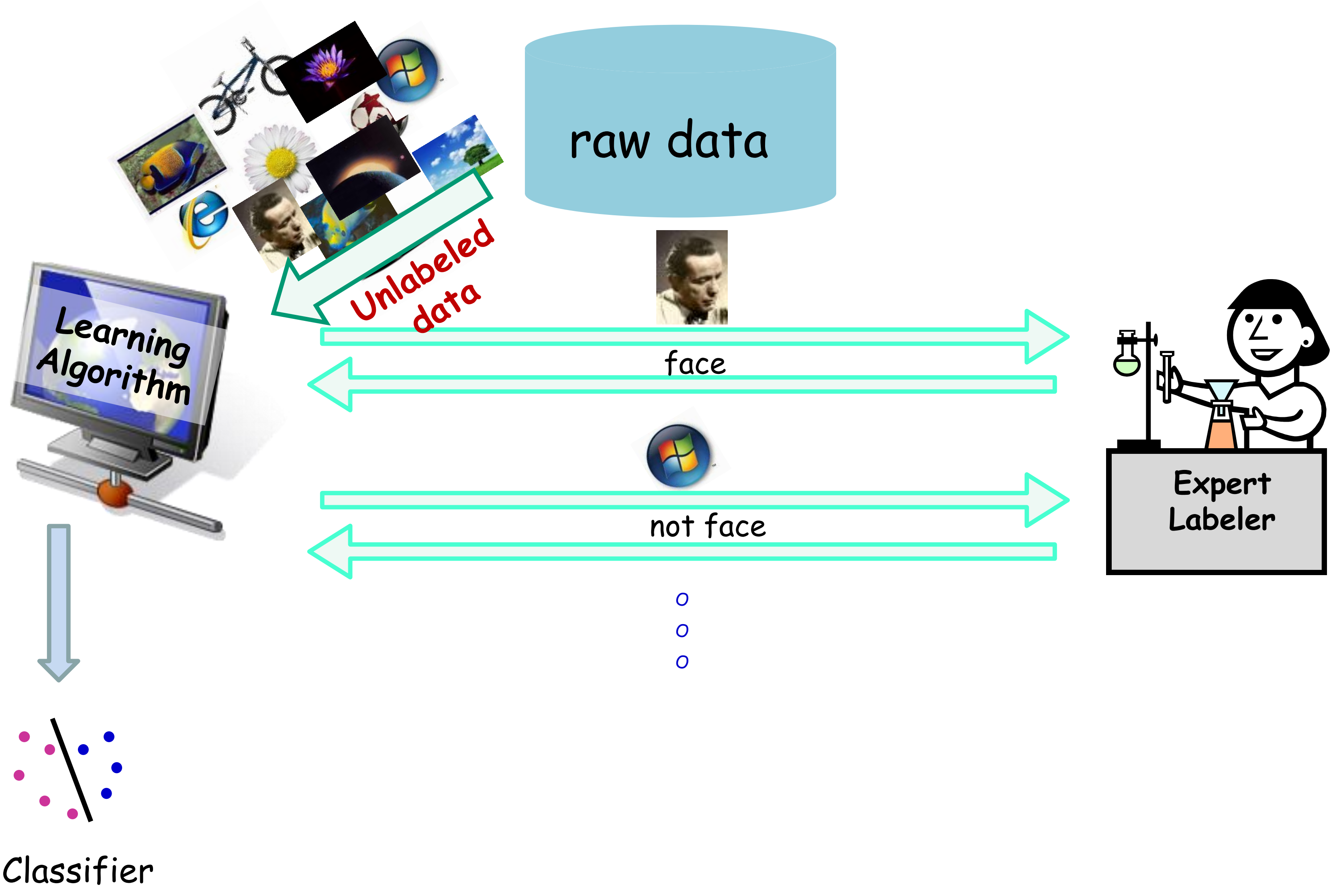
A Label for that Example

⋮

Algorithm outputs a classifier w.r.t D

- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples [pick informative examples to be labeled].

Active Learning



raw data

Unlabeled data

Learning Algorithm

face

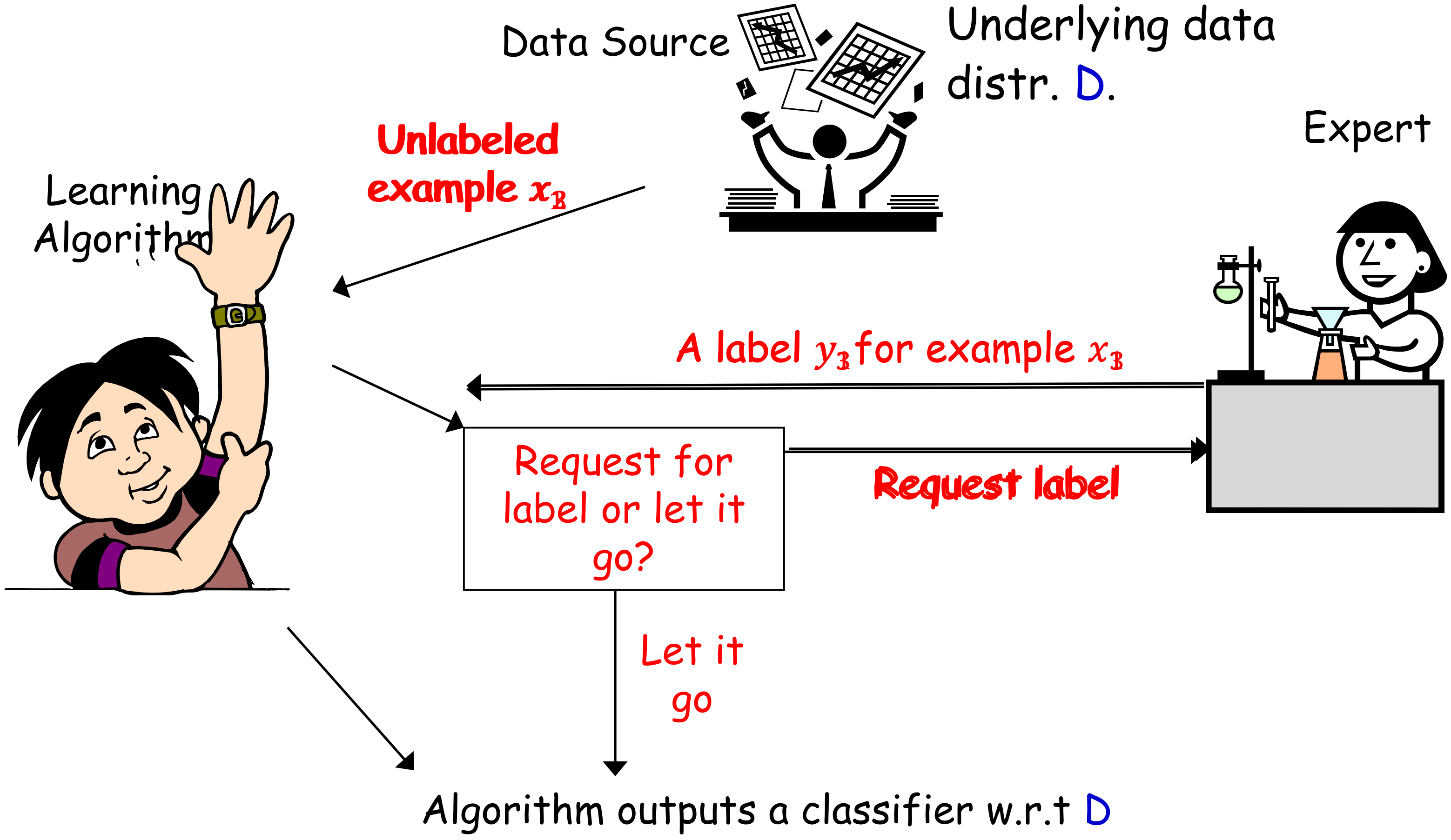
not face

o
o
o

Expert Labeler

Classifier

Selective Sampling Active Learning



Data Source

Underlying data distr. D.

Expert

Learning Algorithm

Unlabeled example x_3

A label y_3 for example x_3

Request for label or let it go?

Request label

Let it go

Algorithm outputs a classifier w.r.t D

- Selective sampling AL (Online AL): stream of unlabeled examples, when each arrives make a decision to ask for label or not.
- Goal: use fewer labeled examples [pick informative examples to be labeled].

What Makes a Good Active Learning Algorithm?

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

- Doesn't make too many label requests.

  Hopefully a lot less than passive learning and SSL.

- Need to choose the label requests carefully, to get informative labels.

Can adaptive querying really do better than passive/random sampling?

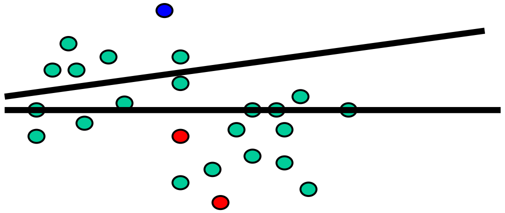
- YES! (sometimes)

- We often need far fewer labels for active learning than for passive.

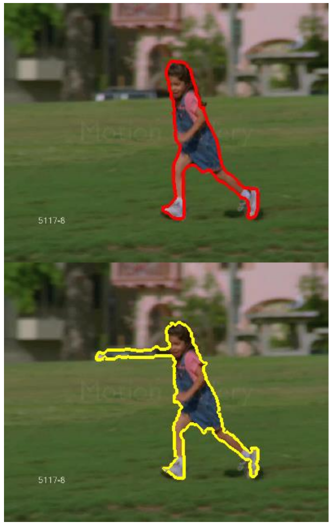
- This is predicted by theory and has been observed in practice.

Active Learning in Practice

- Text classification: active SVM (Tong-Koller, ICML2000).
  - e.g., request label of the example closest to current separator.

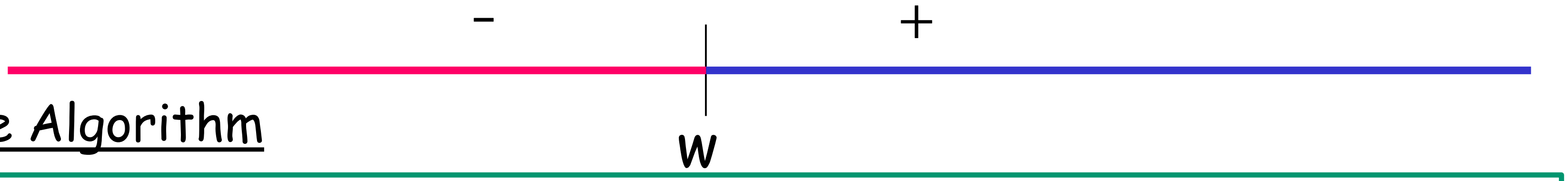


- Video Segmentation (Fathi-Balcan-Ren-Regh, BMVC 11).



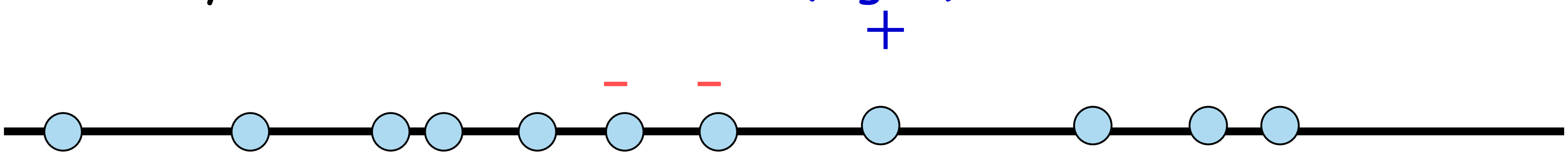
Can adaptive querying help? [CAL92, Dasgupta04]

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



Active Algorithm

- Get N unlabeled examples
- How can we recover the correct labels with $\ll$ N queries?
- Do binary search! Just need O(log N) labels!



- Output a classifier consistent with the N inferred labels.

- $N = O(1/\epsilon)$ we are guaranteed to get a classifier of error $\leq \epsilon$.

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

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

Common Technique in Practice

Uncertainty sampling in SVMs common and quite useful in practice. E.g., [Tong-Koller, ICML 2000; Jain-Vijayanarasimhan-Grauman, NIPS 2010; Schohon Cohn, ICML 2000]

Active SVM Algorithm

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

Common Technique in Practice

Active SVM seems to be quite useful in practice.

[Tong-Koller, ICML 2000; Jain-Vijayanarasimhan-Grauman, NIPS 2010]

Algorithm (batch version)

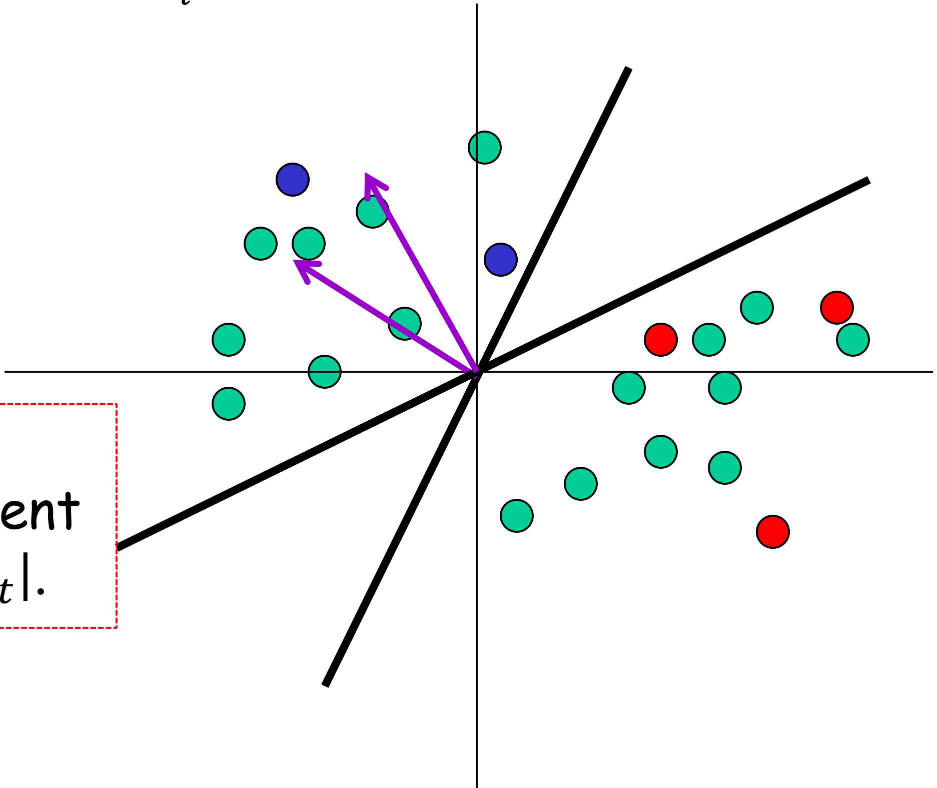
Input $S_u=\{x_1, \ldots, x_{m_u}\}$ drawn i.i.d from the underlying source D

Start: query for the labels of a few random x_is.

For $t = 1, \ldots,$

- Find w_t the max-margin separator of all labeled points so far.
- Request the label of the example closest to the current separator: minimizing $|x_i \cdot w_t|$.

(highest uncertainty)

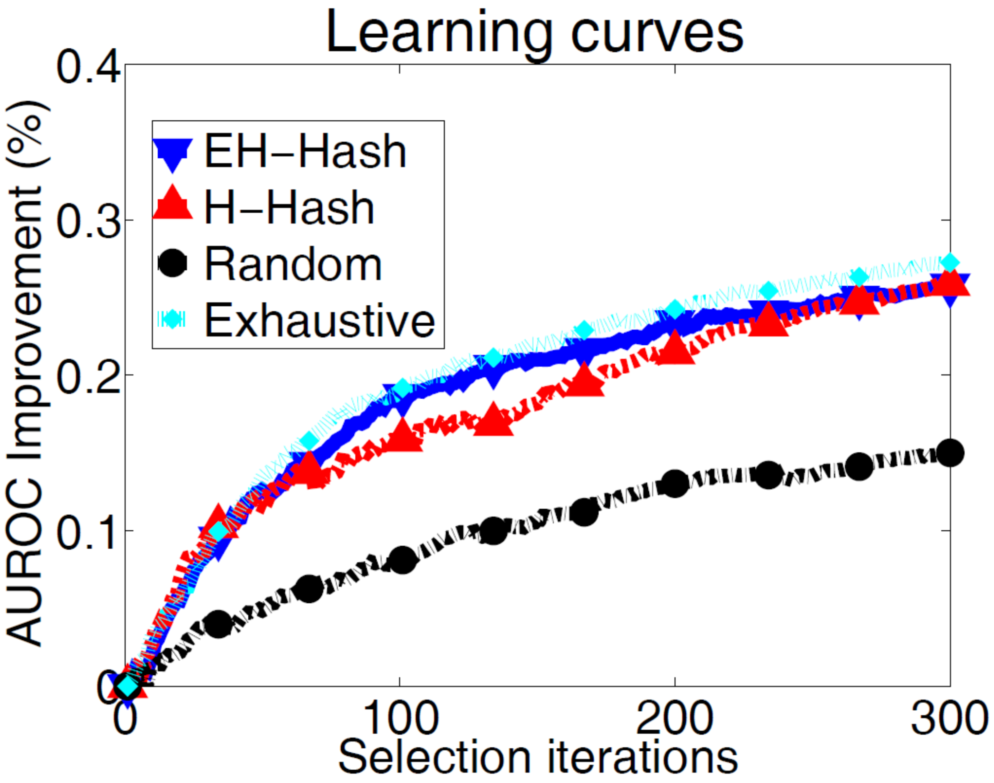


Common Technique in Practice

Active SVM seems to be quite useful in practice.

E.g., Jain-Vijayanarasimhan-Grauman, NIPS 2010

Newsgroups dataset (20.000 documents from 20 categories)

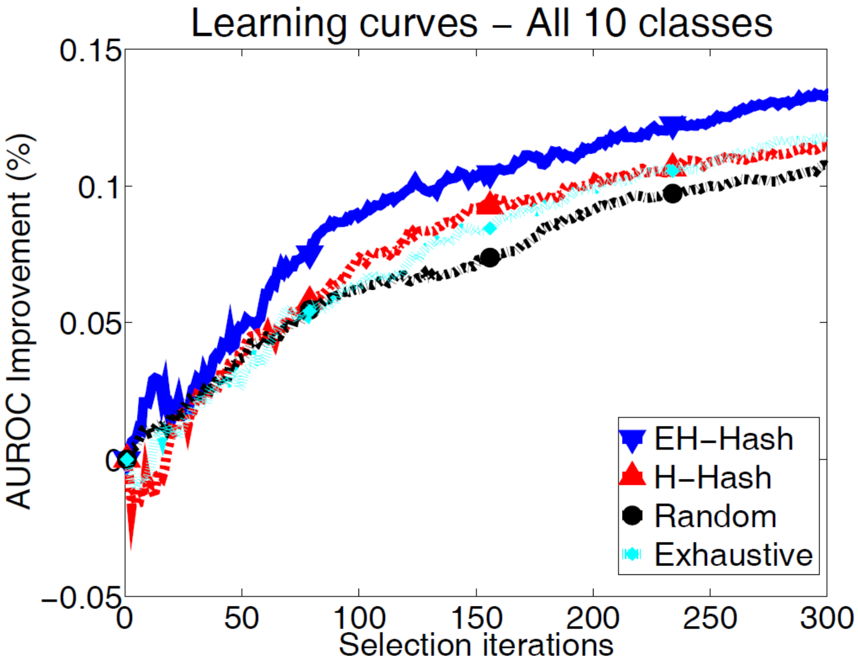


Common Technique in Practice

Active SVM seems to be quite useful in practice.

E.g., Jain-Vijayanarasimhan-Grauman, NIPS 2010

CIFAR-10 image dataset (60.000 images from 10 categories)



Active SVM/Uncertainty Sampling

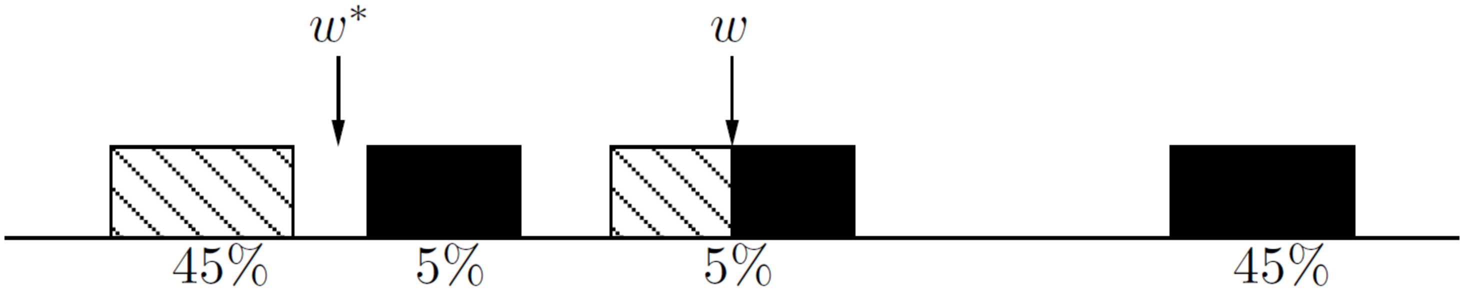
- Works sometimes….

- **However, we need to be very very very careful!!!**

  - Myopic, greedy technique can suffer from sampling bias.

  - A bias created because of the querying strategy; as time goes on the sample is less and less representative of the true data source.

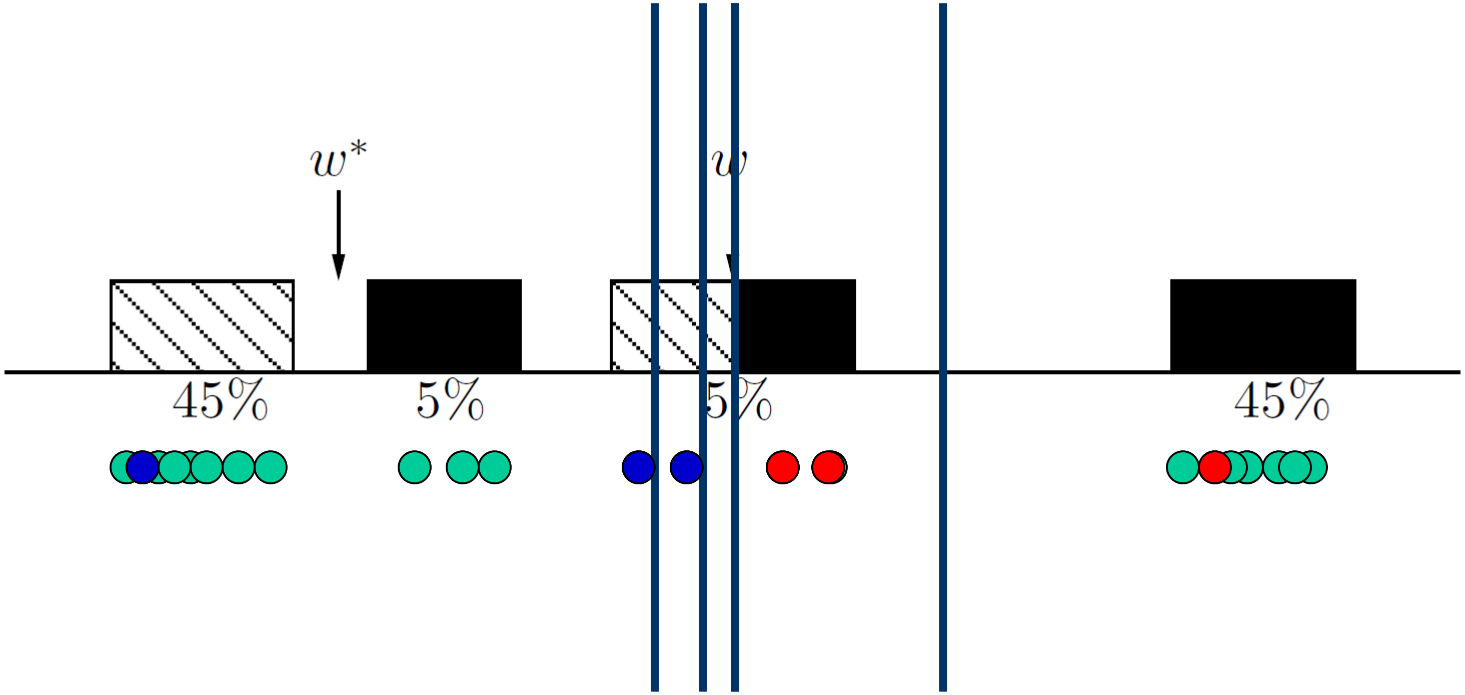
[Dasgupta10]



Active SVM/Uncertainty Sampling

- Works sometimes….

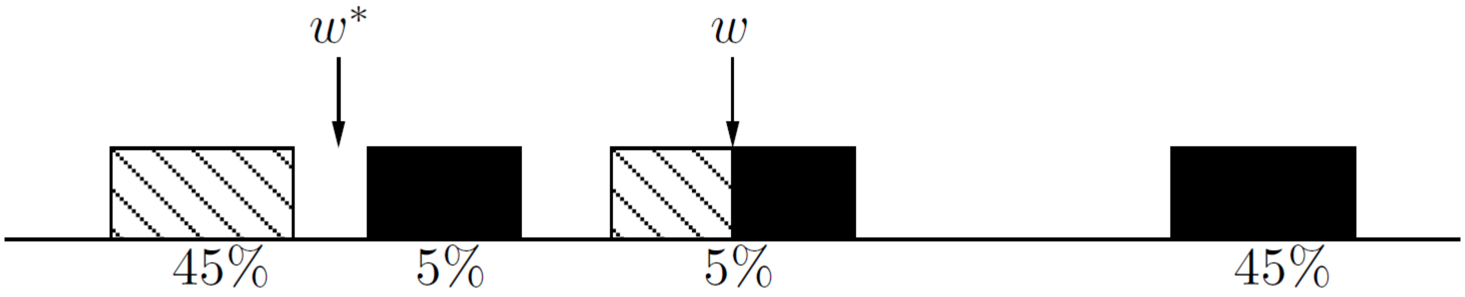
- **However, we need to be very very careful!!!**



Active SVM/Uncertainty Sampling

- Works sometimes….
- **However, we need to be very very careful!!!**
  - Myopic, greedy technique can suffer from sampling bias.
  - Bias created because of the querying strategy; as time goes on the sample is less and less representative of the true source.
  - Observed in practice too!!!!

- **Main tension**: want to choose informative points, but also want to guarantee that the classifier we output does well on true random examples from the underlying distribution.



Safe Active Learning Schemes

Disagreement Based Active Learning

Hypothesis Space Search

[CAL92] [BBL06]

[Hanneke'07, DHM'07, Wang'09 , Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, …]

Version Spaces

- X - feature/instance space; distr. D over X; c^* target fnc
- Fix hypothesis space H.

Definition (Mitchell'82) Assume realizable case: $c^* \in H$.

Given a set of labeled examples (x_1, y_1), …,(x_{m_l}, y_{m_l}), $y_i = c^*(x_i)$

Version space of H: part of H consistent with labels so far.

I.e., $h \in VS(H)$ iff $h(x_i) = c^*(x_i)\ \forall i \in \{1, \dots, m_l\}$.

Version Spaces

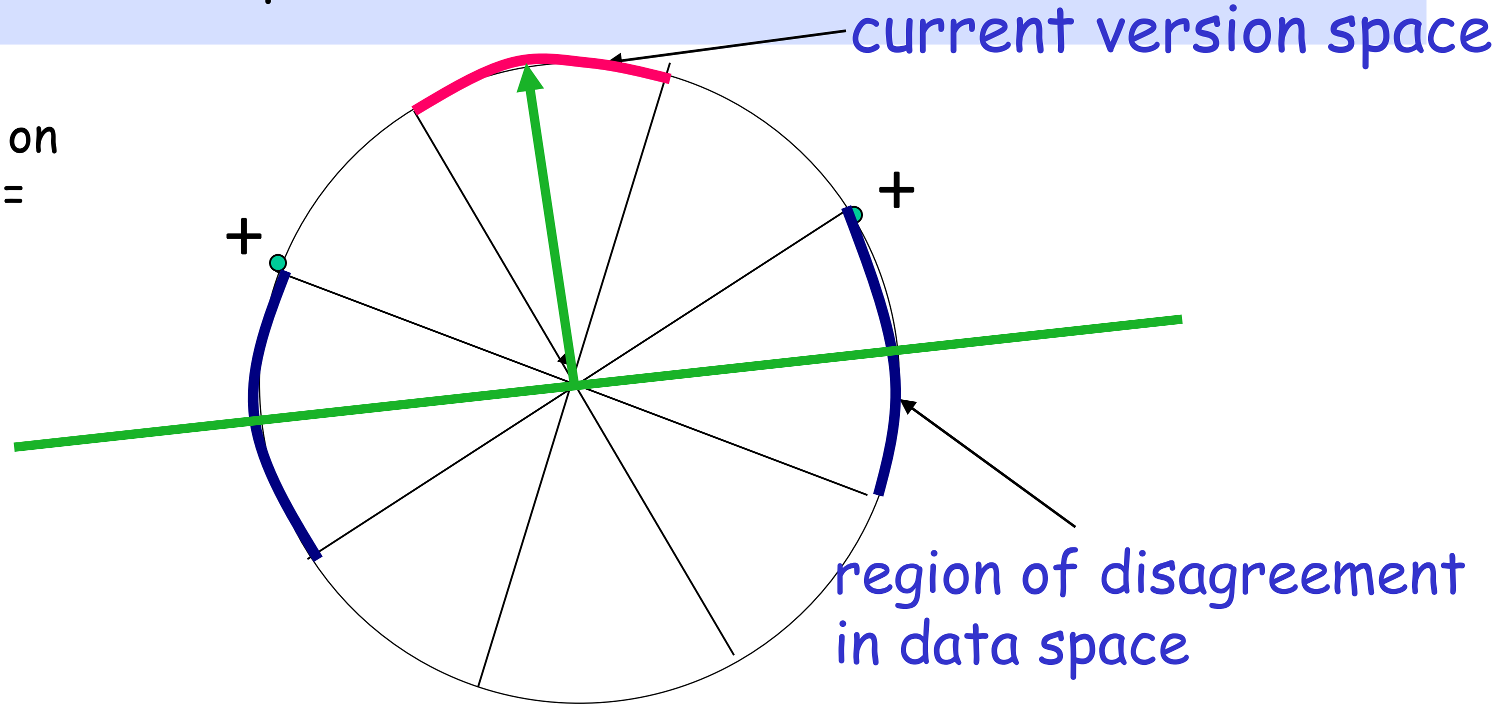
- X – feature/instance space; distr. D over X; c^* target fnc
- Fix hypothesis space H.

Definition (Mitchell'82) Assume realizable case: $c^* \in H$.

Given a set of labeled examples $(x_1, y_1), \ldots, (x_{m_l}, y_{m_l}), y_i = c^*(x_i)$

Version space of H: part of H consistent with labels so far.

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.



Version Spaces. Region of Disagreement

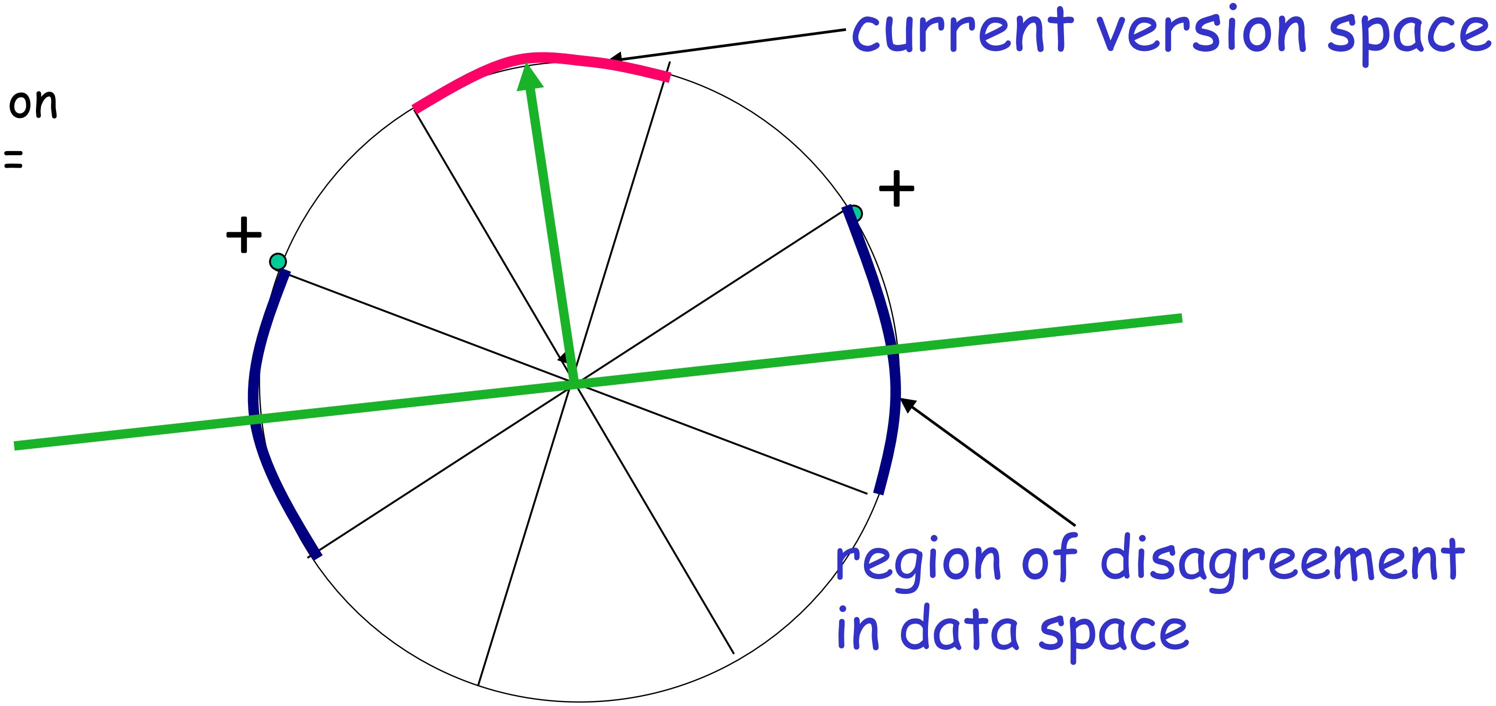
Definition (CAL'92)

Version space: part of H consistent with labels so far.

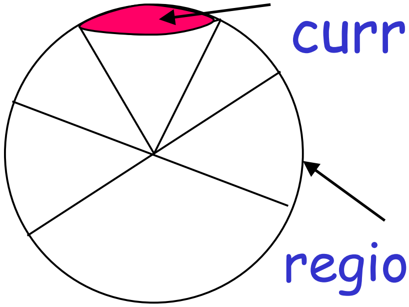
Region of disagreement = part of data space about which there is still some uncertainty (i.e. disagreement within version space)

$x \in X, x \in DIS(VS(H))$ iff $\exists h_1, h_2 \in VS(H), h_1(x) \neq h_2(x)$

E.g.,: data lies on circle in R^2, H = homogeneous linear seps.



Disagreement Based Active Learning [CAL92]



current version space

region of uncertainy

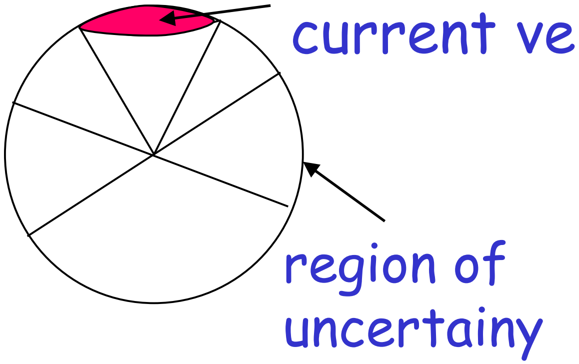
Algorithm:

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

Stop when region of uncertainty is small.

Note: it is active since we do not waste labels by querying in regions of space we are certain about the labels.

Disagreement Based Active Learning [CAL92]



Algorithm:

Query for the labels of a few random x_is.

Let H_1 be the current version space.

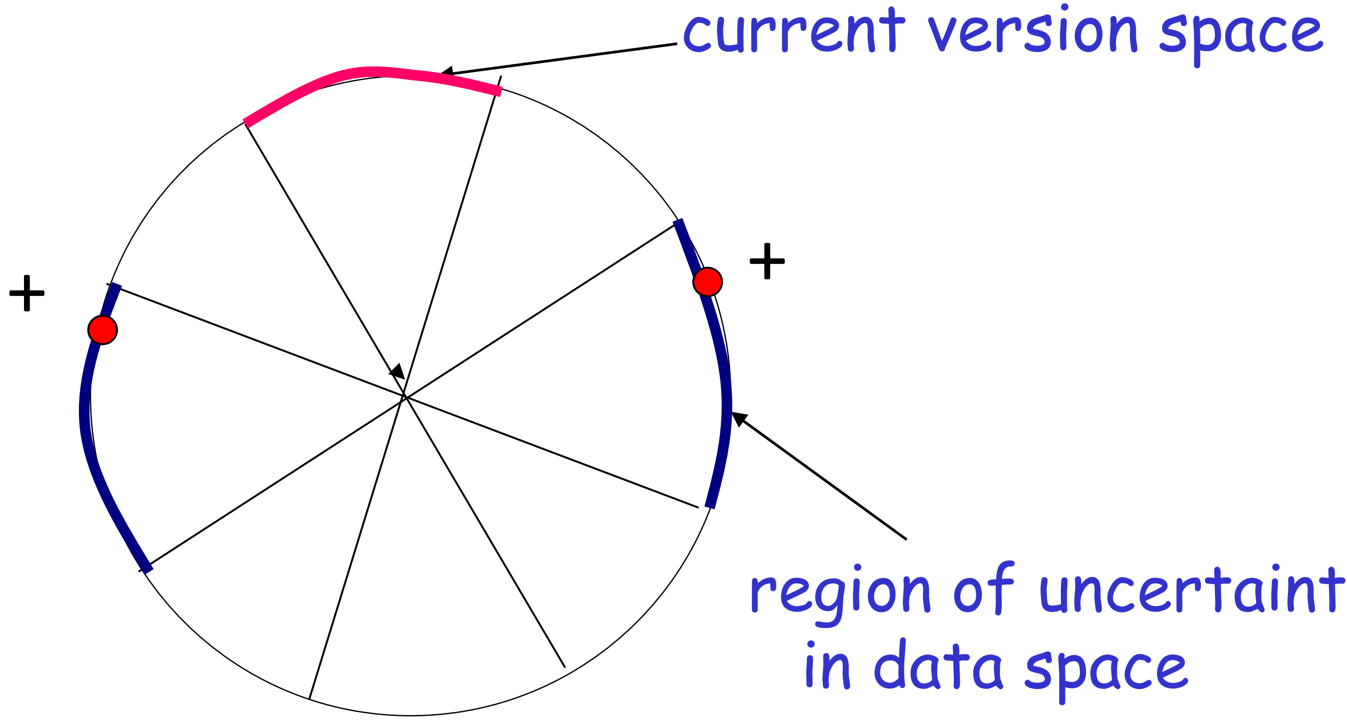
For $\boldsymbol{t = 1}$, ….,

    Pick a few points at random from the current region of disagreement $\mathrm{DIS}(\mathrm{H}_t)$ and query their labels.

    Let H_{t+1} be the new version space.

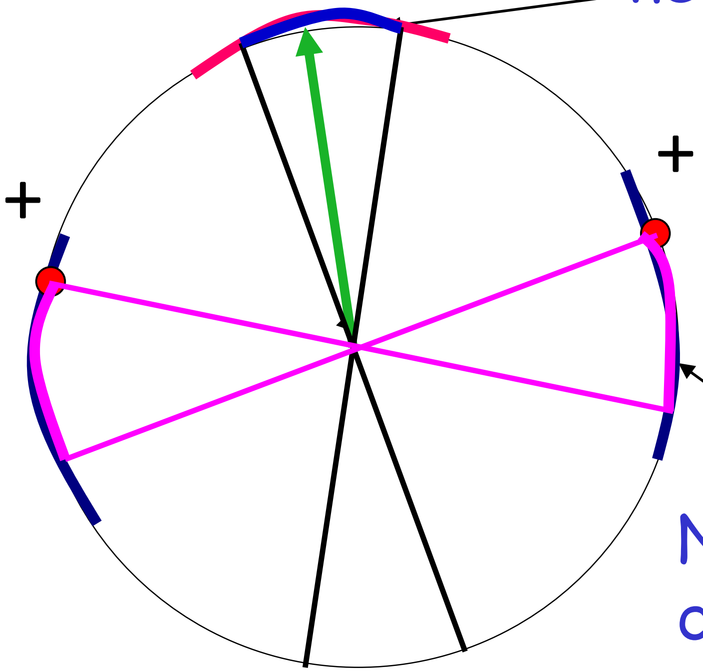
Region of uncertainty [CAL92]

- 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]

- 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)



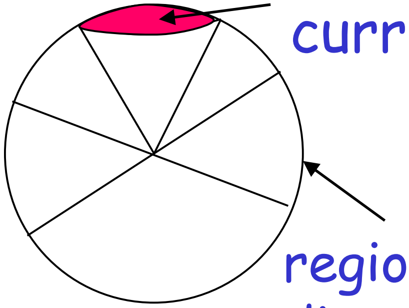
new version space

New region of disagreement in data space

How about the agnostic case where the target might not belong the H?

A^2 Agnostic Active Learner

[Balcan, Beygelzimer, Langford, ICML'06] [Balcan, Beygelzimer, Langford, JCSS'08]



current version space

region of disagreement

Algorithm:

Careful use of generalization bounds;
Avoid the sampling bias!!!!

Let $H_1 = H.$

For $t = 1, \ldots.,$

- Pick a few points at random from the current region of disagreement $\mathrm{DIS}(H_t)$ and query their labels.
- Throw out hypothesis if you are statistically confident they are suboptimal.

When Active Learning Helps. Agnostic case

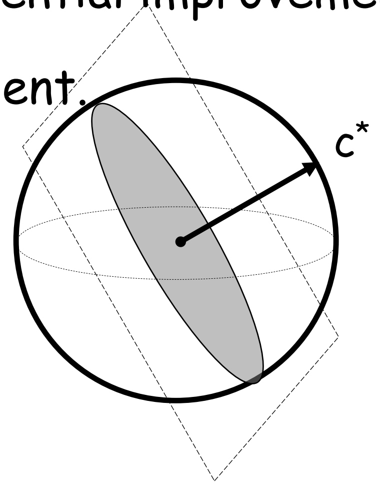
A^2 the first algorithm which is robust to noise.

[Balcan-Beygelzimer-Langford, ICML'06] [Balcan-Beygelzimer-Langford, JCSS'08]

"Region of disagreement" style: Pick a few points at random from the current region of disagreement, query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

Guarantees for A^2 [BBL'06,'08]:

- It is safe (never worse than passive learning) & exponential improvements.
  - C – thresholds, low noise, exponential improvement.
  - C - homogeneous linear separators in R^d, D - uniform, low noise, only $d^2 \log(1/\varepsilon)$ labels.



A lot of subsequent work.

[Hanneke'07, DHM'07, Wang'09 , Fridman'09, Kolt10, BHW'08, BHLZ'10, H'10, Ailon'12, …]

General guarantees for A^2 Agnostic Active Learner

"Disagreement based": Pick a few points at random from the current region of uncertainty, query their labels, throw out hypothesis if you are statistically confident they are suboptimal. [BBL'06]

How quickly the region of disagreement collapses as we get closer and closer to optimal classifier

Guarantees for A^2 [Hanneke'07]:

Disagreement coefficient $\theta_{c^*} = \sup_{r \geq \eta+\epsilon} \frac{\Pr(DIS(B(c^*, r)))}{r}$

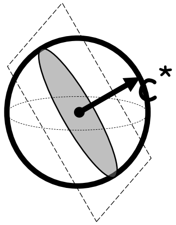
Theorem

$$m = \left(1 + \frac{\eta^2}{\epsilon^2}\right) VCdim(C)\theta_{c^*}^2 \log(\frac{1}{\varepsilon})$$

labels are sufficient s.t. with prob. $\geq 1-\delta$ output h with $err(h) \leq \eta+\epsilon$.

Realizable case: $m = VCdim(C)\theta_{c^*} \log(\frac{1}{\varepsilon})$

Linear Separators, uniform distr.: $\theta_{c^*} = \sqrt{d}$



General guarantees for A² Agnostic Active Learner

"Disagreement based": Pick a few points at random from the current region of uncertainty, query their labels, throw out hypothesis if you are statistically confident they are suboptimal. [BBL'06]

How quickly the region of disagreement collapses as we get closer and closer to optimal classifier

Guarantees for A² [Hanneke'07]:

Disagreement coefficient $\theta_{c^*} = \sup_{r \geq \eta+\epsilon} \frac{\Pr(DIS(B(c^*,r)))}{r}$

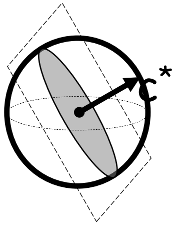
Theorem

$$m = \left(1 + \frac{\eta^2}{\epsilon^2}\right) VCdim(C)\theta_{c^*}^2 \log(\frac{1}{\varepsilon})$$

labels are sufficient s.t. with prob. $\geq 1-\delta$ output h with $err(h) \leq \eta+\epsilon$.

Realizable case: $m = VCdim(C)\theta_{c^*} \log(\frac{1}{\varepsilon})$

Linear Separators, uniform distr.: $\theta_{c^*} = \sqrt{d}$



Disagreement Based Active Learning

"Disagreement based " algos: query points from current region of disagreement, throw out hypotheses when statistically confident they are suboptimal.

- Generic (any class), adversarial label noise.
- Computationally efficient for classes of small VC-dimension

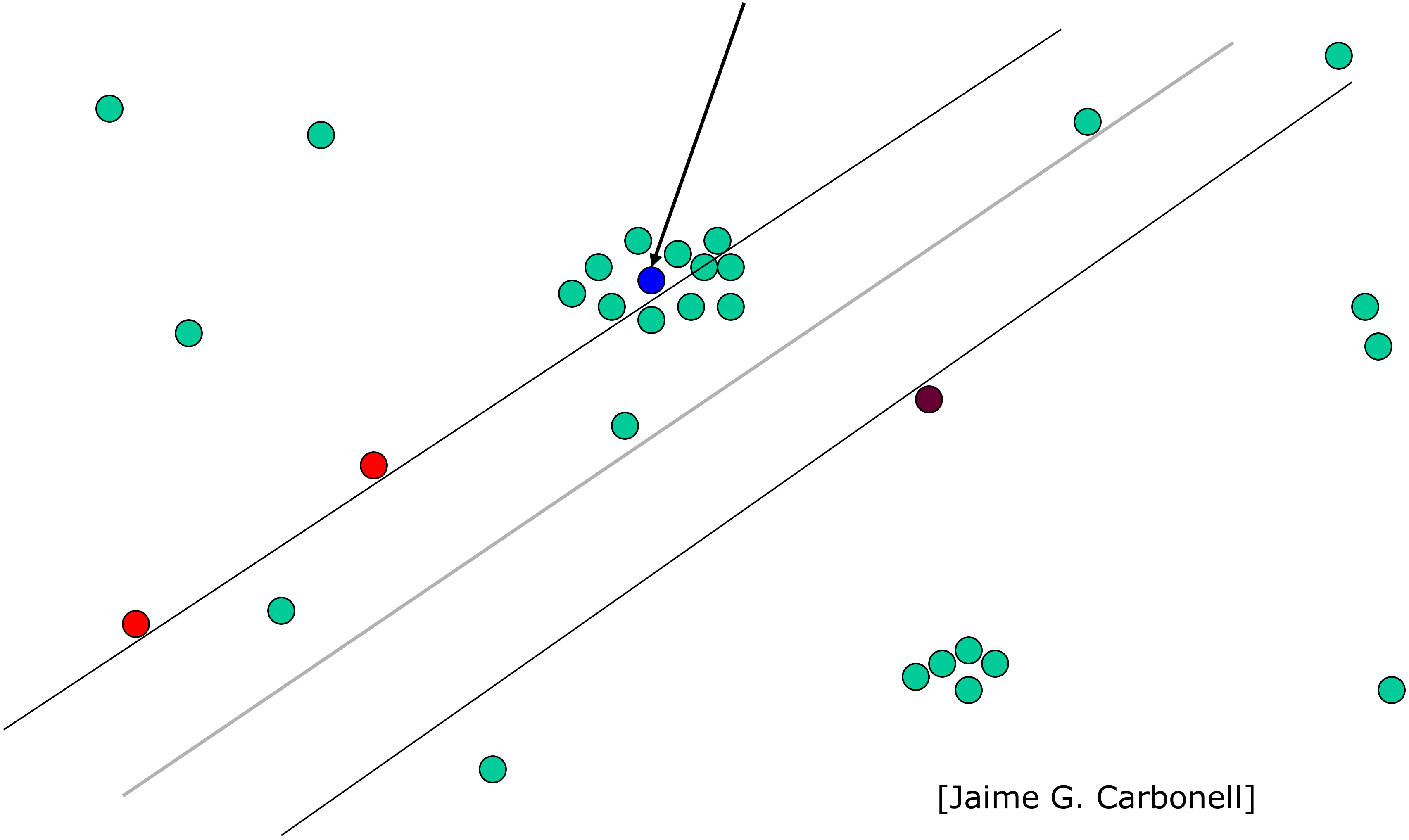
Still, could be suboptimal in label complex & computationally inefficient in general.

Lots of subsequent work trying to make is more efficient computationally and more aggressive too: [Hanneke07, DasguptaHsuMontleoni'07, Wang'09 , Fridman'09, Koltchinskii10, BHW'08, Beygelzimer-Hsu-LangfordZhang'10, Hsu'10, Ailon'12, …]

Other Interesting ALTechniques used in Practice

Interesting open question to analyze under what conditions they are successful.

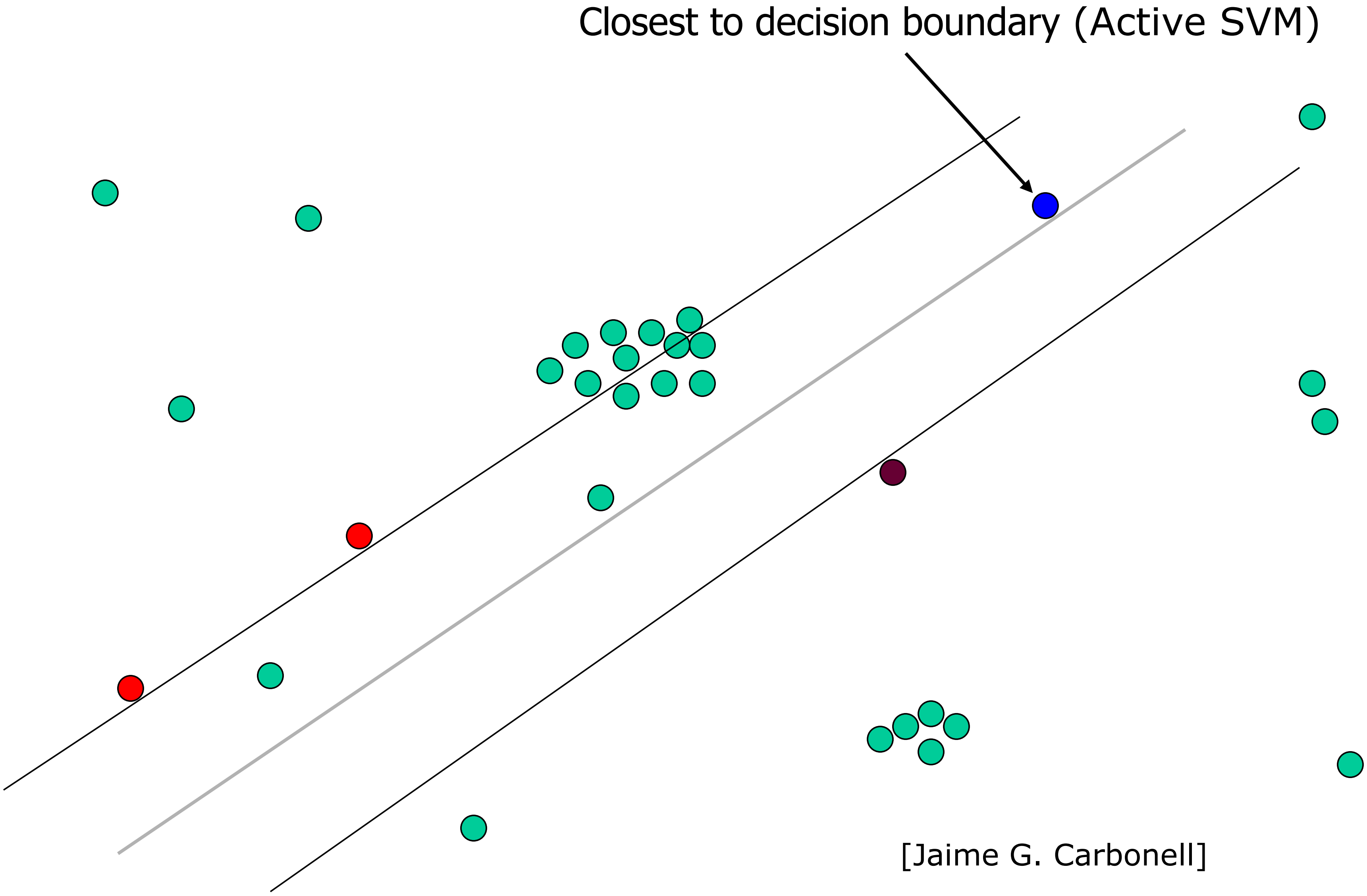
Density-Based Sampling



Centroid of largest unsampled cluster

[Jaime G. Carbonell]

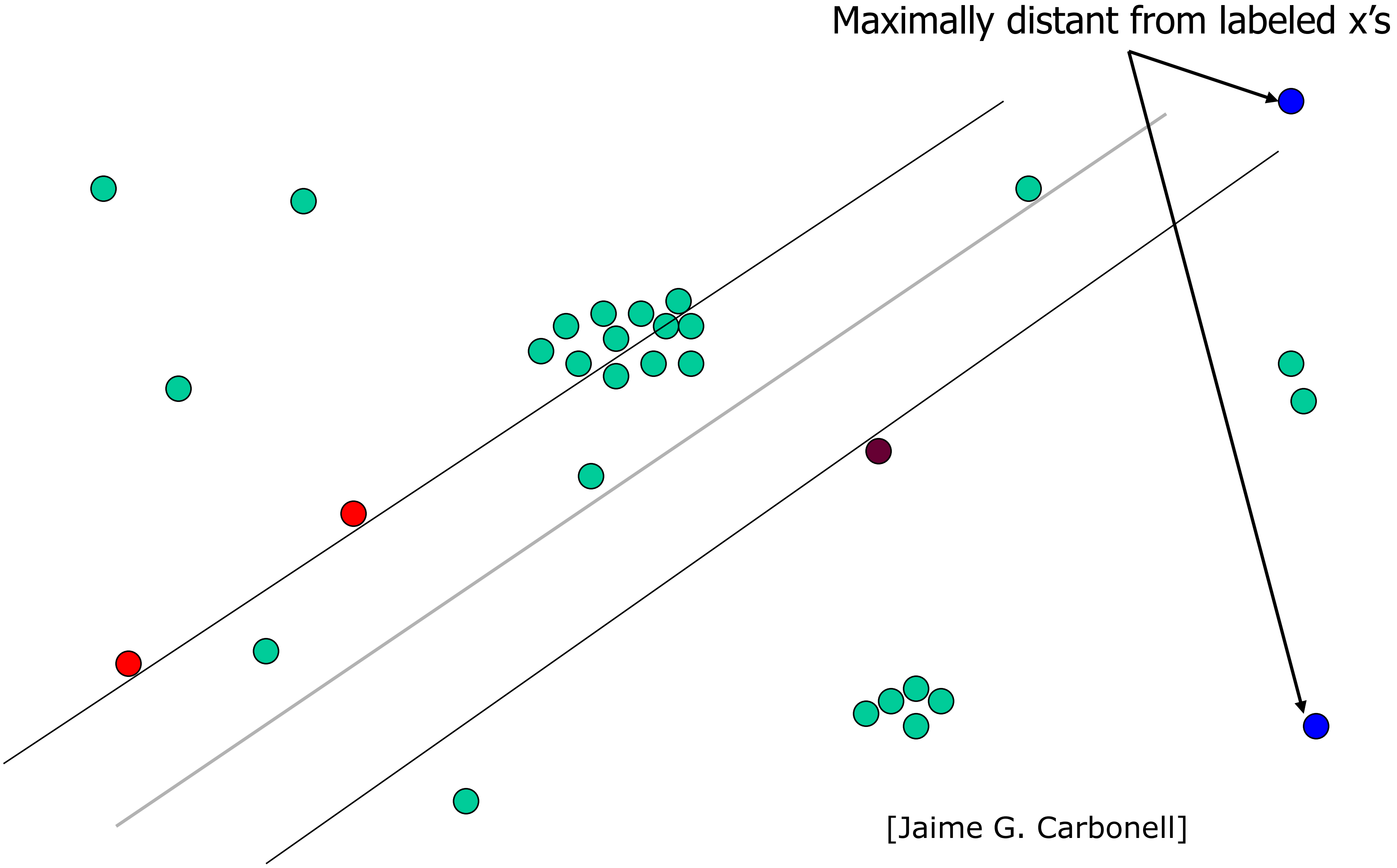
Uncertainty Sampling



Closest to decision boundary (Active SVM)

[Jaime G. Carbonell]

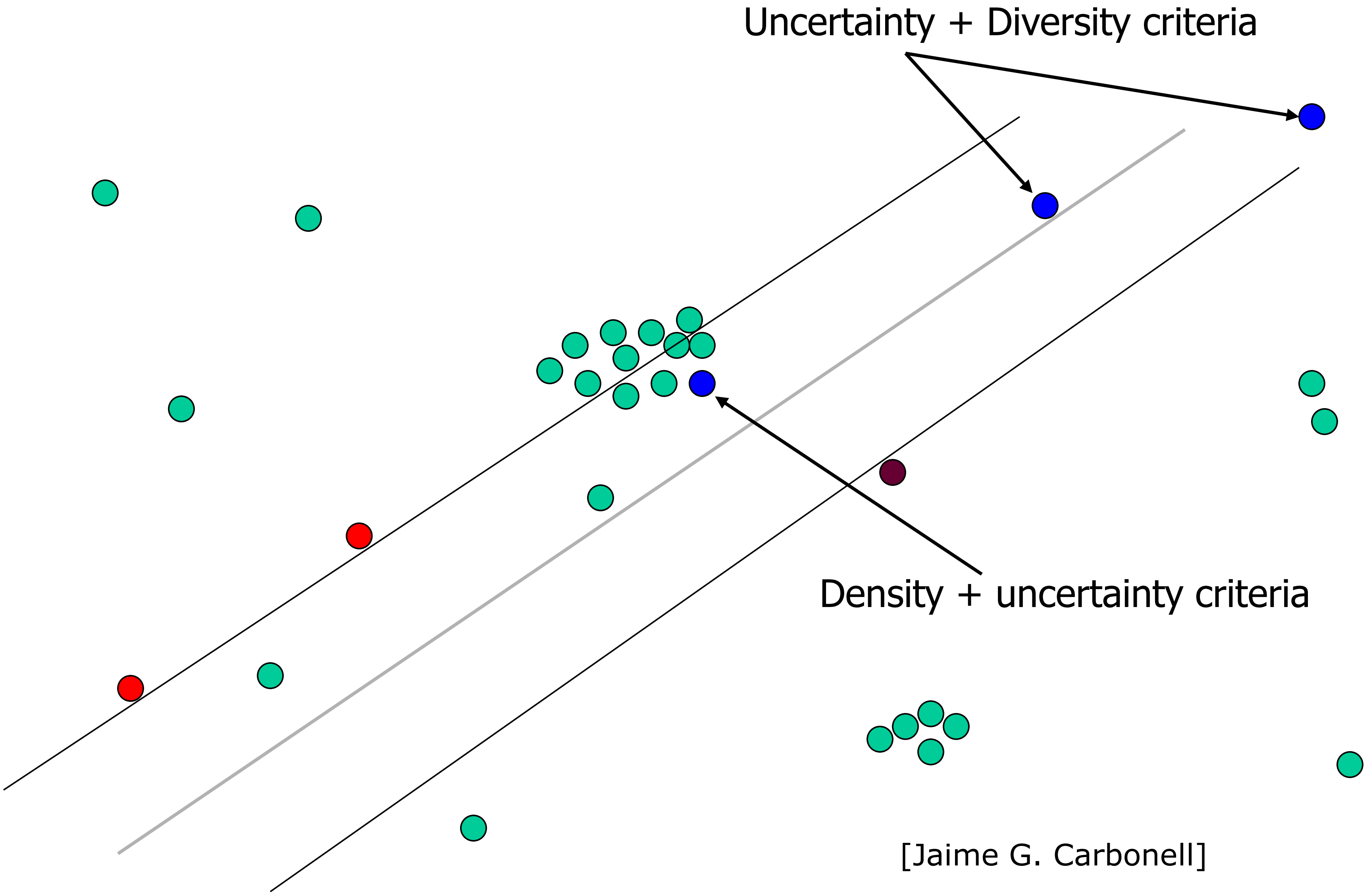
Maximal Diversity Sampling



Maximally distant from labeled x’s

[Jaime G. Carbonell]

Ensemble-Based Possibilities



Uncertainty + Diversity criteria

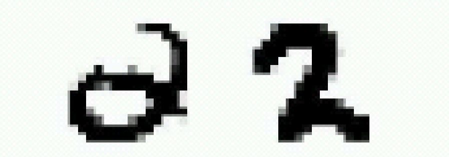
Density + uncertainty criteria

[Jaime G. Carbonell]

Graph-based Active and Semi-Supervised Methods

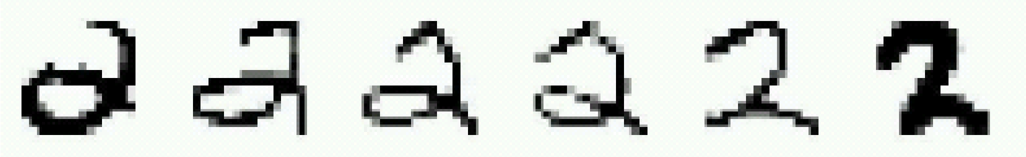
Graph-based Methods

- Assume we are given a pairwise similarity fnc and that very similar examples probably have the same label.
- If we have a lot of labeled data, this suggests a Nearest-Neighbor type of algorithm.
- If you have a lot of unlabeled data, perhaps can use them as “stepping stones”.



not similar

E.g., handwritten digits [Zhu07]:

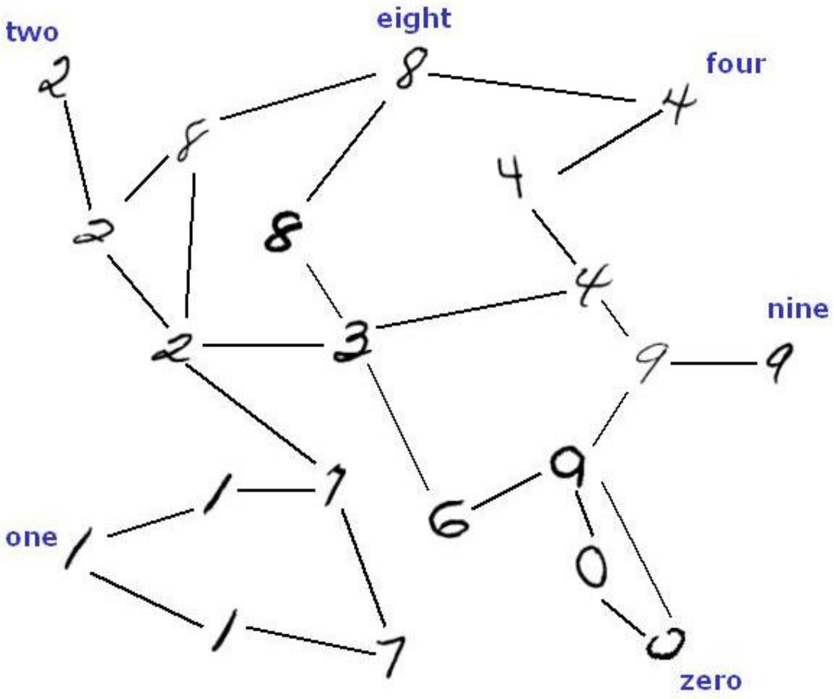


‘indirectly’ similar
with stepping stones

Graph-based Methods

Idea: construct a graph with edges between very similar examples.

Unlabeled data can help "glue" the objects of the same class together.

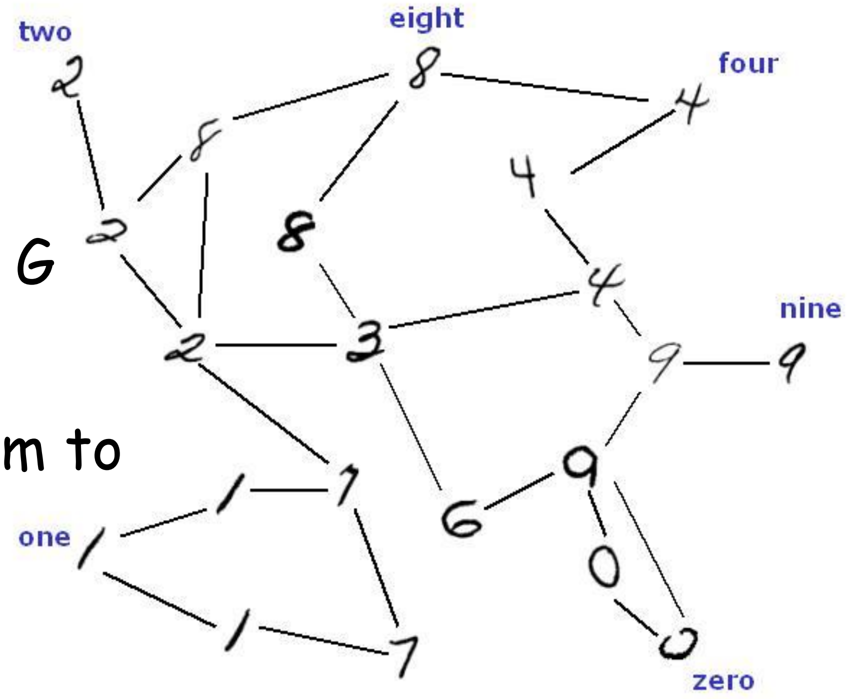


Graph-based Methods

Often, transductive approach. (Given L + U, output predictions on U). Are alllowed to output any labeling of $L \cup U$.

Main Idea:

- Construct graph G with edges between very similar examples.
- Might have also glued together in G examples of different classes.
- Run a graph partitioning algorithm to separate the graph into pieces.



Several methods:

- Minimum/Multiway cut [Blum-Chawla01]
- Minimum "soft-cut" [Zhu-Ghahramani-Lafferty'03]
- Spectral partitioning
- …

SSL using soft cuts

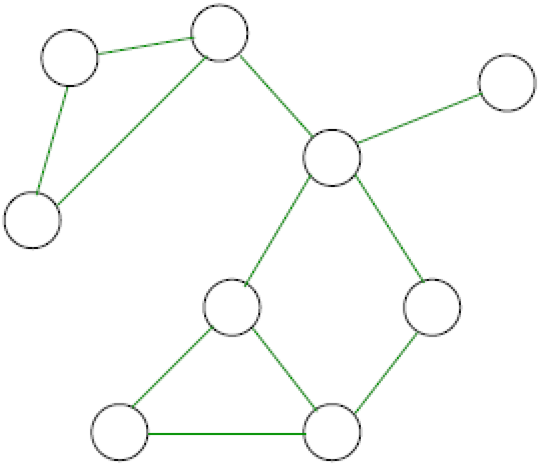
[Zhu-Ghahramani-Lafferty'03]

Solve for label function $f(x) \in [0,1]$ to minimize:

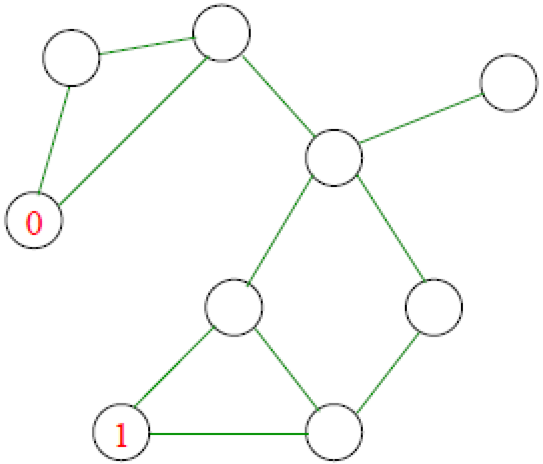
$$J(f) = \underbrace{\sum_{edges\ (i,j)} w_{ij}\big(f(x_i) - f(x_j)\big)^2}_{\substack{\text{Similar nodes get} \\ \text{similar labels} \\ \text{(weighted similarity)}}} + \underbrace{\sum_{x_i \in L} \lambda(f(x_i) - y_i)^2}_{\substack{\text{Agreement with labels} \\ \text{(agreement not strictly enforces)}}}$$

Active learning with label propagation

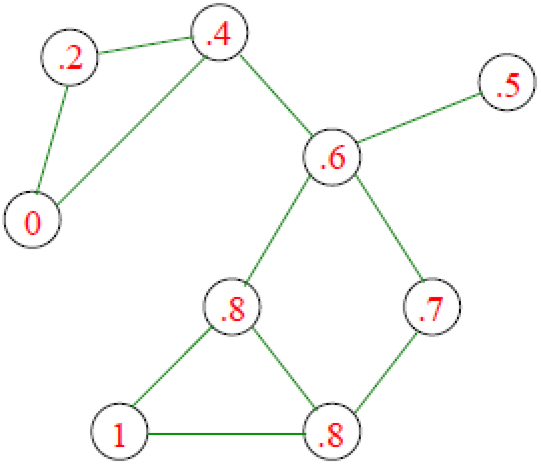
(1) Build neighborhood graph



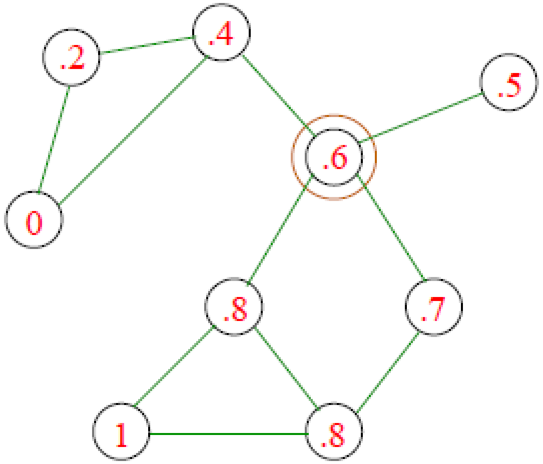
(2) Query some random points



(3) Propagate labels (using soft-cuts)



(4) Make query and go to (3)



How to choose which node to query?

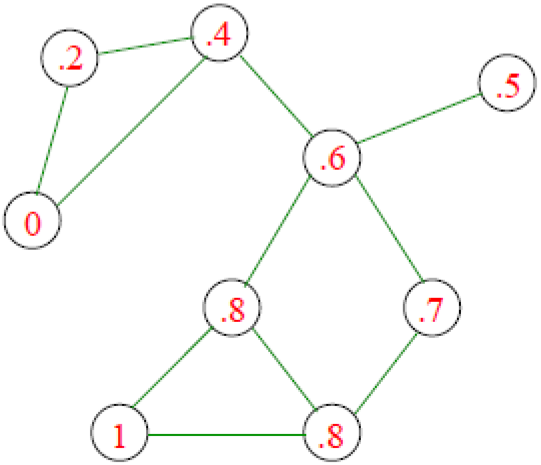
Active learning with label propagation

One natural idea: query the most uncertain point.

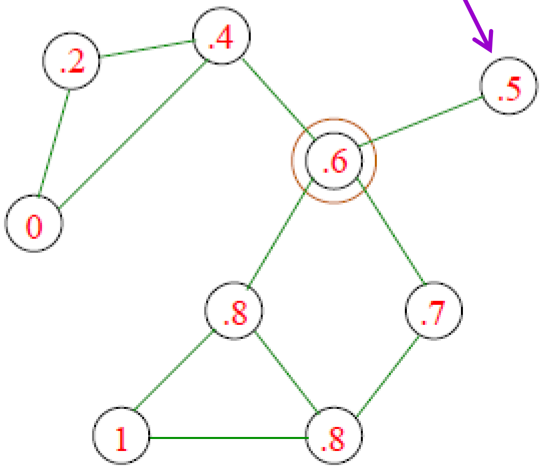
But this has only one edge. Query won't have much impact!

(even worse: a completely isolated node)

(3) Propagate labels (using soft-cuts)



(4) Make query and go to (3)



Active learning with label propagation

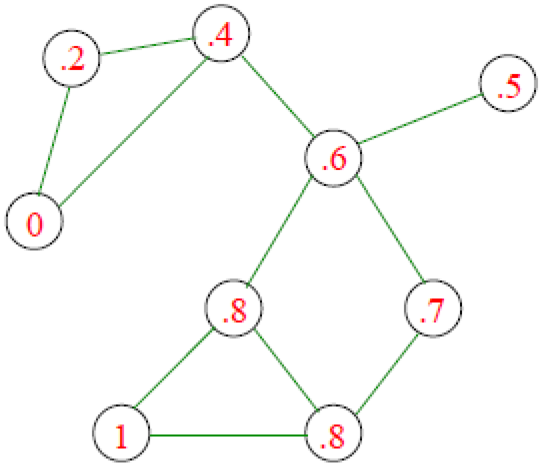
Instead, use a 1-step-lookahead heuristic:

- For a node with label p, assume that querying will have prob p of returning answer 1, $1-p$ of returning answer 0.
- Compute "average confidence" after running soft-cut in each case:

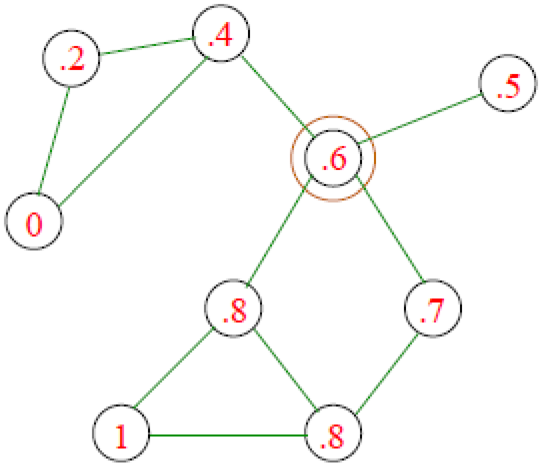
$$p\frac{1}{n}\sum_{x_i}\max\left(f_1(x_i), 1-f_1(x_i)\right)+(1-p)\frac{1}{n}\sum_{x_i}\max\left(f_0(x_i), 1-f_0(x_i)\right)$$

- Query node s.t. this quantity is highest (you want to be more confident on average).

(3) Propagate labels (using soft-cuts)

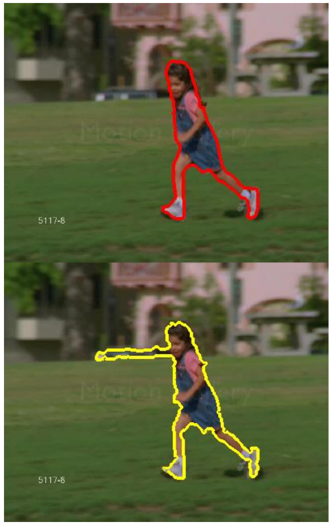


(4) Make query and go to (3)



Active Learning with Label Propagation in Practice

- Does well for Video Segmentation (Fathi-Balcan-Ren-Regh, BMVC 11).



Discussion, Open Directions

- Active learning: important modern learning paradigm.
  - could be really helpful, could provide exponential improvements in label complexity (both theoretically and practically)!
- Common heuristics (e.g., those based on uncertainty sampling). Need to be very careful due to sampling bias.
- Very general sample complexity results, arbitrary concept spaces, high dimensional cases via disagreement based schemes.

Discussion, Open Directions

- Active learning: important modern learning paradigm.
- Very general sample complexity results, arbitrary concept spaces, high dimensional cases.
- Localization developed for label efficiency also useful for handling adversarial examples. [Awasthi-Balcan-Long STOC 2014 & JACM'17]

Open Directions

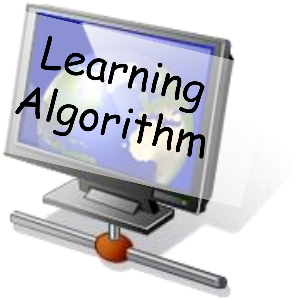
- Active deep learning.
- More general interactions with the expert.

E.g., Local Algorithms for Interactive Clustering.
[Awasthi-Balcan-Voevodski, ICML 2014 & JMLR 2017]

Important direction: richer interactions with the expert.

Better Accuracy

Fewer queries



Natural interaction