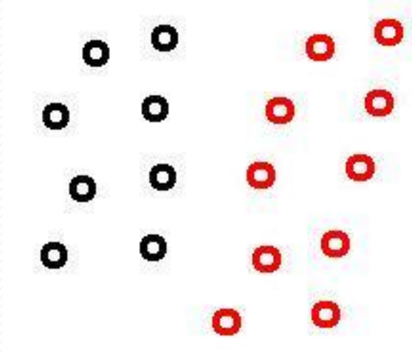
Active Learning Models and Noise

By Sara Stolbach

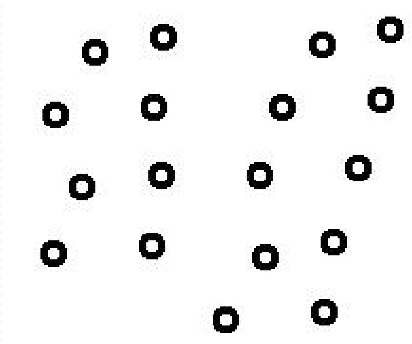
Advanced CLT, Spring 2007

Definition

- In Active Learning the user is given unlabelled examples where it is possible to get any label but it can be costly.
- Pool-Based active learning is when the user can request the label of any example.
- We want to label the examples that **will give us the most information**. i.e. learn the concept in the shortest amount of time.



General Learning Model



Active Learning Model

Pool-Based Active Learning Models

- Bayesian Assumptions - knowledge of a prior upon which the generalization bound is based
  - Query By Committee [F,S,S,T 1997]
- Generalized Binary Search
  - Greedy Active Learning [Dasgupta 2004]
- Opportunistic Priors or algorithmic luckiness
  - a uniform bet over all H leads to standard VC generalization bounds
  - if more weight is placed on a certain hypothesis then it could be excellent if guessed right but worse than usual if guessed wrong,

Query By Committee[F,S,S,T 1997]

QBC Algorithm

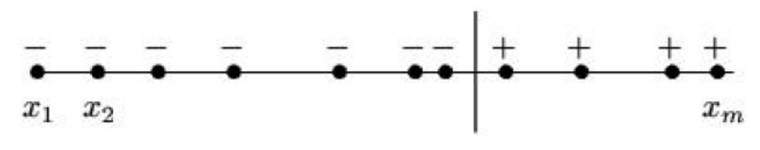
Input: $\epsilon > 0$, $\delta > 0$, **Gibbs**, **Sample**, **Label**
Initialize: $n = 0$, $V_0 = C$
repeat
  Call the **Sample** oracle to get a random instance of x.
  Call **Gibbs** twice to get two predictions p_1 and p_2 for x.
  if $p_1 = p_2$ **then**
    reject the example
  else
    call the **Label**(x) to get $c(x)$, increase n by 1 and set V_n to be all concepts $c' \in V_{n-1}$ where $c'(x) = c(x)$
  end if
until more than t_n consecutive examples are rejected.
Output: the **Gibbs** prediction hypothesis

Query By Committee

- Gibbs Prediction Rule – Gibbs(V,x) predicts the label of example x by randomly choosing $h \in C$ over D, restricted to $V \subset C$, and labeling x according to it.
- Two calls to Gibbs(V,x) can give different predictions.
- It is easy to show that if QBC ever stops then the error of the resulting hypothesis is small with high probability. The real question is will the QBC algorithm stop.
  - It will stop if the number of examples that are rejected between consecutive queries increases with the number of queries (constant improvement)
- The probability of accepting a query or making a prediction mistake is exponentially small compared to the number of queries asked.

Greedy Active Learning[Dasgupta,2004]

- Given unlabeled examples, a simple binary search can be used when d=1 to find the transition from 0 to 1



- Only log m labels are required to infer the rest of the labels.
- Exponential improvement!
- What about in the generalized case? H can classify m points in $O(m^d)$ possibilities; How many labels are needed?
- If binary search were possible, just O(d log m) labels would be needed.

**picture taken from Dasgupta's paper, "Greedy Active Learning"

Greedy Active Learning

- Always ask for the label which most evenly divides the current effective version space.
- The expected number of labels needed by this strategy is at most O(ln |Ĥ|) times that of any other strategy.
- A query tree structure is used; there is not always a tree of average depth O(m).
- The best hope is to come close to minimizing the number of queries and this is done by a greedy approach:
- Algorithm:
  - Let S $\subseteq$ Ĥ be the current version space.
  - For each unlabeled x_i, let S_i^+ be the hypothesis which label x_i positive and S_i^- the ones which label it negative.
  - Pick the x_i for which the positive and negative are most nearly equal in weight; in other words $\min\{(S_i^+), (S_i^-)\}$ is largest.

Active Learning and Noise

- In active learning labels are queried to try to find the optimal separation. The most informative examples tend to be the most noise-prone.
  - QBC
  - Greedy Active Learning
- It can not be hoped to achieve speedups when η is large.
  - Kaariainen shows a lower bound of $\Omega(\eta^2/\varepsilon^2)$ on the sample complexity of any active learner

Comparison of Active Noisy Models

Agnostic Active Learning

- Arbitrary classification noise
- Data sampled i.i.d over some distribution D.
- Algorithm is shown to be successful for certain applications using any η, but exponential improvement if $\eta < \varepsilon/16$

Active Learning using Teaching Dimension

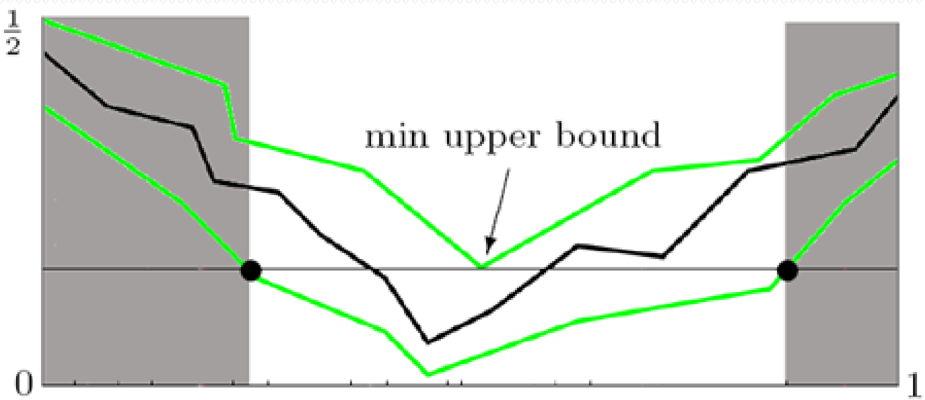
- Arbitrary **persistent** classification noise
- Data sampled i.i.d over some distribution $\mathbf{D}_{XY}$.
- Algorithm is successful for any application using noise rate $v \leq \eta$; not necessarily successful otherwise.

Agnostic Active Learning [B,B,L 2006]

A^2 *Algorithm*

Input: ϵ, Sample Oracle for D, Label Oracle O, H
Initialize: $i = 1$, $D_i = D$ $H_i = H$, $S_i = \emptyset$, and $k = 1$
while $\mathrm{DISAG}_D(H_i)(\min_{h \in H_i} UB(S_i, h, \delta_k) - \min_{h \in H_i} LB(S_i, h, \delta_k)) > \epsilon$ **do**
Set $S_i = \emptyset$, $H_i' = H_i$, $k = k + 1$
while $\mathrm{DISAG}_D(H_i') \geq \frac{1}{2}\mathrm{DISAG}_D(H_i)$ **do**
if $\mathrm{DISAG}_D(H_i')(\min_{h \in H_i'} UB(S_i, h, \delta_k) - \min_{h \in H_i'} LB(S_i, h, \delta_k)) \leq \epsilon$ **then**
Output: $h = \mathrm{argmin}(\min_{h \in H_i'} UB(S_i, h, \delta_k))$
else
$S_i' = 2|S_i| + 1$ sample from D satisfying $\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x)$
$S_i = S_i \cup \{(x, O(x)) : x \in S_i'\}$;
$H_i' = \{h \in H_i : LB(S_i, h, \delta_k) \leq \min_{h \in H_i'} UB(S_i, h', \delta_k)\}$; $k = k + 1$;
end if
end while
$H_{i+1} = H_i'$, $D_{i+1} = D_i$ conditioned on $\exists h_1, h_2 \in H_i : h_1(x) \neq h_2(x)$, $i = i + 1$
end while
Output: $h = \mathrm{argmin}(\min_{h \in H_i'} UB(S_i, h, \delta_k))$

Agnostic Active Learning



- The A^2 algorithm uses an UB and LB subroutine on a subset of examples to calculate the disagreement of a region.
- The disagreement of a region is $Pr_{x \in D}[\exists\, h_1, h_2 \in H_i : h_1(x) \neq h_2(x)]$.
- If all $h \in H_i$ agree on some region it can be safely eliminated thereby reducing the region of uncertainty.
- This eliminates all hypotheses whose lower bound is greater than the minimum upper bound.
- Each round completes when S_i is large enough to reduce half of its region of uncertainty which bounds the number of rounds by log(½)
- A^2 returns $h = \text{argmin}(\min_{h \in H'_i} UB(S, h, \delta))$.

**picture taken from "Agnostic Active Learning" [B,B,L, 2006]

Active Learning &TD [Hanneke 2007]

- Based upon the exact learning MembHalving algorithm [Hegedüs] which uses majority vote of h to continuously minimize V
- **Reduce** repeatedly gets the min specifying set of the subsequence for h_{maj} and V' is all h $\in$ V that did not produce the same outcome of the Oracle in all of the runs. Returns all V/V'
- **Label** gets the minimal specifying set as in reduce and labels those points. It labels the rest of the points which agree on h, h_{maj} and the Oracle using the majority value.

ReduceAndLabel (TDA)

Input: Finite $V \in C_F, U = \{x_1, x_2, \ldots, x_m\} \in X^m$,
values $\epsilon, \delta, \hat{\eta} \in (0, 1]$.

Initialize: $u = \lfloor |U|/(5 \ln |V|) \rfloor$, $V_0 = V, i = 0$

repeat

$i = i + 1$

Let $U_i = \{x_{1+u(i-1)}, x_{2+u(i-1)}, \ldots, x_{ui}\}$

$V_i = \text{Reduce}(V_{i-1}, U_i, \frac{\delta}{48 \ln |V|}, \hat{\eta} + \frac{\epsilon}{2})$

until $|V_i| > \frac{3}{4}|V_{i-1}$ or $|V_i| \leq 1$

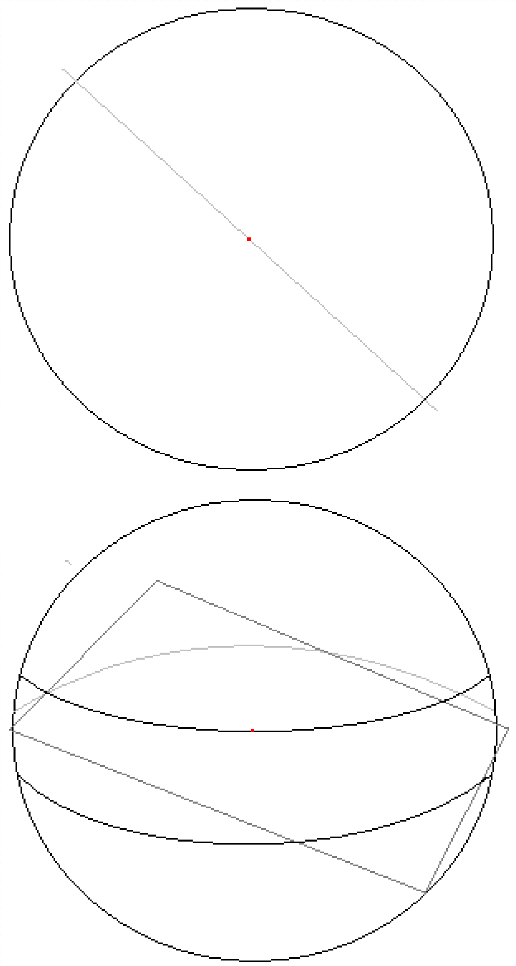
Let $\bar{U} = \{x_{ui+1)}, x_{ui+2)}, \ldots, x_{ui+l}\}$, where $l = \lceil 12\frac{\hat{\eta}}{\epsilon^2} \ln \frac{12|V|}{\delta} \rceil$

$L = \text{Label}(V_{i-1}, \bar{U}, \frac{\delta}{12}, \hat{\eta} + \frac{\epsilon}{2})$

Output: Concept $h \in V_i$ having smallest $er_L(h)$,
(or any $h \in V$ if $V_i = \emptyset$).

An application of Active Learning

- Active learning has been frequently examined using linear separators when the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$.
- **Definition:** X is the set of all data s.t. $X = \{x \in \mathbb{R}^d : ||x|| = 1\}$.
- The data-points lie on the surface area of the sphere.
- The distribution, D, on X is uniform.
- H is the class of linear separators through the origin.
- Any $h \in H$ is a homogeneous hyper-plane.



Comparing the Models

Model	# of Datapoints	# of Labels Queried
QBC	$\mathrm{O}(\frac{d}{\epsilon}\log\frac{1}{\delta\epsilon})$	$O(\frac{d}{\epsilon})$
Modified Perceptron	$O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$	$O(d\log\frac{1}{\epsilon})$
A^2	$\frac{64}{\epsilon^2}\left(2\mathrm{VC}\ln(\frac{12}{\epsilon})+\ln(\frac{4}{\delta})\right)$	$O\left(d\left(d\ln d+\ln\frac{1}{\delta'}\right)\ln\frac{1}{\epsilon}\right)$
TDA	?	?

Extended Teaching Dimension

Definition: $\forall f \in C_f, XTD(f,V,U) = inf(\{t | \exists R \subseteq U : |\{h \in V : h(R) = f(R)\}| \leq 1 \wedge |R| \leq t\}$

- The **teaching dimension** is the minimum number of instances a teacher must reveal to uniquely identify any target concept chosen from the class.
- The **extended teaching dimension** is a more restrictive form; The function of the minimal subset, f(R), can be satisfied by only one hypothesis, h(R), and the size of the subset is at most the size of XTD.

TDA Bounds

Theorem: *Let XTD be as defined above and* $X = \{0,1\}^d$ *and no datapoints lie on the separator. The bound on the number of labels queried in* $C, D, \epsilon, \delta, \eta$ *for linear separators under the uniform distribution in* X *is* $> (\frac{2^d}{\sqrt{d}})(\frac{\eta^2}{\epsilon^2} + 1)(d \log \frac{1}{\epsilon} + log \frac{1}{\delta})(\log \frac{d}{\epsilon\delta})$

- It is known that the TD for linear separators is 2^d [A,B,S 1995].
- The linear separator goes through the origin, therefore only the points lying near it need to be taught. This is roughly a TD of $2^d / \sqrt{d}$.
- The XTD is even more restrictive so it is probably worse.

Comparing the Models

Model	# of Datapoints	# of Labels Queried
QBC	$\mathrm{O}(\frac{d}{\epsilon}\log\frac{1}{\delta\epsilon})$	$O(\frac{d}{\epsilon})$
Modified Perceptron	$O(\frac{d}{\epsilon}\log\frac{1}{\epsilon})$	$O(d\log\frac{1}{\epsilon})$
A^2	$\frac{64}{\epsilon^2}\left(2\mathrm{VC}\ln(\frac{12}{\epsilon})+\ln(\frac{4}{\delta})\right)$	$O\left(d\left(d\ln d+\ln\frac{1}{\delta'}\right)\ln\frac{1}{\epsilon}\right)$
TDA	$\left[224\frac{\eta+\epsilon/2}{\epsilon^2}\ln\frac{48\ln\lvert 2(\frac{4e}{\epsilon}\ln\frac{4e}{\epsilon}^2)\rvert}{\delta}\right]\times$ $(5\ln\lvert 2(\frac{4e}{\epsilon}\ln\frac{4e}{\epsilon}^2)\rvert)$	$O>((\frac{2^d}{\sqrt{d}})(\frac{\eta^2}{\epsilon^2}+1)\times$ $(d\log\frac{1}{\epsilon}+log\frac{1}{\delta})(\log\frac{d}{\epsilon\delta}))$

Open Questions

- What are the bounds of A^2 for axis-aligned rectangles?
- Can the concept of Reduce and Label in TDA be used to write an algorithm that does not rely on the exact teaching dimension?
- Can a general algorithm be written which would produce reasonable results in all the applications.
- Can general bounds be created for A^2?