Learning Accurate and Interpretable Decision Trees

Maria-Florina Balcan, Dravyansh Sharma

Carnegie Mellon University

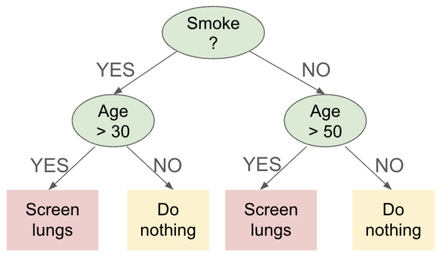
Best Paper from Sister Conferences Track

IJCAI 2025 MONTREAL

Decision Trees

Interpretability of model predictions is a major challenge in modern machine learning!

- Decision trees are highly interpretable.
- But learning optimal decision trees is hard!



Several heuristic approaches are known for training decision trees.

- How to select the best heuristic?
- How much data is needed to learn provably good decision trees?

New Learning Perspective

- *Data-driven algorithm design* [1,2] family of algorithms given by real-valued hyperparameters.
- **Goal**: learn hyperparameters from a collection of multiple datasets coming from the same domain $(X_1, y_1), \ldots, (X_N, y_N)$.
- Domain $\iff$ fixed, unknown distribution $\mathcal{D}$.

Formally, given a bound on the maximum tree size t, a finite family $\mathcal{F}$ of node functions,

How many samples N are enough to learn a near-optimal hyperparameter $\hat{\lambda}$? (λ^* is optimal)

$$\mathbb{E}_{(X,y)\sim\mathcal{D}}[\ell_{(X,y)}(\hat{\lambda}) - \ell_{(X,y)}(\lambda^*)] \leq \epsilon$$

[1] M.-F. Balcan. *Data-Driven Algorithm Design* (book chapter). In *Beyond Worst Case Analysis of Algorithms*, Tim Roughgarden (Ed). Cambridge University Press, 2020.

[2] D. Sharma. *Data-driven algorithm design and principled hyperparameter tuning in machine learning*. PhD Thesis, 2024.

Splitting criterion

Top-down learning

Inputs: Node function class $\mathcal{F}$, tree size t, splitting criterion G
1. Start with leaf node
2. While (at most t leaf nodes)
   Split leaf node l using node function f which maximizes "splitting criterion" G

Key decision: *Which node to split next and how?*

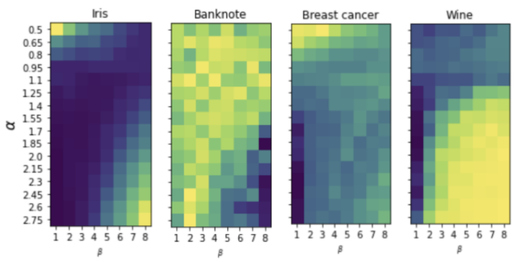
New parameterized family, (α, β)-Tsallis entropy

$$g^{\text{TSALLIS}}_{\alpha,\beta}(P) := \frac{C}{\alpha - 1}\left(1 - \left(\sum_{i=1}^{c} p_i^{\alpha}\right)^{\beta}\right)$$

- $g^{\text{TSALLIS}}_{2,1} \iff$ Gini impurity
- $g^{\text{TSALLIS}}_{1,1} \iff$ Entropy
- $g^{\text{TSALLIS}}_{\frac{1}{2},2} \iff$ Kearns-Mansour criterion

Theorem. We can learn to tune (α, β) using $O\left(\frac{t \log |\mathcal{F}| t}{\epsilon^2}\right)$ problem samples.

Proof insight. Analyse accuracy as a function of (α, β) on a fixed instance (X, y);
Induction over top-down rounds, bounding the number of distinct behaviors in each round;
Over t rounds, $\tilde{O}(|\mathcal{F}|^{2t} t^{2t})$ distinct behaviors, which implies pseudo-dimension is $O(t \log |\mathcal{F}| t)$.



Different (α, β) work best for different datasets.

Bayesian Trees

Two-phase randomized algorithm.

Prior

1. Start with a single root node
2. Split the node with probability
$$p_{\text{SPLIT}} = \sigma(1 + d)^{-\phi}$$
3. Select uniformly random splitting rule at each node if split
4. Repeat step 2 for each new node

Stochastic search

1. T^0 = initial skeleton with random rules according to Prior
2. $T^* \leftarrow$ obtained by small modification to T^i
3. $T^{i+1} = T^*$ with probability $q(T^i, T^*)$ based on Dirichlet posterior, $T^{i+1} = T^i$ otherwise

Question: *How to set hyperparameters σ, ϕ?*

Insight Analyze the accuracy as a function of hyperparameters for fixed random bits;
piecewise constant with exponential boundaries and at most $t^2 N^2$ pieces over N problem samples.

Theorem. $O(\log t/\epsilon^2)$ dataset samples are sufficient to learn near-optimal parameters σ, ϕ.

Pruning

Post-processing to

- Reduce overfitting
- Increase interpretability

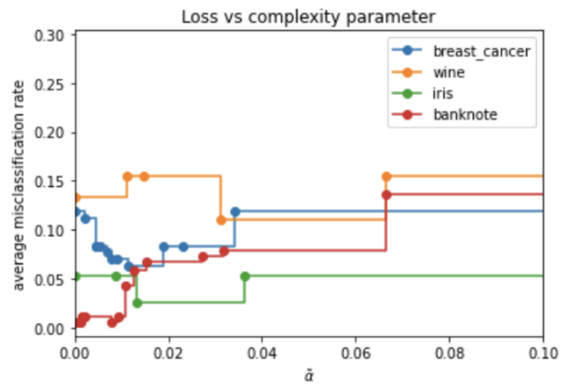
Min cost-complexity pruning

- Maximizing accuracy on training set typically leads to large trees
- Add tree size as a penalty term in training loss

Cost-complexity:

$$R_{(X,y)}(T) = \ell_{(X,y)}(T) + \alpha|\text{leaves}(T)|$$

Theorem. $O(\log t/\epsilon^2)$ dataset samples are sufficient to learn near-optimal α.

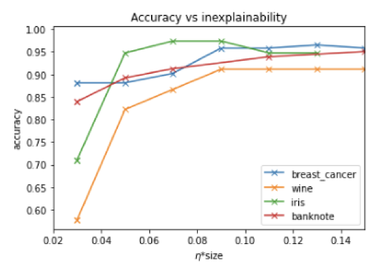


Interpretability

Modified objective:

$$R_{(X,y)}(T) = \ell_{(X,y)}(T) + \eta|\text{leaves}(T)|$$

- Unlike min cost-complexity pruning, also modify test loss
- η controls the accuracy-interpretability trade-off
- Tune splitting/pruning hyperparameters simultaneously to maximize the modified objective



Conclusion

- We design a new parameterized family of splitting criteria encompassing known ones.
- Sample complexity bounds for learning splitting criterion, Bayesian prior and pruning when building a decision tree.
- Optimize both accuracy and interpretability.

Read and Discuss!

Balcan and Sharma, *Learning accurate and interpretable decision trees*. UAI 2024.