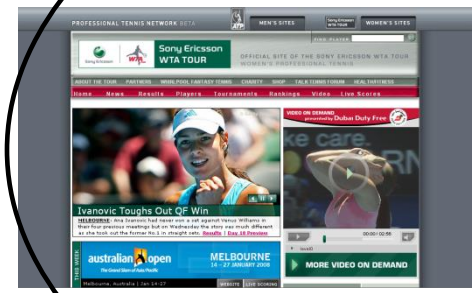
Interactive Machine Learning

Maria-Florina Balcan

Machine Learning

Image Classification

Document Categorization



Speech Recognition

Protein Classification

Branch Prediction

Fraud Detection

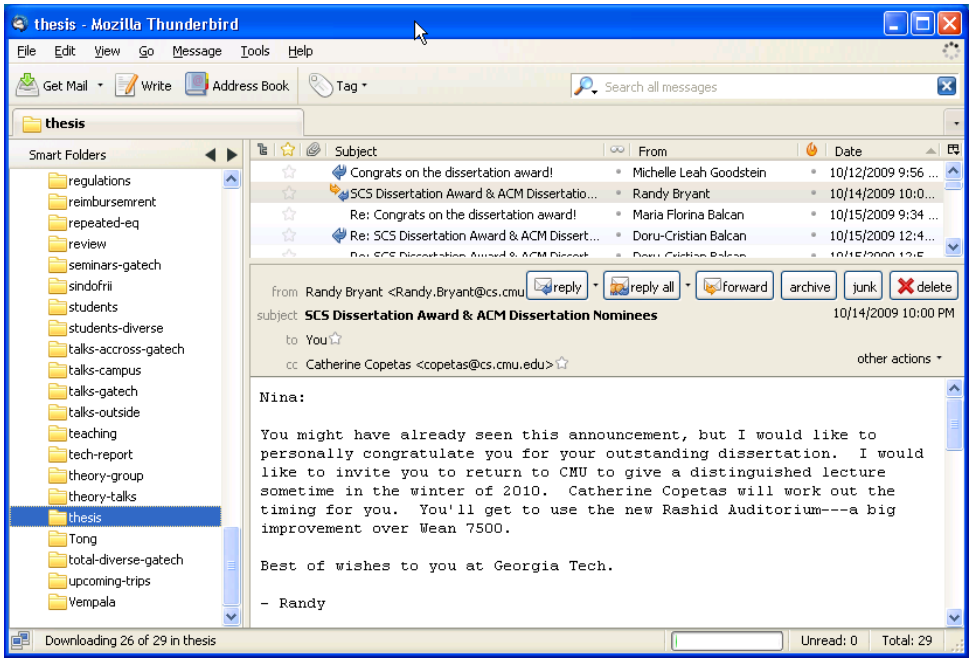
Spam Detection

Supervised Classification

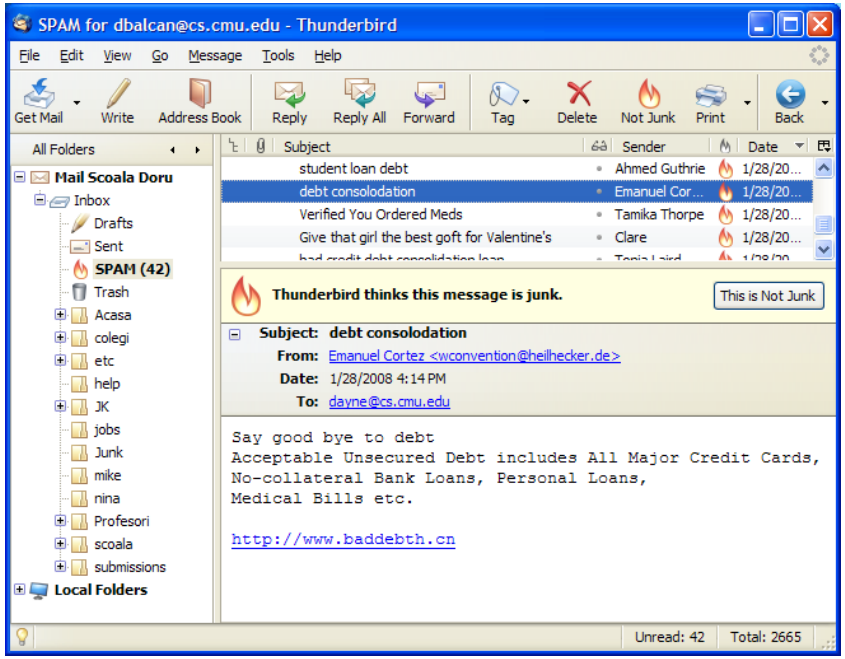
Example: Decide which emails are spam and which are important.

Supervised classification

Not spam



spam



<u>Goal</u>: use past emails to produce good rule for future data.

Example: Supervised Classification

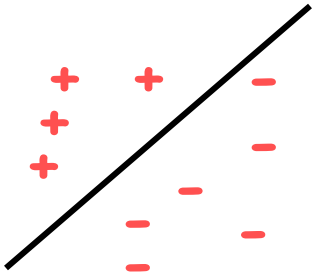
Represent each message by features. (e.g., keywords, spelling, etc.)

	"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.

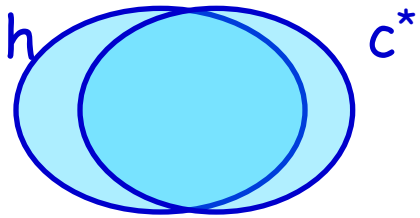
Supervised Learning Formalization

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.

Supervised Learning

Infinite C, realizable

Theorem

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

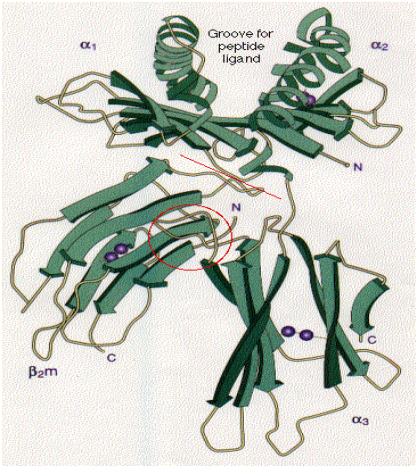
labeled examples are sufficient s.t. with prob. at least $1-\delta$, all $h \in C$ with $\widehat{err}(h) = 0$ have $err(h) \leq \varepsilon$.

- Lots of work on tighter bounds [e.g., Haussler,Littlestone,Wartmuth'94'; Bartlett,Bousquet,Mendelson'05; Gine and Koltchinski'06].
- Lower bounds that match upper bounds on worst case distr.
- Sample complex. pretty well understood up to pesky gaps
- Powerful algorithms as well (AdaBoost & SVM)

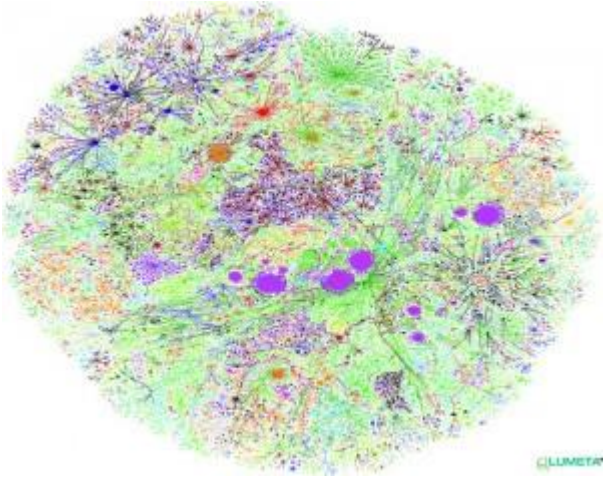
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



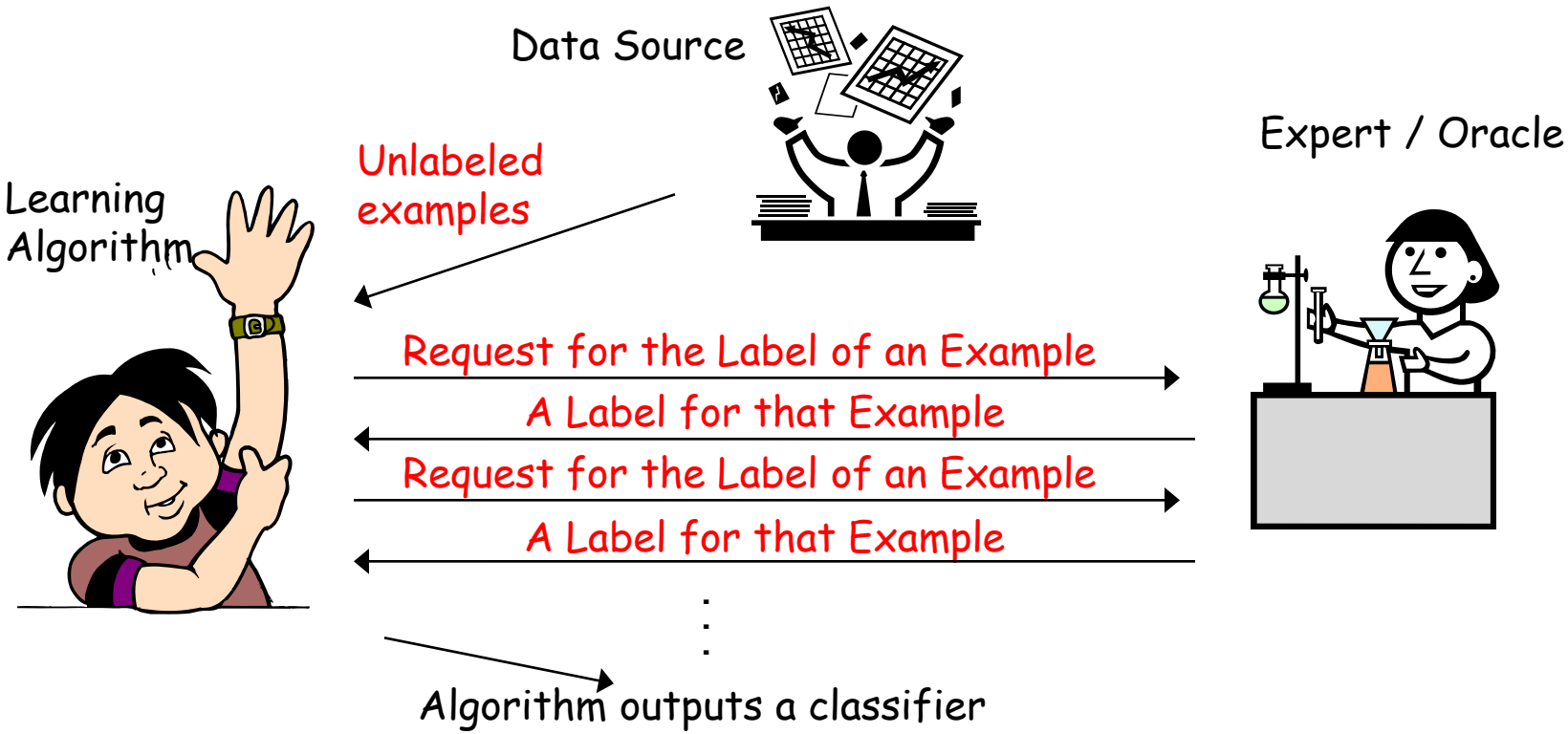
Protein sequences



Billions of webpages

Images

Active Learning



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

Active Learning

Do unlabeled data, interaction help?

When, why, by how much?

- Foundations lacking a few years ago.
- Significant progress recently
  - Mostly on sample complexity.
- Still lots of exciting open questions.

Outline of My Talk

- Brief history, disagreement based active learning.
- Aggressive active learning of linear separators.
- Interactive learning more generally.

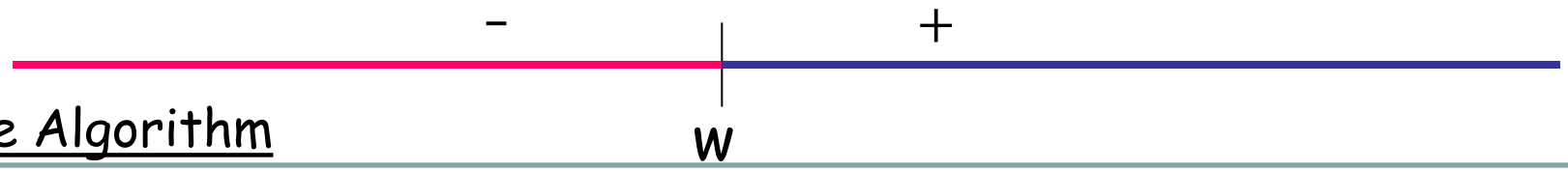
Interactive learning

- Passive learning (sample complexity & algorithms)
- Distributed learning (communication efficient protocols)

Can adaptive querying help? Noise-Free Case

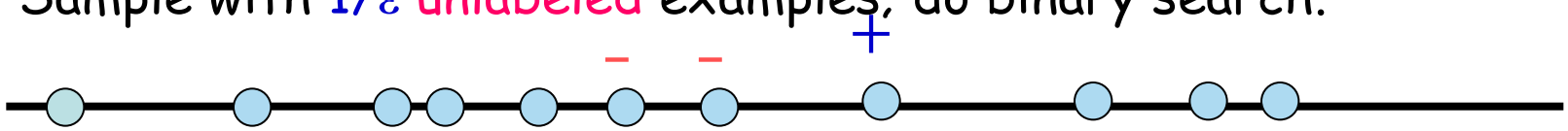
- Useful in practice, e.g., Active SVM [Tong & Koller, ICML '99]

- Canonical theoretical example [CAL92, Dasgupta04]



Active Algorithm

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



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

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

Can adaptive querying help? Noise-Free Case

- Useful in practice, e.g., Active SVM [Tong & Koller, ICML '99]

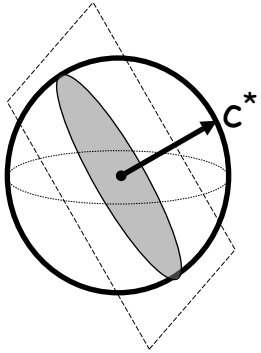
- Canonical theoretical example [CAL92, Dasgupta04]

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

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

- Homogeneous linear sep, uniform distribution
  - QBC [Freund et al., '97]
  - Active Perceptron [Dasgupta, Kalai, Monteleoni '05]

Exponential improvement.



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

When Active Learning Helps. Agnostic case

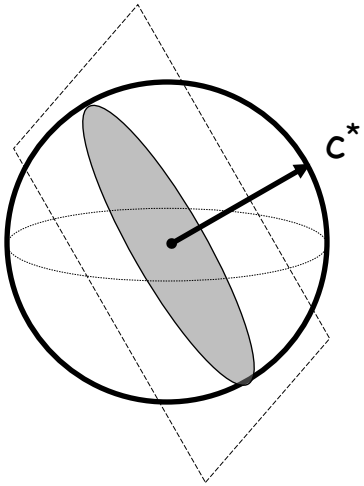
A^2 the first algorithm which is robust to noise.

[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 uncertainty, query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

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

- Fall-back & 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.

E.g., Hanneke'07, DasguptaHsuMonteleoni'07, Wang'09, Friedman'09, Koltchinskii10, BalcanHannekeWortman'08, BeigelzimerHsuLangfordZhang'10, Hanneke'10, El-Yaniv-Wiener'12, Minsker'12, AilonBegleiterExra'12….

A^2 first active algo robust to noise [BBL'06,'08]

"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.

Guarantees for A^2 [Hanneke'07]:

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

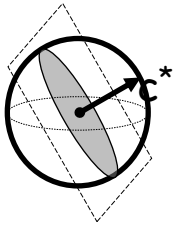
Theorem

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

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.: $\theta_{c^*} = \sqrt{d}$



When Active Learning Helps

Lots of exciting activity in recent years.

- Several specific analyses for r le case. E.g., linear separators, uniform dist
  - QBC [Freu
  - Active ptron [Dasgupta, Kalai, Monteleoni '05]

Very Specific.

- Generic algos that work even in the agne under various noise conditions
  - A^2 [Balcan, Beygelzimer, [Hanneke'10] [Wang'09]
  - DKM algo [D [Hanneke '10]
  - Koltc rchinksi'10]

Typically suboptimal in query complexity.

Aggressive Active Learning

This talk: C - homogeneous linear sets in R^d, D – logconcave

- Realizable: exponential improvement, only $O(d \log 1/\varepsilon)$ labels to find a hypothesis with error ε. [Bounded noise].
- Tsybakov noise: polynomial improvement.

Optimal bounds

[Balcan, Broder, Zhang, COLT'07] ; [Balcan,Long, COLT'13]

Log-concave distributions: log of density fnc concave

- wide class: includes uniform distr. over any convex set, Gaussian distr., Logistic, etc
- played a major role in sampling, optimization, integration, learning [LV'07, KKMS'05,KLT'09]

Margin Based Active-Learning, Realizable Case

Algorithm

Draw m_1 unlabeled examples, label them, add them to $W(1)$.

iterate $k=2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
  - $W(k)=W(k-1)$.
  - sample m_k unlabeled samples x satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
  - label them and add them to $W(k)$.

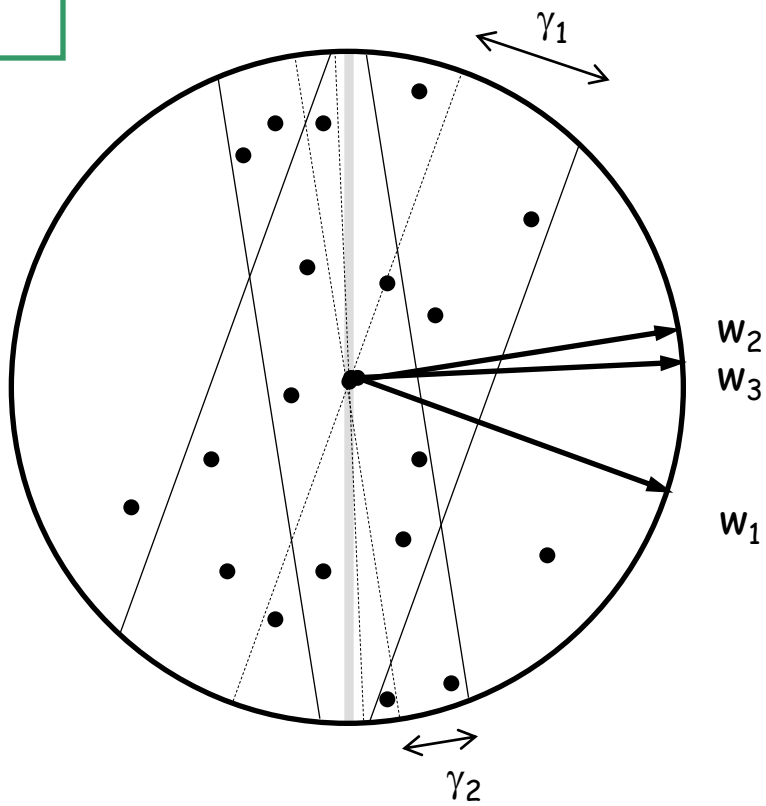
end iterate

Margin Based Active-Learning, Realizable Case

Draw m_1 unlabeled examples, label them, add them to $W(1)$.

iterate $k = 2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
- $W(k)=W(k-1)$.
- sample m_k unlabeled samples x satisfying $|w_{k-1}^T \cdot x| \leq \gamma_{k-1}$
- label them and add them to $W(k)$.



Margin Based Active-Learning, Realizable Case

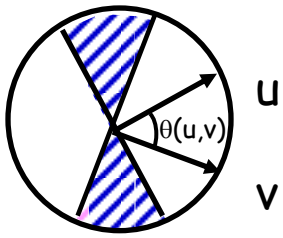
Theorem P_X log-concave in R^d.

If $\gamma_k = O\left(\frac{c}{2^k}\right)$ and $m_k = O\left(d + \log\log(1/\epsilon)\right)$ then after $s = \log\left(\frac{1}{\epsilon}\right)$ iterations w_s has error $\leq \varepsilon$.

Linear Separators, Log-Concave Distributions

Fact 1

$$d(u,v) \approx \frac{\theta(u,v)}{\pi}$$

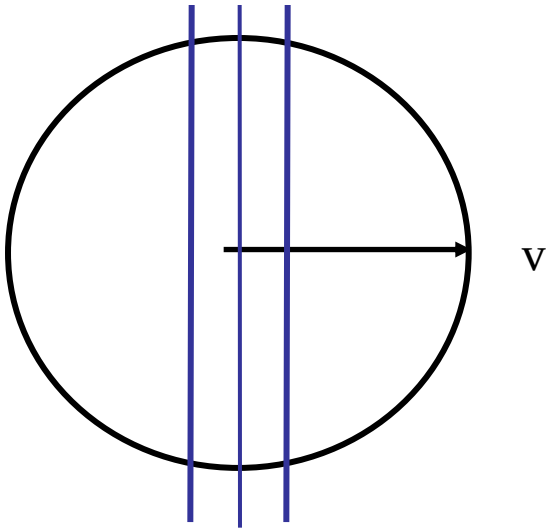


Proof idea:

- project the region of disagreement in the space given by u and v
- use properties of log-concave distributions in 2 dimensions.

Fact 2

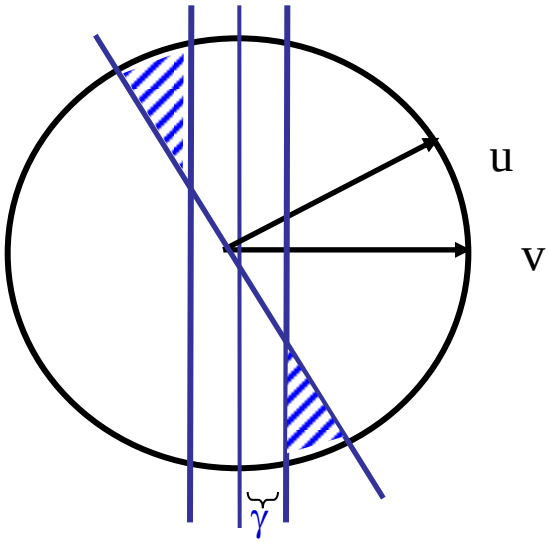
$$\Pr_x\left[|v \cdot x| \leq \gamma\right] \leq \gamma.$$



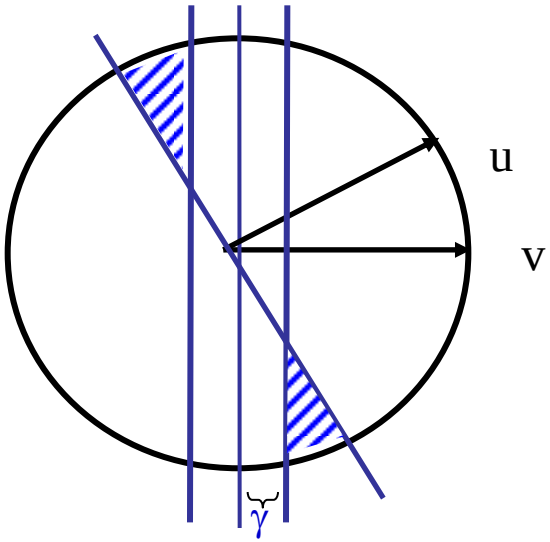
Linear Separators, Log-Concave Distributions

$\boxed{\text{Fact 3}}$ If $\theta(u, v) = \beta$ and $\gamma = C\beta$

$$\Pr_x\left[(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma\right] \leq \frac{\beta}{4}.$$



Linear Separators, Log-Concave Distributions



$\boxed{\text{Fact 3}}$ If $\theta(u,v) = \beta$ and $\gamma = C\beta$

$$\Pr_x[\overbrace{(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma}^{E}] \leq \frac{\beta}{4}.$$

Proof idea:

- project the region of disagreement in the space given by u and v
- Note that each x in E has $||x|| \geq \gamma/\beta$ $= c_2$

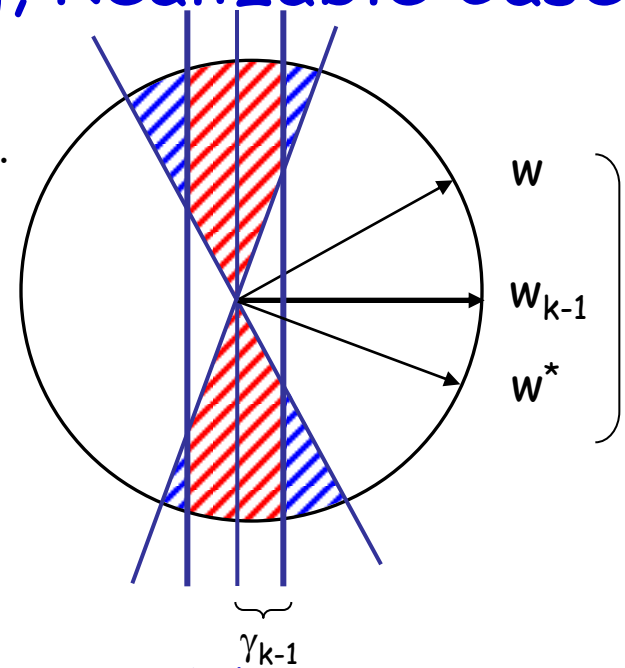
$$\Pr_x[x \in E] = \sum_{i=1}^{\infty} \Pr[E \cap (B((i+1)c_2) - B(ic_2))]$$

$$\leq C\beta(i+1)^2 \exp[-Ci]$$

Margin Based Active-Learning, Realizable Case

iterate $k=2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
- $W(k)=W(k-1)$.
- sample m_k unlabeled samples x satisfying $|w_{k-1}^T \cdot x| \leq \gamma_{k-1}$
- label them and add them to $W(k)$.



Proof Idea

Induction: all w consistent with $W(k)$ have error $\leq 1/2^k$;

so, w_k has error $\leq 1/2^k$.

For $\gamma_k = O\left(\frac{c}{2^k}\right)$

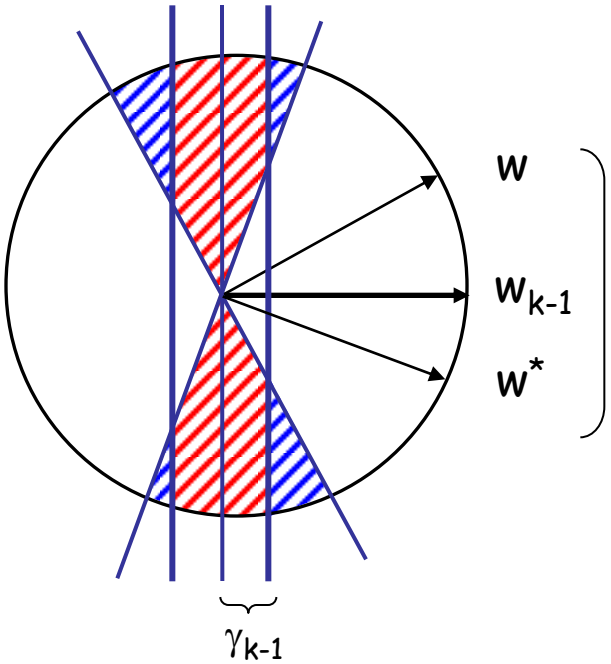
$$\mathrm{err}(w) = \overbrace{\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1})}^{\leq 1/2^{k+1}} + \Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1})$$

Proof Idea

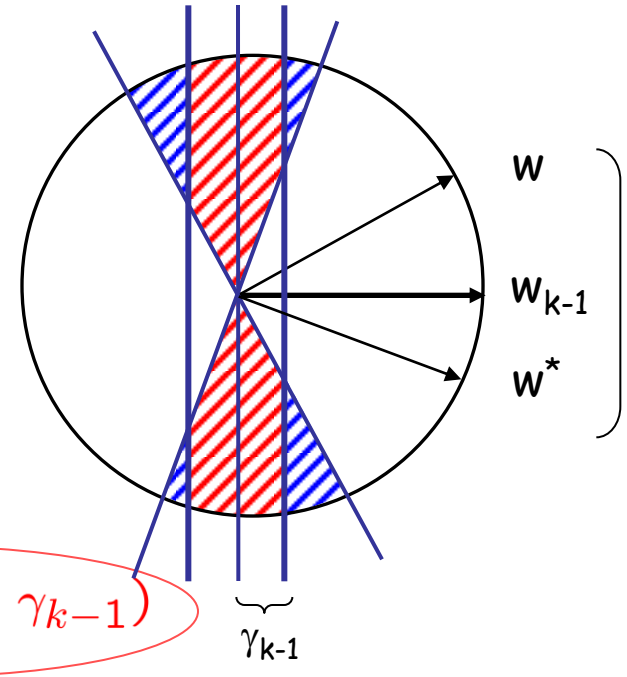
Under the uniform distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$

$$\mathrm{err}(w) = \overbrace{\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1})}^{\leq 1/2^{k+1}} +$$

$$\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1})$$



Proof Idea



Under the uniform distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$

$$\mathrm{err}(w) = \underbrace{\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1})}_{\leq 1/2^{k+1}} +$$

$$\Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \underbrace{\Pr(|w_{k-1} \cdot x| \leq \gamma_{k-1})}_{\leq C\gamma_{k-1}}.$$

Enough to ensure

$$\Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C_1$$

Can do with only $m_k = O(d + \log\log(1/\epsilon))$ labels.

Passive Learning

Theorem

Any passive learning algo that outputs w consistent with

$\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$ examples, satisfies err(w) $\leq \epsilon$, with proab. 1-δ.

- Solves open question for the uniform distr. [Long'95,'03], [Bshouty'09]
- First tight bound for poly-time PAC algos for an infinite class of fns under a general class of distributions. [Ehrenfeucht et al., 1989; Blumer et al., 1989]

Passive Learning

Theorem

Any passive learning algo that outputs w consistent with

$\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$ examples, satisfies err(w) $\leq \epsilon$, with proab. 1-δ.

High Level Idea

- Run algo online, use the intermediate w to track the progress
- Performs well even if it periodically builds w using some of the examples, and only uses borderline cases for preliminary classifiers for further training.
- Carefully distribute δ, allow higher prob. of failure in later stages [once w is already pretty good, it takes longer to get examples that help to further improve it]

Passive Learning

Theorem

Any passive learning algo that outputs w consistent with

$\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$ examples, satisfies err(w) $\leq \epsilon$, with proab. 1-δ.

High Level Idea

$$m_k = C_2\,(d + \log((1 + s - k)/\delta)) \quad b_k = C_1/2^k$$

$$\sum_{k=1}^{s} 2^k\,(d + \log((1 + s - k)/\delta))) =$$

$$O(2^s\,(d + \log(1/\delta)) + \sum_{k=1}^{s} 2^k \log(1 + s - k)$$

$O(1/\epsilon)$

Margin Based AL, Bounded Noise

$\boxed{\text{Guarantee}}$ Assume P_X log-concave in R^d.

Assume that $|P(Y=1|x)-P(Y=-1|x)| \geq \beta$ for all x.

Assume w^* is the Bayes classifier.

Then $\beta \cdot \frac{\theta(w,w^*)}{\pi} \leq \mathrm{err}(w) - \mathrm{err}(w^*)$

The previous algorithm and proof extend naturally, and get again an exponential improvement.

Bounded Noise, Exponential improvement

$\boxed{\text{Guarantee}}$ Assume P_X log-concave in R^d

Assume that $|P(Y=1|x)-P(Y=-1|x)| \geq \beta$ for all x.

Assume w^* is the Bayes classifier.

Then $\beta \cdot \frac{\theta(w,w^*)}{\pi} \leq \mathrm{err}(w) - \mathrm{err}(w^*)$

The previous algorithm and proof extend naturally, and get again an exponential improvement.

$$\mathrm{err}(w) - \mathrm{err}(w^*) \leq (\mathrm{err}(w|S_1) - \mathrm{err}(w^*|S_1))\,\Pr(S_1) \;+$$

$$\Pr_X\left[(w \cdot x)(w^* \cdot x) < 0, x \in S_2\right].$$

Margin Based AL, Summary

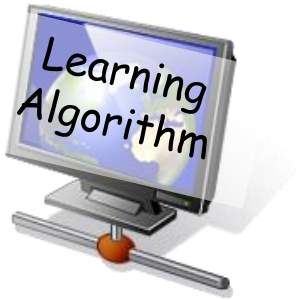
- Extensions to nearly log-concave distributions, noisy settings. Matching Lower Bounds.
- Broadens class of pbs for which AL provides exponential improvement in $1/\epsilon$ (without additional increase on d).
- Interesting implications to passive learning:

  - First tight bound for a poly-time PAC algo for an infinite class of fns under a general class of data distributions. [Balcan-Long, COLT'13]
  - New algo that tolerates malicious noise at significantly better levels than previously known even for passive learning [KKMS'05, KLS'10]; also active (much better label sample complexity). [Awasthi-Balcan-Long'13]

Important direction: richer interactions with the expert.

Better Accuracy

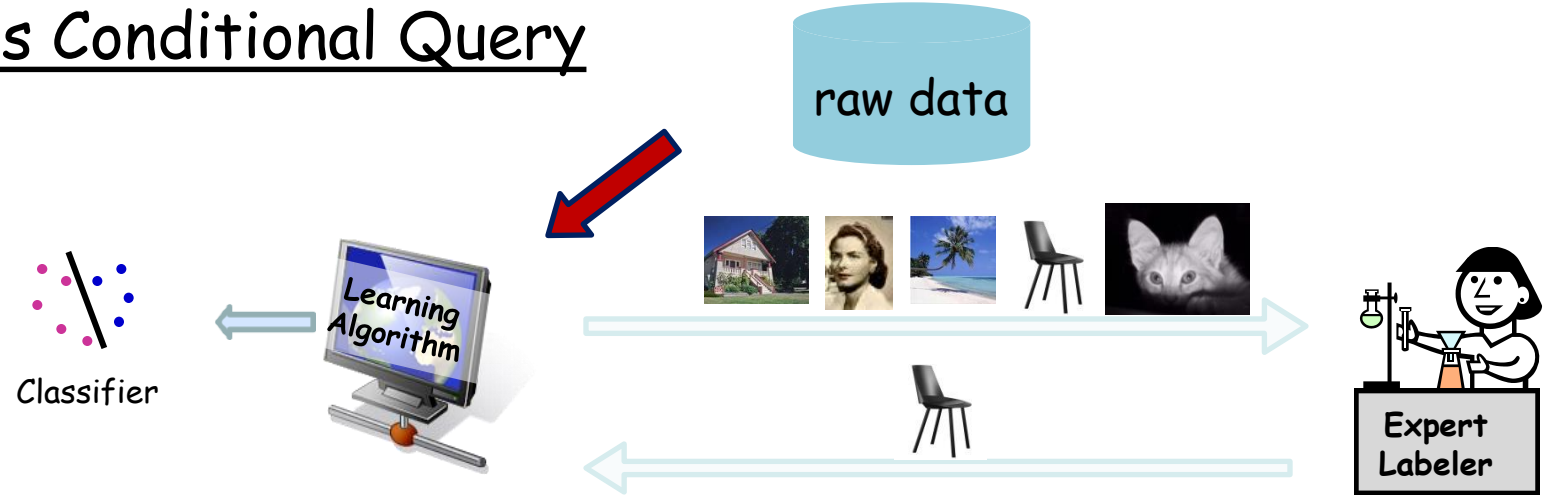
Fewer queries



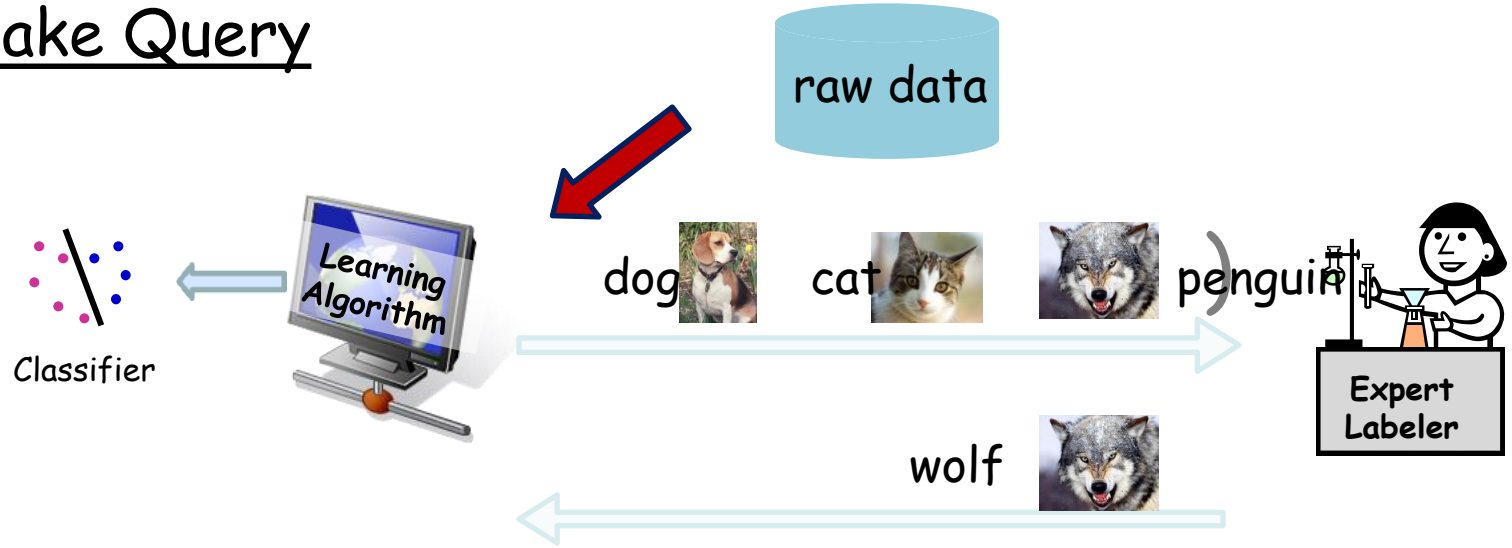
Natural interaction

New Types of Interaction [Balcan-Hanneke COLT'12]

Class Conditional Query

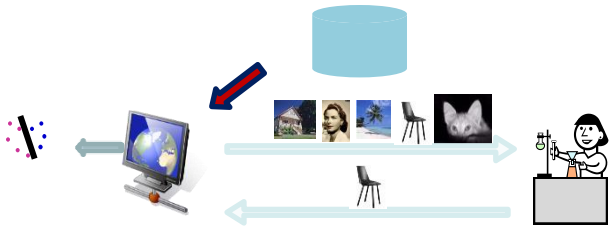


Mistake Query



Class Conditional & Mistake Queries

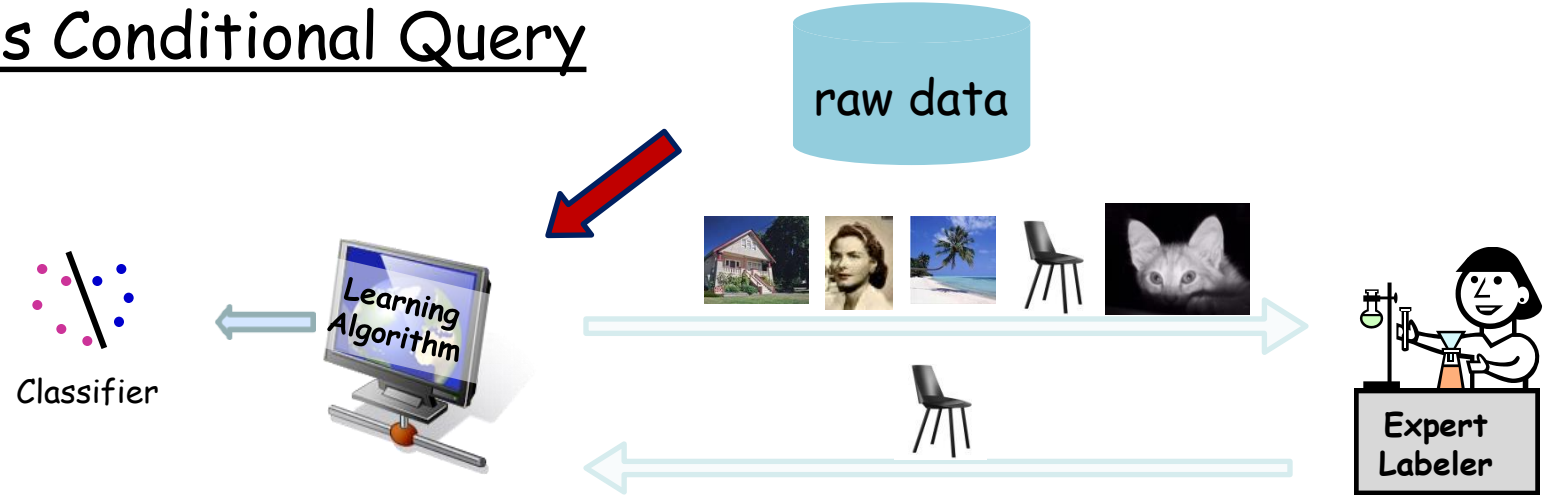
- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.



- Realizable (Folklore): much fewer queries than label requests.

Class Conditional Query

Class Conditional Query



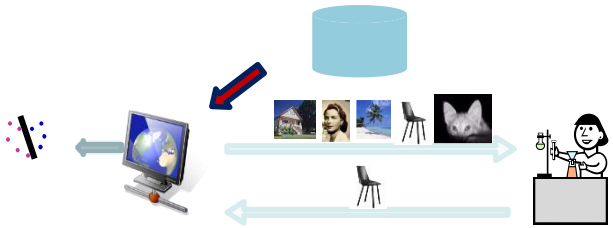
Folklore Result (Realizable Case):

$kd \log\left(\frac{k}{\epsilon}\right)$ queries sufficient to learn an ϵ – accurate classifier, assuming that the target belongs to C of dimension d.

- Run halving algorithm

Class Conditional & Mistake Queries

- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.



- Realizable (Folklore): much fewer queries than label requests.

Balcan-Hanneke, COLT'12

Tight bounds on the # of CCQs to learn with noise (agnostic, bounded noise, one-sided noise).

Formal Model

- X instance space, $Y = \{1,2,\dots,k\}$ label space.
- $D_{\{X,Y\}}$ fixed target distribution, $D_{\{X\}}$ marginal over X.
- An i.i.d sequence (x_1, y_1), $(x_2, y_2), (x_3, y_3), \dots,$ each with distr $D_{\{X,Y\}}$.

- Algo can access x_i values; info about y_i obtained via CCQs.

  CCQ($S = \{x_{i_1}, \dots, x_{i_m}\}$, label l)
  - If $y_{i_j} \neq l$, for all j, expert says "none";
  - otherwise, returns an arbitrary x_{i_j} s.t. $y_{i_j} = l$

- FindMistake(S,h) based on k CCQs
  - For each l, query the set $\{x \in S: h(x) \neq l\}$ for label l
  - If received back an example (x, l), return (x, l)

- C, $\mathrm{Ndim}(C) = d$ (Natarajan dimension)

  largest m s.t. $\exists\, (a_1, b_1, c_1), \dots, (a_m, b_m, c_m) \in X \times Y \times Y$, $b_i \neq c_i$ s.t. $\{b_1, c_1\} \times \cdots \times \{b_m, c_m\} \subseteq \{h(a_1), \dots, h(a_m)\}$

Agnostic Case, Upper Bound

Phase 1: Output h of error $10(\eta + \epsilon)$ with $\tilde{O}\left(kd \log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Input $U = (x_1, x_2, \ldots, x_{ps})$, V- ϵ-cover. $s = \frac{1}{16\eta}$, $N = \log\left(\frac{\log|V|}{\delta}\right)$; $b = 1$

While b.

- Draw $S_1, \ldots, S_N$ of size s uniformly from U.
- For each i, Find-Mistake $(S_i, \mathrm{plur}(V))$. If mistake record it $(\tilde{x}_i, \tilde{y}_i)$.
- If mistakes on more than $N/3$ sets, remove from V every h that makes a mistake on more than $N/9$ examples $(\tilde{x}_i, \tilde{y}_i)$; else $b = 0$.

Output $\mathrm{plur}(V)$.

I.e., eliminate h when makes at least one mistake in some number out of several sets chosen at random from U.

- rather than eliminating h for a mistake, as in halving.

Agnostic Case, Upper Bound

Phase 1: Output h of error $10(\eta + \epsilon)$ with $\tilde{O}\left(kd \log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Input $U = (x_1, x_2, \ldots, x_{ps})$, V- ϵ-cover. $s = \frac{1}{16\eta}$, $N = \log\left(\frac{\log|V|}{\delta}\right)$; $b = 1$

While b.

- Draw $S_1, \ldots, S_N$ of size s uniformly from U.
- For each i, Find-Mistake $(S_i, \mathrm{plur}(V))$. If mistake record it $(\tilde{x}_i, \tilde{y}_i)$.
- If mistakes on more than $N/3$ sets, remove from V every h that makes a mistake on more than $N/9$ examples $(\tilde{x}_i, \tilde{y}_i)$; else $b = 0$.

Output $\mathrm{plur}(V)$.

<u>Proof Sketch</u>

- Best h in V survives (does not make mistakes on too many sets).
- If $\mathrm{err}(\mathrm{plur}(V)) \geq 10\,\eta$, then $\mathrm{plur}(V)$ make mistakes on enough sets; so, a constant fraction of V will make mistakes on more sets than the best classifier & eliminate a constant fraction of V.

Upper Bound, Agnostic Case

$\text{Agnostic}(C, \eta) = \{D_{X,Y}: \inf_{h \in C} \text{err}(h) \leq \eta\}$

η Passive QC

For any class C, $QC(\epsilon, \delta, C, \text{Agnostic}(C, \eta)) = \tilde{\Theta}\left(d\left(\frac{\eta}{\epsilon}\right)^2\right)$ [Const k]

Phase 1: Robust halving algorithm

Output h, $\text{err}(h) \leq 10(\eta + \epsilon)$; only $\tilde{O}\left(kd \log\left(\frac{1}{\epsilon}\right)\right)$ queries.

Phase 2: A simple refining algorithm

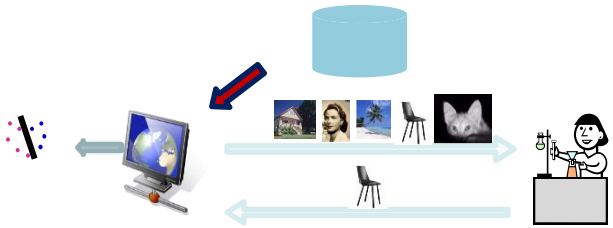
$\tilde{O}\left(kd\left(\frac{\eta}{\epsilon}\right)^2\right)$ queries to turn h into one of error $\eta + \epsilon$

Pick $\tilde{O}\left(d\frac{\eta}{\epsilon^2}\right)$ unlabeled samples, and repeatedly run FindMistake to find all the mistakes h makes.

We recover the true labels of this unlabeled sample and run ERM.

Class Conditional & Mistake Queries

- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.



- Realizable (Folklore): much fewer queries than label requests.

Balcan-Hanneke, COLT'12

Tight bounds on the # of CCQs to learn with noise (agnostic, bounded noise, one-sided noise).

- Adapted to obtain the first agnostic communication efficient in a distributed learning setting. [Balcan-Blum-Fine-Mansour'COLT12]

Discussion, Open Directions

Active Learning

- Rates for aggressive AL with optimal query complexity for general settings.
  - [BL'13], [BBZ'07], aggressive localization in instance space leads to aggressive localization in concept space; can it be generalized?
- Computationally efficient algos for noisy settings.
  (recent progress [BF'13], [ABL'13])

More general queries

- Efficient algorithms for CCQs.
- More general/natural queries.
- Shallow/Deep learning strategies.