Literatre Review of Active Learning Convergence Rates

Simon D. Nguyen [1] **Kentaro Hoffman** [1] **Tyler H. McCormick** [1 2]

1. Selective Sampling Using the Query by Committee Algorithm (Freund, Seung, and Tishby 1997)

Main Result The following holds with probability larger than $1-\delta$ over the random choice of the target concept, the sequence of examples, and the choices made by QBC.

- The number of calls to Sample that QBC makes is smaller than

$$m_0 = \max\left\{\frac{4d}{e\delta}, \frac{160(d+1)}{g\epsilon}\max\left(6, \ln\frac{80(d+1)}{\epsilon\delta^2 g}\right)^2\right\} \tag{1}$$

- The number of calls to label that QBC makes is smaller than

$$n_0 = \frac{10(d+1)}{g}\ln\frac{4m_0}{\delta} \tag{2}$$

such that

- d: VC dimension
- g : expected information gain of queries made by QBC is uniformly lower bounded by $g > 0$. A uniform lower bound on the information means that for any version space (the set of hypotheses consistent with our training data) that can be reached by QBC with non-zero probability, the expected information gain from the next query of QBC is larger than g.

Notes

1. Results are in number of observations to be labeled
2. Done in a two member committee
3. Achieves exponential improvement in label efficiency.
4. Restrictive Assumption: concepts are assumed to be deterministic and noiseless.
5. Defined for concept class (linear separators):

$$c_{\vec{w}}(\vec{x}, t) = \begin{cases} 1 & \text{if } \vec{w}\cdot\vec{x} \geq t, \\ 0 & \text{if } \vec{w}\cdot\vec{x} < t, \end{cases} \tag{3}$$

2. Rates of Convergence in Active Learning (Henneke 2011)

Main Result (one of many) When allowed n label requests, there exists a finite universal constant c such that, with probability $\geq 1-\delta$, $\forall n \in \mathbb{N}$

$$Error(\hat{h}_n) - \nu \leq c\sqrt{\frac{\nu^2\theta^2\left(d\log n + \log\delta^{-1}\right)\cdot\log\left(\frac{n+2\nu\theta}{\nu\theta}\right)}{n}} + 2\exp\left\{-\frac{n}{c\theta^2\left[d\log\theta + \log(n\delta^{-1})\right]}\right\}$$

such that

- θ: Disagreement coefficient. It is a measure of complexity that quantifies disagreement among a set of classifiers.
- d: VC dimension of hypothesis class
- $\nu := \inf_{h\in\mathbb{C}} er(h)$: noise rate of the hypothesis class (best achieveable error)
- n: n label requests
- $\delta \in (0, 0.5)$: Confidence parameter

Notes

- Results are in achieveable generalization error
- Primarily goes the route of disagreement coefficient
- Results are form Algorithm 1, which uses confidence bounds to eliminate suboptimal classifiers and focuses sampling on a region R.
- Further results in the paper extend the above results to show even better rates under Tsybakov's noise conditions. All are dependent upon VC dimension d.
- Tsybakov's noise conditions basically quantify how much noise exists as you move further/closer to the decision boundary parameterized by κ Tsybakov's Noise Conditions: Characterized by parameter $\kappa \geq 1$. Value $\kappa = 1$ means noiseless/bounded noise (probability "jumps" at the boundary) with $\kappa > 1$ means the noise is unbounded (probability approaches 0.5 near the boundary).

3. Minimax Bounds for Active Learning (Castro and Nowak, 2007)

Theorem 3 (Upper Bound on Active Learning): Consider an active learning strategy using a piecewise polynomial interpolation.

Let $\rho = (d-1)/\alpha$ then

$$\limsup_{n\to\infty} \sup_{P\in BF(\alpha,\kappa,L,C,c)} \mathbb{E}[R(\hat{G}_n)] - R(G^*) \le c_{\max}\left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+\rho-2}}. \tag{4}$$

Notes

- The error rate is for a *theoretical active learning algorithm* that:
  - Constructs a grid with spacing M over the first $d-1$ dimensions
  - Samples N points actively along vertical line segments through this grid
  - Uses one-dimensional change-point detection on each line
  - Builds a piecewise polynomial approximation of the decision boundary
- $BF(\alpha,\kappa,L,C,c)$ represents distributions with:
  - Decision boundaries that are graphs of Hölder smooth functions with parameter α
  - Noise characterized by parameter $\kappa \ge 1$
  - Assumes Tsybakov's Noise Conditions

4. Information, prediction, and query by committee (Seung et al. 1992)

Main Result The probability that one of the two committee members makes a mistake on a randomly chosen example with respect to a randomly chosen

$$\left(3 + O(e^{-c_1 n})\right) \cdot \frac{n}{d}\exp\left\{-\frac{c}{2(d+1)}n\right\} \tag{5}$$

such that

- d: VC dimension
- c: lower bound on the expected information gain
- n number of queries asked so far
- c_1 some constant.

Notes

- *Prediction error decreases exponentially* with n queries when information gain is lower bounded.