Sufficient Conditions for Agnostic Active Learnable

Liwei Wang

Peking University

Supervised learning problem

Training examples

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \qquad x_i \in X, \quad y_i \in Y = \{1, 2, \ldots, L\}$

i.i.d. from an underlying joint distribution D_{XY}

Hypothesis space: H

Generalization error: $err_D(h) := P_D(h(x) \neq y)$

Best hypothesis in $\boldsymbol{H}$: $\boldsymbol{h}^*$

$$\nu = err_D(h^*) = \arg\min_{h \in H} err_D(h).$$

Sample complexity:

Number of examples needed to learn a hypothesis ε-close to $\boldsymbol{h}^*$

i.e. $err_D(\hat{h}) \leq \nu + \varepsilon.$

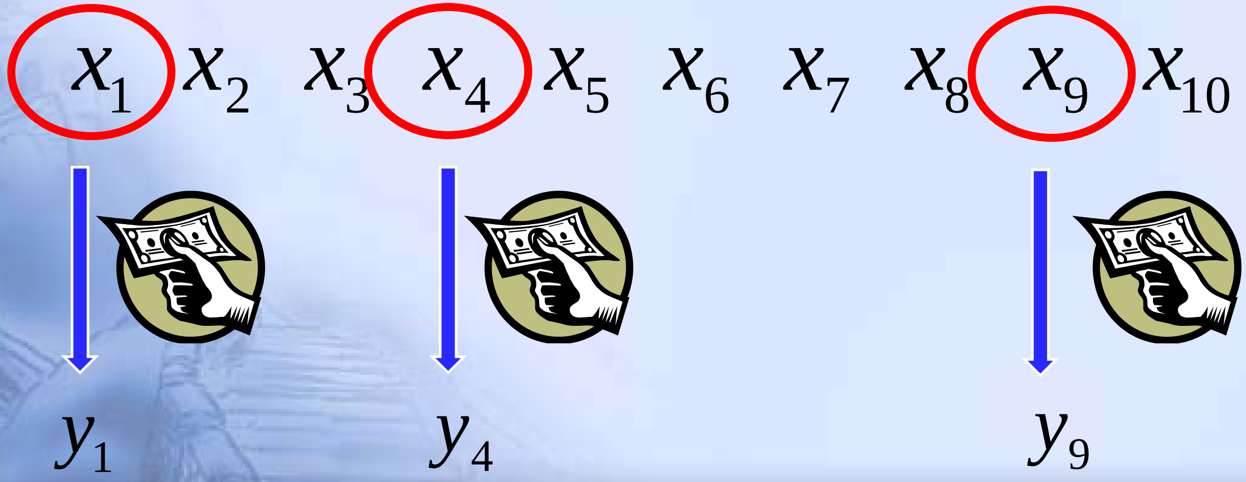
Sample complexity for supervised learning:

Agnostic: $O(1/\varepsilon^2)$ Realizable: $O(1/\varepsilon)$

Active Learning:

- Labeling costs human efforts; But **unlabeled** examples are often cheap.
- Can we reduce the number label requests by choosing the most informative data to label?

Pool-based active learning:



x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_9 x_{10}

y_1 y_4 y_9

Fundamental question on active learning:

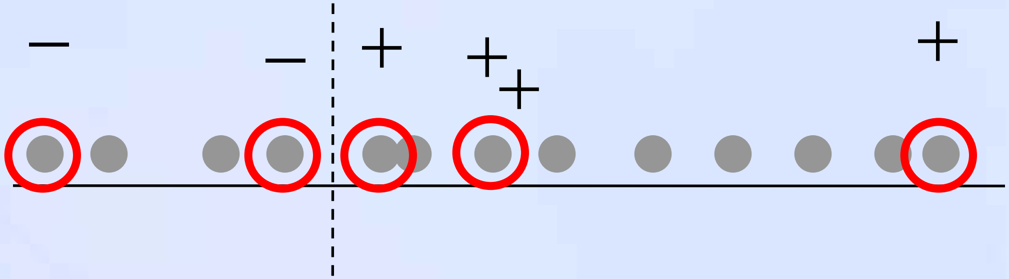
Does active learning requires strictly fewer labels than supervised learning?

Positive example:

- Leaning threshold on a line interval:

  There is a threshold perfectly separates the two classes.

  Data has a uniform distribution on [0,1].



Sample complexity: $O(\log \frac{1}{\varepsilon})$

- Leaning homogenous linear separators in R^d

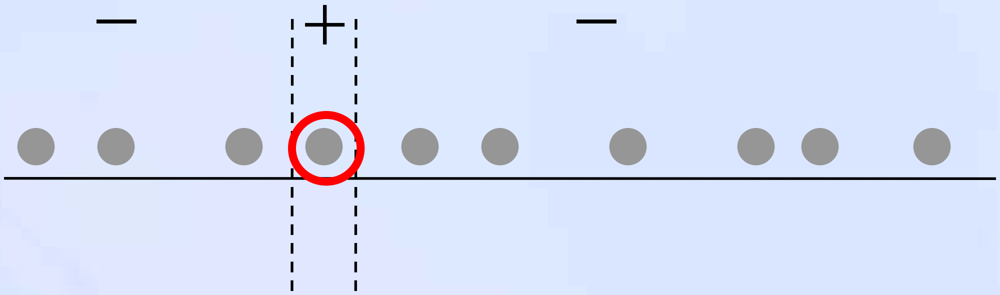
  sample complexity: $O(\log^d \frac{1}{\varepsilon})$

  [Dasgupta 2005]

Negative example:

- Learning intervals on [0,1].

Positive samples lie in an interval on [0,1], negative samples lie elsewhere.



Label complexity: $O(1/\varepsilon)$. No advantage!

Lower bound on the sample complexity: $O(\nu^2/\varepsilon^2)$

Active learning does not always help! [Kaarininen 2006]

A natural question:

- Under what condition does active learning help? Is there any intuitively reasonable conditions under which active learning does help?

We study *sufficient* conditions under which active learning is strictly superior than supervised learning.

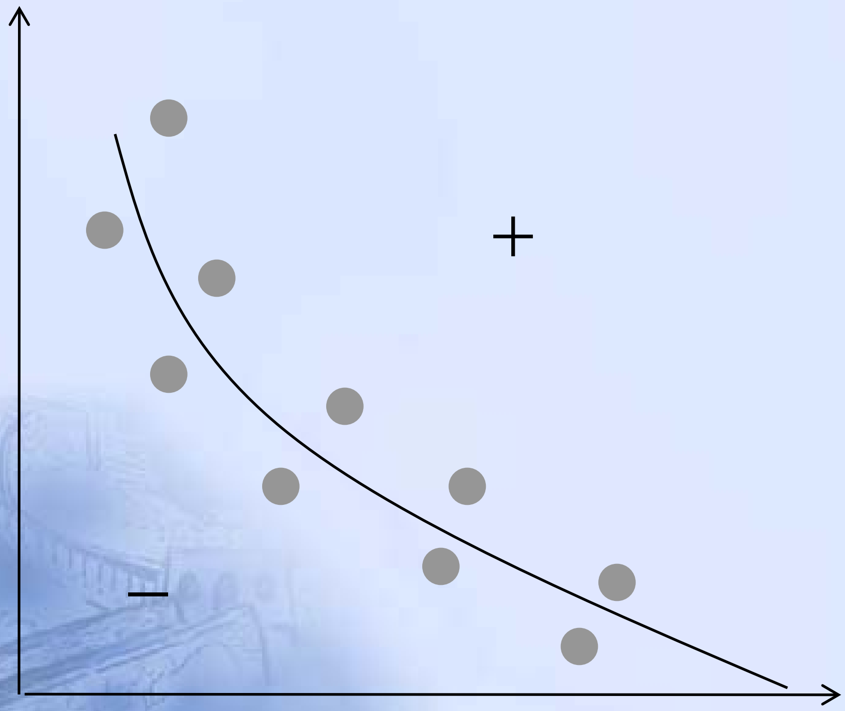
Main result (informal):

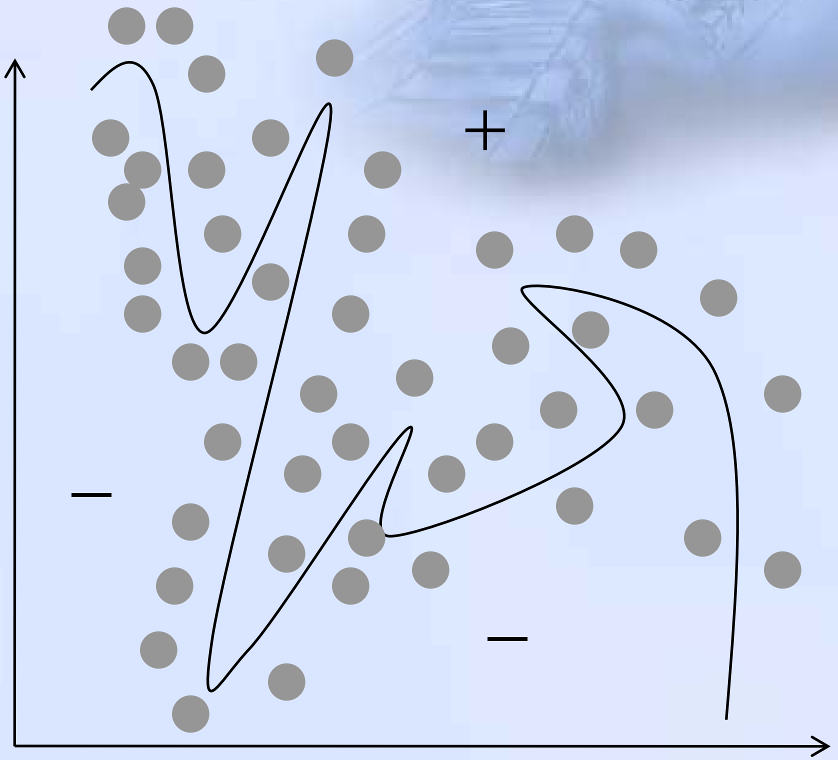
Smoothness is important for active learning.

If the classification boundary is smooth, then under some noise condition active learning requires strictly less labels than supervised learning.

The improvement is polynomial for finite smoothness.

Exponential for infinite smoothness.





Smooth functions:

- K-norm: $\|f\|_K := \max_{|\mathbf{k}|\le K-1} \sup_x \left|D^{\mathbf{k}} f(x)\right| + \max_{|\mathbf{k}|=K-1} \sup_x \frac{\left|D^{\mathbf{k}} f(x) - D^{\mathbf{k}} f(x')\right|}{\|x-x'\|}$,

  where $\mathbf{k} = (k_1, k_1, \ldots, k_d)$, $D^{\mathbf{k}} = \frac{\partial^{|\mathbf{k}|}}{\partial^{k_1} x_1 \ldots \partial^{k_d} x_d}$.

- Kth order smooth functions:

$$K < \infty: \quad F_C^K := \{f : \|f\|_K \le C\}.$$

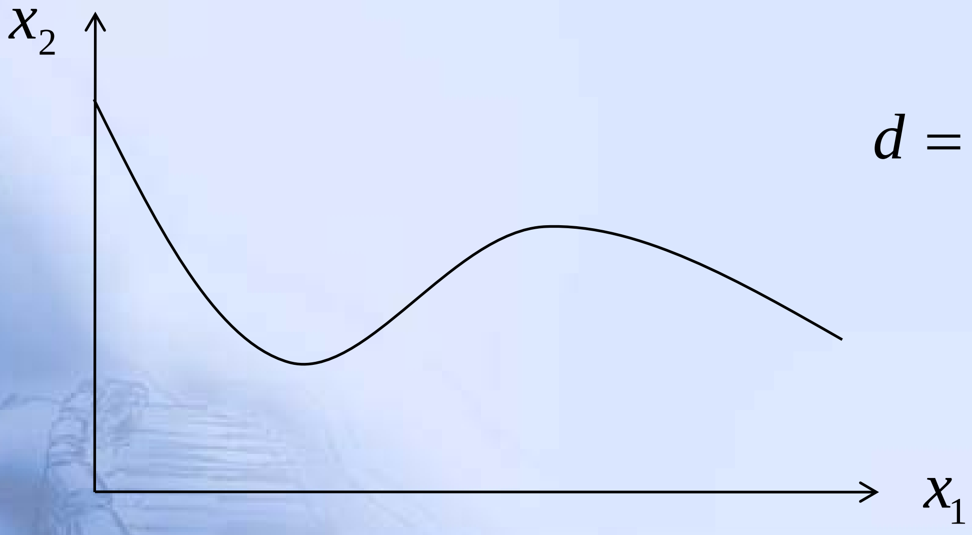
Kth order smooth functions have uniformly bounded partial derivatives up to order K-1, and the Kth order partial derivatives are Lipschitz.

$$K = \infty: \quad F_C^\infty := \{f : \ f \in F_C^K, \ K = 1,2,\ldots\}.$$

The smooth boundary hypothesis space:

- Hypotheses with smooth boundaries:

A set of hypotheses H_C^K defined on $[0,1]^{d+1}$ is said to have Kth order smooth boundaries, if for every $h \in H_C^K$ the corresponding classification boundary is the graph of a Kth order smooth function on $[0,1]^d$.



$d = 1$

- Main results
  - Thm: Let the hypothesis space be H_C^K. Assume that the Bayes classifier h^* of the learning problem is in H_C^K; D_X has a density bounded from both above and below by a Kth order smooth function; and the Tsybakov noise condition (exponential form) holds. Then there is active learning algorithm that outputs a classifier that is ε -close to h^*.

If $K < \infty$, the label complexity is $\tilde{O}\left(\left(\frac{1}{\varepsilon}\right)^{\frac{2d}{K+d}}\right)$;

If $K = \infty$, the label complexity is $O\left(polylog\left(\frac{1}{\varepsilon}\right)\right)$.

Supervised learning: $O(1/\varepsilon^2)$

The agnostic active learning algorithm A^2

[Balcan, Beygelzimer, Langford]

- Maintain hypotheses that with high probability are not too worse than the best classifier (according to the labeled examples).
- Ask for labels of those examples that current hypotheses have disagreement on its label.
- Never much worse than supervised learning.
- The sample complexity is characterized by the *Disagreement Coefficient* θ [Hanneke]

$$O\left(\theta^2\left(\frac{\nu^2}{\varepsilon^2}+1\right)polylog\left(\frac{1}{\varepsilon}\right)\right).$$

- Disagreement Coefficient (DC):
  - Depends on both the learning problem and the hypothesis space.
  - Definition of DC:

Let $\rho(\cdot,\cdot)$ be the pseudo-metric on a hypothesis space H induced by D_X . That is, for $h, h' \in H$, $\rho(h, h') = P_{X \sim D_X}(h(X) \neq h'(X))$. Let $B(h^*, r) = \{h' \in H : \rho(h^*, h') \leq r\}$. DC $\theta(\varepsilon)$ is defined as:

$$\theta(\varepsilon) = \sup_{r \geq \varepsilon} \frac{P_{X \sim D_X}(X \in DIS(B(h^*, r)))}{r},$$

where $DIS(B(h^*, r)) = \bigcup_{h \in B(h^*, r)} \{x \in X : h(x) \neq h^*(x)\}$.

DC for smooth problems:

- Thm:

  Let the hypothesis space be H_C^K. If the distribution D_X has a density $p(\cdot)$ such that there exists a Kth order smooth function $g(\cdot)$ and two constants $0 \le \alpha \le \beta$ such that $\alpha g \le p \le \beta g$, then

for $K < \infty$ $$\theta(\varepsilon) = O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{d}{K+d}}\right),$$

for $K = \infty$, $$\theta(\varepsilon) = O\left(\log^d\left(\frac{1}{\varepsilon}\right)\right).$$

The Tsybakov noise condition [Tsybakov]

$$P_D\left(\left|\eta(X)-1/2\right|\le\frac{1}{T}\right)\le c_1 e^{-c_2 T},\qquad \eta(X)=P\left(Y=1\mid X\right).$$

The label complexity of A^2 under Tsybakov noise:

$$O\left(\theta^2 polylog\left(\frac{1}{\varepsilon}\right)\right)$$

Without noise assumption:

$$O\left(\theta^2\left(\frac{\nu^2}{\varepsilon^2}+1\right)polylog\left(\frac{1}{\varepsilon}\right)\right).$$

Future Direction

- Computationally efficient active learning algorithms that yield the same sample complexity bound.

Thanks