Multi-group Learning for Hierarchical Groups

Samuel Deng [1] Daniel Hsu [1]

[1]Department of Computer Science, Columbia University

Problem Setup

Cast of characters. Consider an input space $\mathcal{X}$ and a label space $\mathcal{Y} = \{0, 1\}$...

- Arbitrary joint distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$.
- "Benchmark" hypothesis class $\mathcal{H}$ of functions $h : \mathcal{X} \to \{0, 1\}$.
- $\mathcal{G} \subseteq \mathcal{X}$ is a collection subsets of the input space (*groups*).
- Can assume that $\mathcal{H}$ is an arbitrary class with finite VC dimension $d_{\mathcal{H}}$; $\mathcal{G}$ is finite and exponentially large or has VC dimension $d_{\mathcal{G}}$.
- Our notion of *test/generalization error*:

$$L_{\mathcal{D}}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}\{f(x) \neq y\}] = \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y].$$

Our notion of *empirical/sample error* over dataset $S = \{(x_i, y_i)\}_{i=1}^n$:

$$L_S(f) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{f(x_i) \neq y_i\}$$

NOTE: In this poster, we focus on zero-one loss, but this generalizes to arbitrary bounded loss functions.

Motivation. Traditional learning theory is concerned with *aggregate* performance over $\mathcal{D}$. No assurance for *individual-level* guarantees:

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y] < \epsilon \quad \text{where } (x, y) \sim \mathcal{D}.$$

A middle-ground between on-average and individual-level guarantees: consider a rich collection of subsets of the input space, $\mathcal{G} \subseteq \mathcal{X}$, and ensure:

$$L_{\mathcal{D}}(f \mid g) := \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x) \neq y \mid x \in g] < \epsilon_g \quad \text{for all } g \in \mathcal{G}.$$

In *agnostic (PAC) learning*, for any $\epsilon \in (0, 1)$, given $n = \text{poly}\left(\frac{1}{\epsilon}, d_{\mathcal{H}}\right)$ i.i.d. *training examples* $(x_i, y_i) \sim \mathcal{D}$, goal is to find $\hat{f} : \mathcal{X} \to \{0, 1\}$ such that, with high probability over the i.i.d. training examples,

$$L_{\mathcal{D}}(\hat{f}) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Learning theory 101: For finite VC classes, empirical risk minimization (ERM) is necessary and sufficient, with optimal sample complexity $d_{\mathcal{H}}/\epsilon^2$.

Multi-group (agnostic PAC) learning

For any $\epsilon \in (0, 1), \gamma \in (0, 1)$, given $n = \text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\gamma}, d_{\mathcal{H}}, d_{\mathcal{G}}\right)$ i.i.d. examples $(x_i, y_i) \sim \mathcal{D}$, goal is to find $\hat{f} : \mathcal{X} \to \{0, 1\}$ such that, with high probability over the i.i.d. training examples,

$$L_{\mathcal{D}}(\hat{f} \mid g) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h \mid g) + \epsilon_g \quad \text{for all } g \in \mathcal{G}.$$

Why is this interesting? We can no longer resort to ERM over all the data!

There may be no single $h \in \mathcal{H}$ that is good for all groups simultaneously!

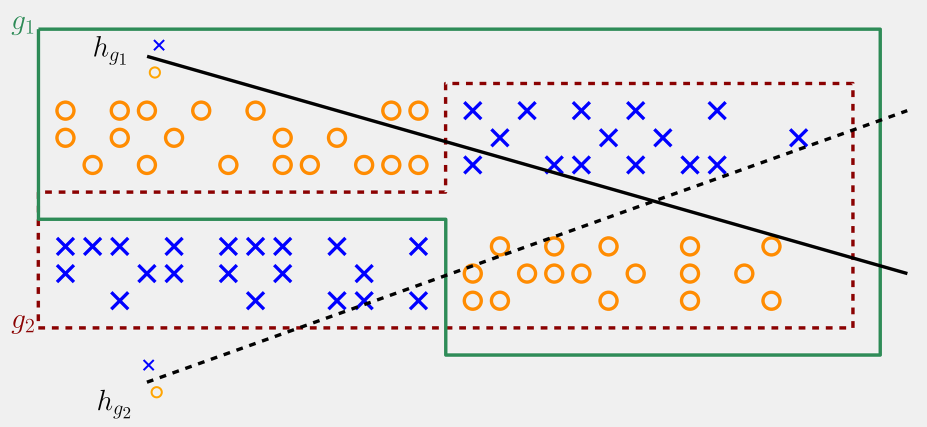

Figure 1. **No best h for all groups simultaneously.** Letting $\mathcal{H}$ be the class of halfspaces, the groups g_1 (indicated by the green solid line) and g_2 (indicated by the red dotted line) overlap, but their optimal predictors h_{g_1} and h_{g_2} are much different.

Hierarchically structured groups

Our main assumption. A collection of groups $\mathcal{G}$ is *hierarchically structured* or *laminar* if, for every pair of distinct groups $g, g' \in \mathcal{G}$, exactly one of the following holds:

1. $g \cap g' = \emptyset$ (g and g' are disjoint).
2. $g \subset g'$ (g is contained in g').
3. $g' \subset g$ (g' is contained in g).

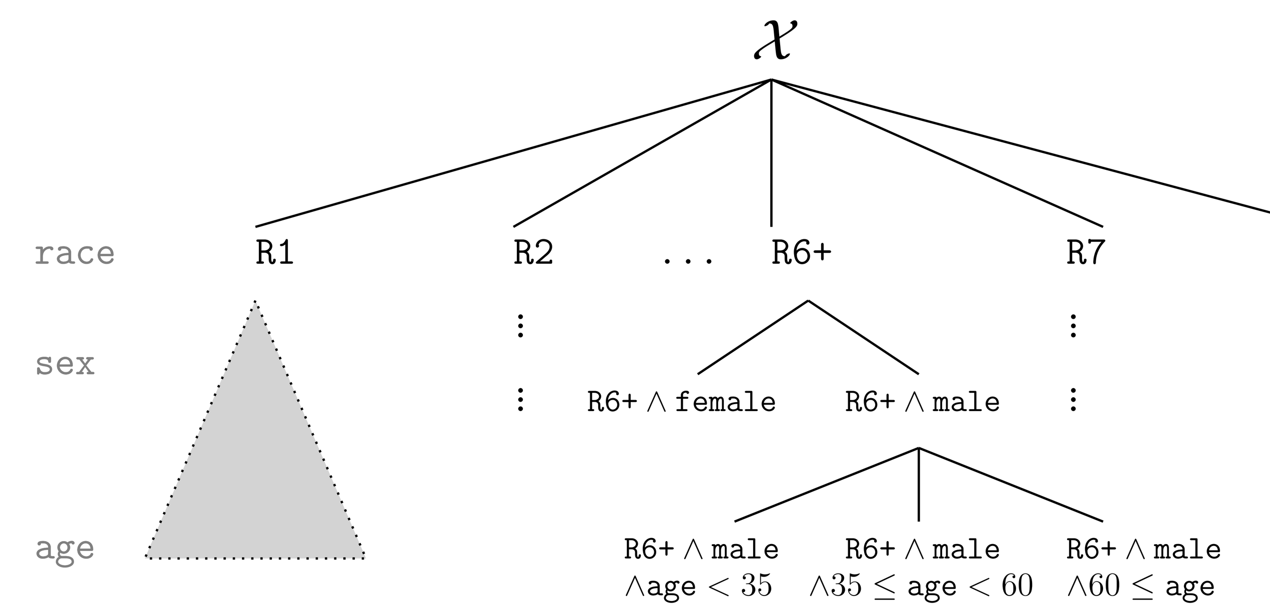

Figure 2. **Example of a hierarchically structured tree.** Each level of the tree above corresponds to a demographic attribute (`race`, `sex`, and `age`). Proceeding down the tree yields increasingly granular subgroups. The leaves are the most granular level, with subgroups such as `R6+` ∧ `male` ∧ `age` < 35.

Goal and existing results

Goal: We want an *interpretable (simple)*, *computationally efficient*, and *statistically efficient* classifier. Previous work traded off between these goals:

- *Rothblum & Yona (2021) [2]*: Boosting-style algorithm with sample complexity of
$$n = \frac{1}{\epsilon^8\gamma}\text{polylog}\left(\frac{|\mathcal{H}| \times |\mathcal{G}|}{\epsilon}\right).$$
- *Tosh & Hsu (2022) [3] and Globus-Harris, Kearns, Roth (2022) [1]*: (Non-optimal) decision list algorithm with sample complexity of
$$n = \frac{d_{\mathcal{H}} + d_{\mathcal{G}}}{\epsilon^3\gamma^2}\log\left(\frac{1}{\epsilon}\right).$$
- *Tosh & Hsu (2022) [3]*: Complex, uninterpretable online-to-batch algorithm ensembling n base classifiers, with sample complexity of
$$n = \frac{1}{\epsilon^2\gamma}\left(d_{\mathcal{H}}\log\frac{1}{\epsilon} + \log|\mathcal{G}|\right)$$

Theorem: Near-optimal sample complexity with hierarchical groups

Suppose $\mathcal{H}$ is a benchmark hypothesis class $\mathcal{H}$ with VC dimension $d_{\mathcal{H}}$ and $\mathcal{G} \subseteq \mathcal{X}$ is a collection of hierarchically structured groups. Let $\epsilon_g \in (0, 1)$ be a desired level of accuracy for each group $g \in \mathcal{G}$. There exists a learning algorithm requiring

$$n_g := \frac{1}{\epsilon_g^2}\left(d_{\mathcal{H}}\log\frac{1}{\epsilon} + \log|\mathcal{G}|\right)$$

samples for each $g \in \mathcal{G}$ that achieves **multi-group agnostic PAC learning.**

Algorithm

Algorithm 1 `MGL-Tree`

Require:
1: S, a training dataset.
2: Collection of hierarchically structured groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$.
3: Error rates $\epsilon_n(g) \in (0, 1)$ for all $g \in \mathcal{G}$

Ensure: Decision tree $f : \mathcal{X} \to \{0, 1\}$.
4: Order $\mathcal{G}$ into a *hierarchical tree* $\mathcal{T}_{\mathcal{G}}$.
5: Initialize the root: $f^{\mathcal{X}} := \hat{h}^{\mathcal{X}}$.
6: **for** each node $g \in \mathcal{T}_{\mathcal{G}} \setminus \{\mathcal{X}\}$ in breadth-first order **do**
7: Compute the ERM classifier $h \in \mathcal{H}$ for g:
$$\hat{h}^g \in \arg\min_{h\in\mathcal{H}} L_S(h \mid g).$$
8: **if** $L_S(f^g(x) \mid g) - L_S(\hat{h}^g \mid g) - \epsilon_n(g) \geq 0$ **then**
9: Set $f^g := \hat{h}^g$.
10: **else**
11: Set $f^g := f^{\text{pa}(g)}$, where $\text{pa}(g)$ denotes the parent node of g.
12: **end if**
13: **end for**
14: **return** $f : \mathcal{X} \to \{0, 1\}$, a decision tree predictor.

The appropriate setting of $\epsilon_n(g)$ for each group comes from the Theorem (possibly conservative):

$$\epsilon_n(g) = 18\sqrt{\frac{2d\log(16|\mathcal{G}|n/\delta)}{n_g}}.$$

Some experimental results

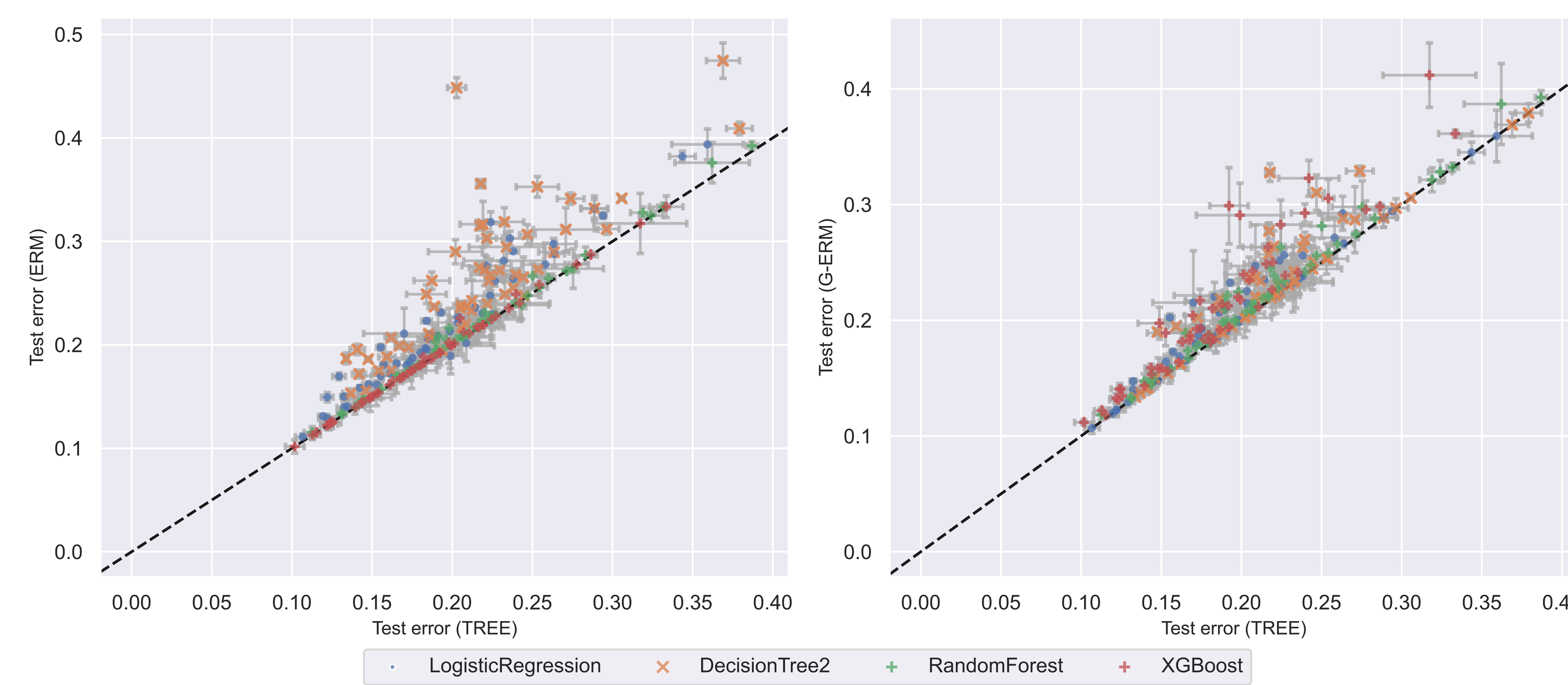

Figure 3. **Test accuracy on `race-sex-age` groups for CA Employment (top row) and CA Income (bottom row).** Each point in the plot represents the test error on a specific group. The $y = x$ line represents equal error between our algorithm and the competing method; points *above* the $y = x$ line are groups where our algorithm exhibits better generalization.