Active Learning for Decision Trees with Provable Guarantees

Arshia Soltani Moakhar, Tanapoom Laoaron, Faraz Ghahremani,
Kiarash Banihashem, MohammadTaghi Hajiaghayi

ICLR
International Conference On
Learning Representations

Motivation

Active learning

minimizes labeling effort.

Critical when labels are expensive (medical diagnosis)

Decision tree

Decision trees are interpretable, efficient, and widely used.

The Gap

Existing learning theory doesn't cover decision trees.

Contributions

- First analysis of the disagreement coefficient for decision trees:

$$\theta = O(ln^d(n))$$

- First active learning algorithm achieving (1+ε)-multiplicative error guarantee for classification

Disagreement Coefficient

It measures how many data points have uncertain labels within a set of plausible hypotheses: $\theta_h := \sup_{r>0} \frac{|\mathrm{DIS}_S(B_H(h,r))|}{rn}$

Why does it matter?

- Bounding θ for decision trees → bounding their active learning cost
- Smaller θ → more efficient active learning.
- Prior work showed θ is finite for decision trees, but never explicitly computed it.

Bounding θ for Decision Trees

Theorem 1.1 The disagreement coefficient is: $\theta = O(ln^d(n))$ Under two assumptions

- each node tests a feature distinct from its ancestors
- the data has a grid-like structure

Approach

- Decompose a tree h into LineTrees: simpler classifiers that isolate one leaf at a time.
- Analyze disagreement of each LineTree, then recombine.

Assumptions Are Necessary

- Without distinct features per node:

$$\theta = \Omega(n^{1/dim})$$

- Without grid structure:

$$\theta = \Omega(n)$$

A (1+ε)-Multiplicative Error Algorithm

- Additive error algorithms can't be efficiently adapted

Theorem 1.2 General Binary Classification Algorithm returns a (1+ε)-approximate classifier w.p. ≥ 1-δ using:

$$O(ln(n)\cdot\theta^2\cdot(V_H\cdot ln\theta + ln(lnn/\delta)) + (\theta^2/\varepsilon^2)\cdot(V_H\cdot ln(\theta/\varepsilon) + ln(1/\delta)))$$

- For decision trees this means a poly-logarithmic label complexity

Lower Bounds: Near-Optimality

Theorem 4.3 Any active learning algorithm requires at least $\Omega(ln(1/\delta) \cdot 1/\varepsilon^2)$ queries to return a (1+ε)-approximate decision stump with probability > 1-δ.

Dependence on ε cannot be improved beyond log factors