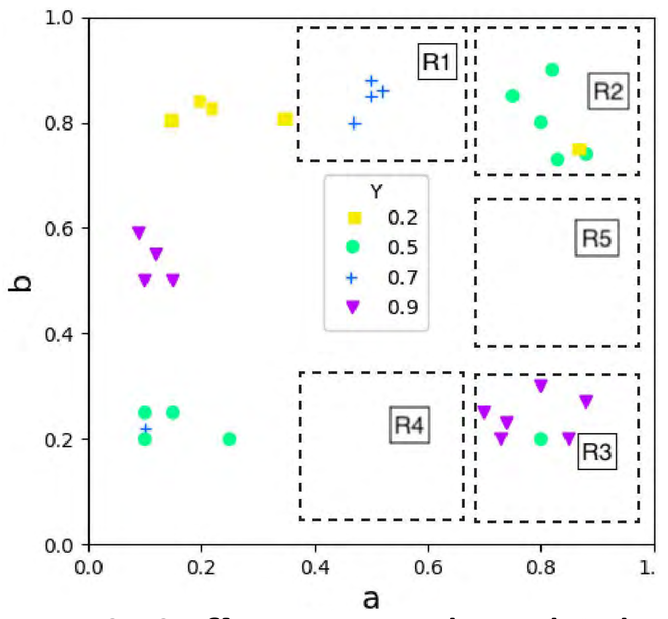
A Unified Framework for **K**nowledge **I**ntensive **G**radient **B**oosting: Leveraging Human Experts for Noisy Sparse Domains

Harsha Kokel, Phillip Odom, Shuo Yang, Sriraam Natarajan

UT Dallas · LinkedIn · Georgia Tech · STARLinG LAB

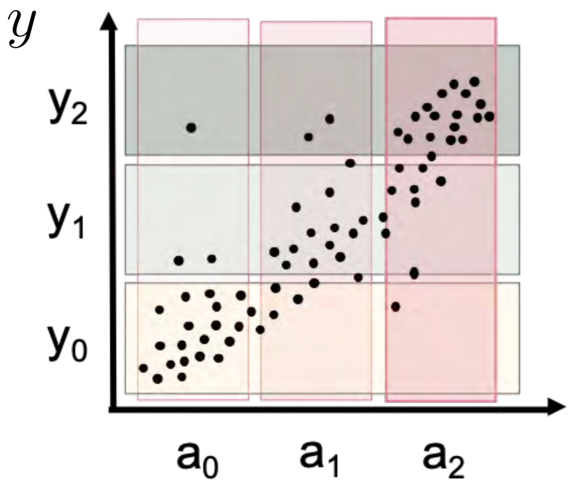
Motivation

- Can we leverage domain knowledge – like monotonic influence for regions where we do not have data (R5) or data is noisy (R1-R2) while using boosting methods?



- **Monotonic influence**: when higher values of variable a stochastically results in higher values of variable y

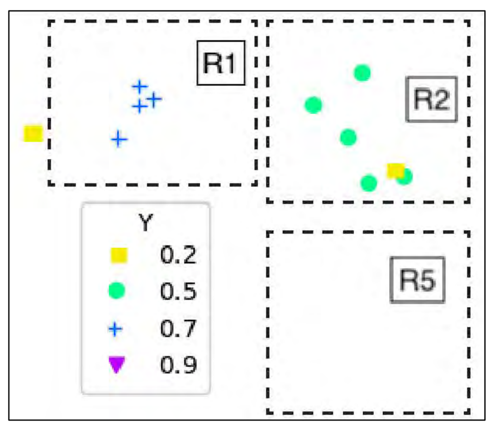
$$a \prec^{Q+} y$$



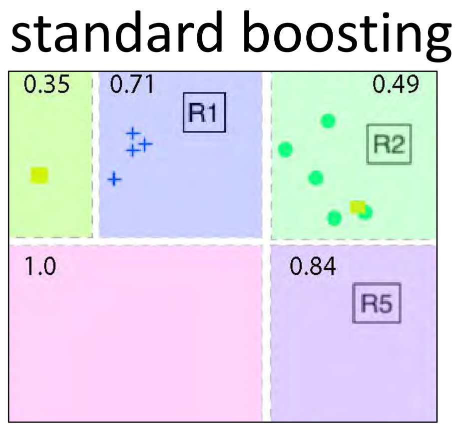
- **Why monotonic influence**?
  Most of the market trends can be expressed as monotonic or synergic influence

Key aspect - Avoids overfitting

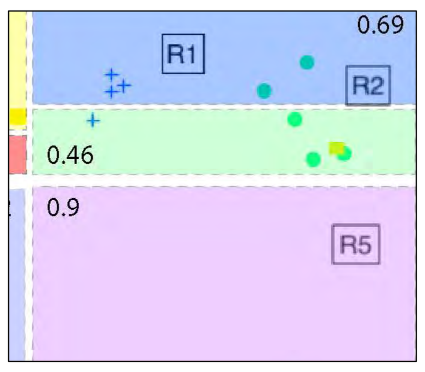
- When the data violates the advice (region R1 & R2 below), LGBM monotonic boosting approach overfits the data, KiGB on the other hand allows for trade-off between data and advice



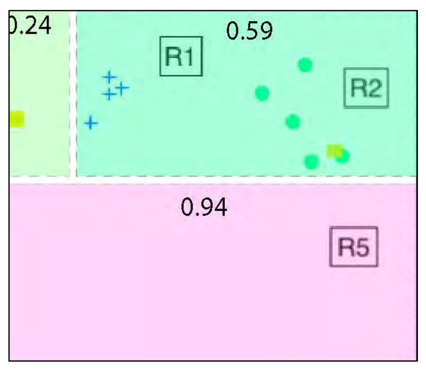
standard boosting



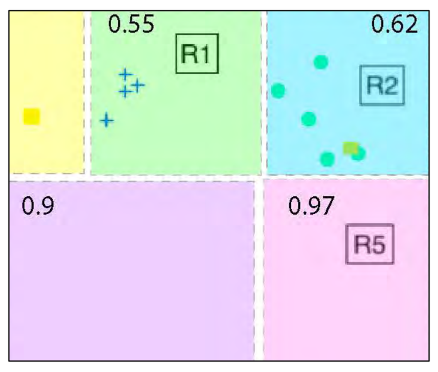
monotonic boosting



KiGB with $\lambda = 6$

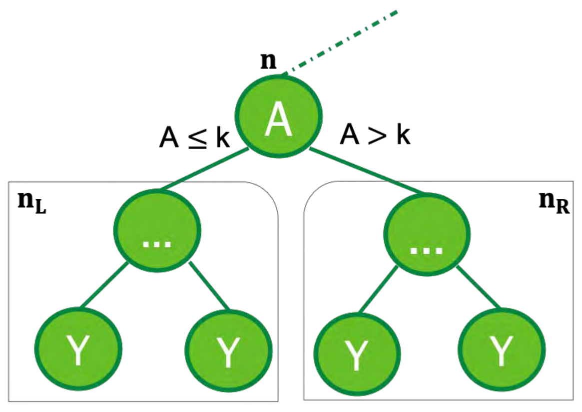


KiGB with $\lambda = 8$



Formulation

- When monotonic influence is indicated by an expert, the **expected value of left subtree should be no more than the right subtree**



$$\mathbb{E}_\psi[\mathbf{n_L}] \le \mathbb{E}_\psi[\mathbf{n_R}]$$

$$\underbrace{\mathbb{E}_\psi[\mathbf{n_L}] - \mathbb{E}_\psi[\mathbf{n_R}]}_{\zeta_n} - \varepsilon < 0$$

violation margin/slack

- Objective

relative importance of advice

$$\underset{\psi_t}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{N} (\tilde{y}_i - \psi_t(x_i))^2}_{\text{loss function w.r.t data}} + \frac{\lambda}{2} \underbrace{\sum_{n \in \mathcal{N}(x_c)} \max(\zeta_n \cdot |\zeta_n|, 0)}_{\text{loss function w.r.t. advice}}$$

- Update equation

$$\psi_t^{\ell}(\mathbf{x}) = \underbrace{\frac{1}{|\ell|} \sum_{i=1}^{N} \tilde{y}_i \cdot \mathbb{I}(x_i \in \ell)}_{\text{mean}} +$$

$$\underbrace{\frac{\lambda}{2} \sum_{n \in \mathcal{N}(x_c)} \mathbb{I}(\zeta_n > 0)\zeta_n \cdot \left(\frac{\mathbb{I}(\ell \in \mathbf{n_R})}{|\mathbf{n_R}|} - \frac{\mathbb{I}(\ell \in \mathbf{n_L})}{|\mathbf{n_L}|} \right)}_{\text{penalty for advice violation}}$$

Algorithm

Knowledge-intensive Gradient Boosting (KiGB)

INPUT: Data $(\mathbf{x}, y)$, # trees M, monotonic features $\mathbf{x}_c$, λ, ε
OUTPUT: $\psi(\mathbf{x})$

function KiGB($\mathbf{x}, y, M, \mathbf{x}_c, \lambda, \varepsilon$)
  $\psi(\mathbf{x}) = \psi_0(\mathbf{x}) = \text{mean}(y)$
  for $m = 1$ **to** M **do**
    $\tilde{y} = y - \psi(\mathbf{x})$ ▹ Compute gradient
    $\psi_m(\mathbf{x}) = \text{tree}(\tilde{y}, \mathbf{x})$ ▹ Learn the next tree
    for ℓ **in** ψ_m **do**
      $\psi_m^{\ell}(\mathbf{x}) = \psi_m^{\ell}(\mathbf{x}) + \text{penalty}^{\ell}(\mathbf{x}_c, \lambda, \varepsilon)$
      ▹ Update leaf values
    end for
    $\psi(\mathbf{x}) = \psi(\mathbf{x}) + \psi_m(\mathbf{x})$ ▹ Update the function
  end for
  return $\psi(\mathbf{x})$
end function

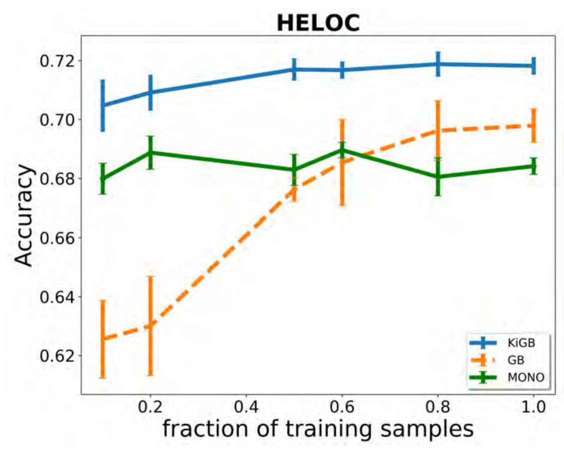
Code and paper is available on the website

Results

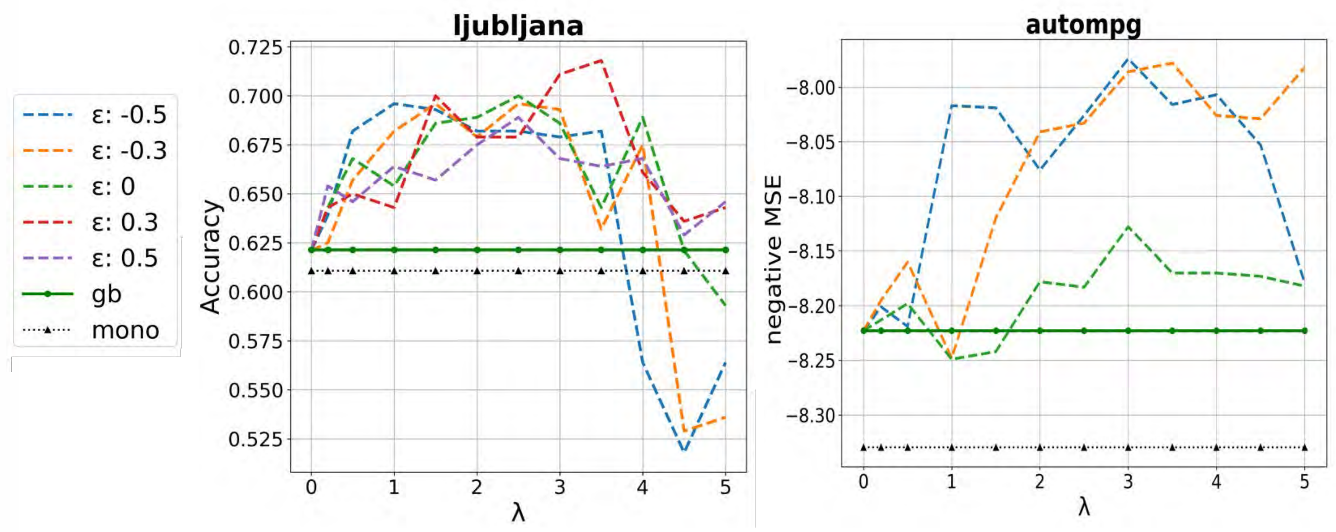
- **Baselines**
  - LGBM – LightGBM gradient boosting
  - LMC – LightGBM monotonic boosting
  - SGB – Scikit-learn gradient boosting
  - MONO – Monoensemble (Bartley et al. 2019)
- Comparison with 15 standard (10 regression and 5 classification) datasets is in the paper
- **KiGB** performs better than standard boosting and existing monotonic boosting approaches in most of the domains
- Comparison on **real world dataset** – Logistics dataset from **Turvo** Inc. and HELOC dataset from **FICO** (xML challenge Neurips 2018)

Dataset	KiGB	LGBM	LMC
Logistics(mse)	**1.851**	1.898	1.889
Dataset	KiGB	SGB	MONO
HELOC (accuracy)	**0.717**	0.7	0.688

- **KiGB** achieves a **jump start** and **higher asymptote** in performance



- ε and λ allows for **trade-off between data and advice**, thus making KiGB robust to noisy data/advice
- **KiGB** is robust to choice of hyperparameters: λ and ε



Conclusion

- **First** unified framework to handle qualitative constraints inside FGB
  - Regression and classification
- Easily extendable to relational settings
- Allows for a natural human-machine interaction
- Next step – Actively soliciting such advice during boosting

Human knowledge can help gradient boosting achieve **efficient and effective** performance in **noisy sparse domains**.

TURVO

This research was supported by Turvo Inc. and CwC Program Contract W911NF-15-1-0461 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO)

https://starling.utdallas.edu
https://www.utdallas.edu/~hkokel