An Analysis of Active Learning with Uniform Feature Noise

Aaditya Ramdas, Aarti Singh, Barnabas Poczos, Larry Wasserman

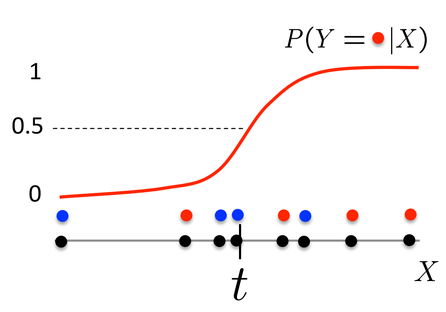
MACHINE LEARNING DEPARTMENT

Carnegie Mellon University




Introduction

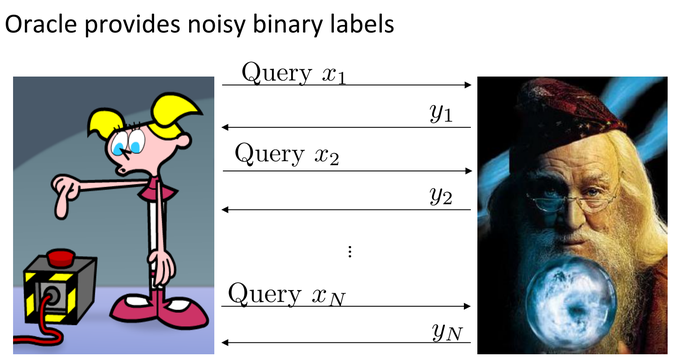
Active 1-D Threshold Learning



Aim: minimize **label complexity**

(N = # queries needed to find decision boundary)

Oracle provides noisy binary labels



$y_i \in \{0,1\}$ and $\mathbb{E}[Y|X=x] = P(Y=1|X=x)$

Point error: $|\hat{t}_N - t|$

Errors-In-Variables

1. Classical Model (observe W,Y – infer m)

$$W = X + \delta$$
$$Y = m(X) + \epsilon \qquad W \leftarrow X \rightarrow Y$$

2. Berkson Model (observe W,Y – infer m)

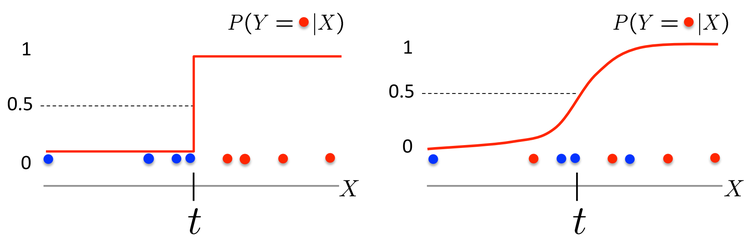
Seemingly easier to interpret active learning

$$X = W + \delta$$
$$Y = m(X) + \epsilon \qquad W \rightarrow X \rightarrow Y$$

(motivation: errors in measurement, communication, oracle)

Active Learning

Tsybakov's Noise Condition

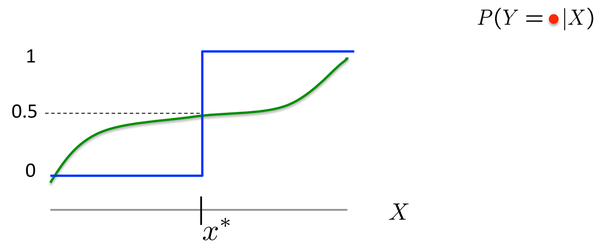


$$|P(Y=\bullet|X=x) - 1/2| \geq c|x-t|^{k-1}$$

Characterizes growth of the regression function around the threshold t by a polynomial of distance to t.

When does Active Learning help?

Flatter is harder



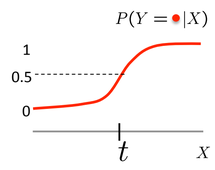
- JUMP : Binary search - exponentially fast! $k = 1$
- FLAT : No intelligent queries – passive learning $k \to \infty$

$$|P(Y=\bullet|X=x) - 1/2| \geq c|x-t|^{k-1}$$

Castro-Nowak's Minimax Rates

If Tsybakov's Noise Condition (TNC) holds

$$|P(Y=\bullet|X=x) - 1/2| \geq c|x-t|^{k-1}$$



then, minimax optimal active learning rate is

$$\mathbb{E}|\widehat{t}_N - t| \asymp N^{-\frac{1}{2k-2}}$$

Eg: exponentially fast when $k = 1, N^{-1/2}$ when $k = 2$, constant when $k = \infty$

and minimax optimal passive learning rate is

$$\mathbb{E}|\widehat{t}_N - t| \asymp N^{-\frac{1}{2k-1}}$$

Eg: 1/N when $k = 1, N^{-1/3}$ when $k = 2$, constant when $k = \infty$

Active Learning + Feature Noise

Formal Setup

Let the domain be $[-1, 1]$, regression function m.
Unique t s.t. $m(t) = 1/2$ (Bayes' optimal classifier).

1. User chooses W, requests label.
2. Oracle receives noisy W, namely $X = W + U$
3. Oracle returns Y, where $P(Y = +|X = x) = m(x)$

We take noise $U \sim \text{Unif}[-\sigma, \sigma]$, known σ.
Loss measure is point error $L(\hat{t}, t) = |\hat{t} - t|$.

Assume querying within σ of boundary disallowed.

Passive learning : $W \sim \text{Unif}[-1, 1]$ or a grid.

Minimax Risk

For $k \geq 1$, define $\mathcal{P}(k, \sigma)$ as the set of functions $m(x)$ satisfying for some threshold t,

T. $C|x-t|^{k-1} \geq |m(x) - 1/2| \geq c|x-t|^{k-1}$ whenever $|m(x) - 1/2| \leq \epsilon_0$

M. $m(t+\delta) = 1/2 - m(t-\delta)$ for all $\delta \leq \sigma$

B. t is at least σ away from $\{-1, 1\}$.

Minimax Risk under point error loss:

$$R_N(P(k,\sigma)) = \inf_{S \in \mathcal{S}_N} \sup_{P \in \mathcal{P}(k,\sigma)} \mathbb{E}|\hat{t}_N - t|$$

where $\mathcal{S}_N$ is the set of active/passive strategies with access to N oracle queries.

Summary of Main Result

Under the Berkson error model, given N labels sampled (A)ctively or (P)assively when the true regression function lies in $\mathcal{P}(k, \sigma)$ for known k, σ, the minimax risk is

1. $\mathcal{R}_N^P(\mathcal{P}(k,\sigma)) \asymp \begin{cases} N^{-\frac{1}{2k-1}} & \text{if } \sigma \prec N^{-\frac{1}{2k-1}} \\ \sigma^{-(k-\frac{3}{2})}\sqrt{\frac{1}{N}} & \text{otherwise} \end{cases}$

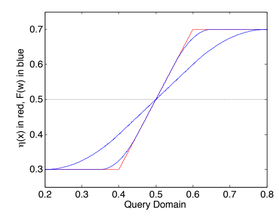
2. $\mathcal{R}_N^A(\mathcal{P}(k,\sigma)) \asymp \begin{cases} N^{-\frac{1}{2k-2}} & \text{if } \sigma \prec N^{-\frac{1}{2k-2}} \\ \sigma^{-(k-2)}\sqrt{\frac{1}{N}} & \text{otherwise} \end{cases}$

Discussion

Remarks about Main Result

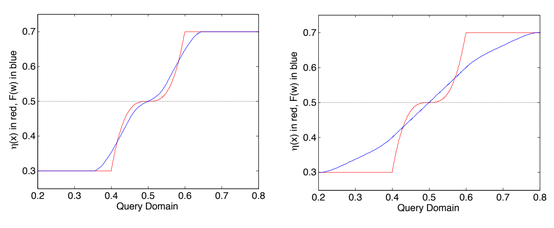
1. When σ is zero, we get Castro-Nowak's noiseless rates.
2. When $\sigma <$ noiseless error, we get the same noiseless rate.
3. When σ is large, and $k > 2$ (flat regression function), we get an improvement in rate with larger noise level σ!
4. On the one hand, this is explained by the unflattening of the regression function when convolved with uniform noise.
5. On the other hand, the function class gets simpler with larger σ because of $m(x)$'s local anti-symmetry around t.

Unflattening due to Convolution



Original in red, convolved in blue, k = 2

Small noise (left) and large noise (right)



Original in red, convolved in blue, k = 4

Remarks about Proof Techniques

1. We use Castro-Nowak's proof technique involving Fano's inequality to derive information theoretic lower bounds.
2. We demonstrate two well-separated functions in $\mathcal{P}(k, \sigma)$ that are hard to differentiate, i.e. the joint distribution of observations under both functions have small KL divergence.
3. Passive upper bound – a histogram variant (estimate at each bin averages over a region of width σ instead of bin-width).
4. Active upper bound – any optimal passive subroutine can be applied recursively to get an optimal active algorithm.