Sparse Learning with CART

Jason M. Klusowski

Princeton University, Operations Research and Financial Engineering

Objectives

- When do decision trees adapt to the sparsity of a predictive model?

Introduction

- Training data
$$\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}, \quad (\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$$
- Predictor for decision tree T
$$\widehat{Y} = \widehat{Y}(T, \mathcal{D})$$
- Prediction error
$$\mathrm{Err}(\widehat{Y}(T)) = \mathbb{E}_{(\mathbf{X}',Y')}[(Y' - \widehat{Y}(\mathbf{X}'))^2]$$
for independent copy $(\mathbf{X}', Y')$

CART decision trees

- CART [1] methodology based on recursively minimizing impurity
- For regression, impurity in node $\mathrm{t} \in T$ is
$$\widehat{\Delta}(\mathrm{t}) = \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \overline{Y}_{\mathrm{t}})^2,$$
where $N(\mathrm{t}) = \#\{\mathbf{X}_i \in \mathrm{t}\}$ and $\overline{Y}_{\mathrm{t}} = \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} Y_i$
- Optimal direction $\hat{\jmath}$ and split point $\hat{s}$ obtained by maximizing reduction in impurity
$$\widehat{\Delta}(s, \mathrm{t}) = \widehat{\Delta}(\mathrm{t}) - \frac{N(\mathrm{t}_L)}{N(\mathrm{t})} \widehat{\Delta}(\mathrm{t}_L) - \frac{N(\mathrm{t}_R)}{N(\mathrm{t})} \widehat{\Delta}(\mathrm{t}_R),$$
where
$$\mathrm{t}_L = \{\mathbf{X} \in \mathrm{t} : X_j \le s\}, \quad \mathrm{t}_R = \{\mathbf{X} \in \mathrm{t} : X_j > s\}$$
are left and right child nodes
- Tree output $\widehat{Y}(\mathbf{x}) = \overline{Y}_{\mathrm{t}}$ for $\mathbf{x}$ in terminal node t

Main results

- Consider pruned tree
$$\widehat{T} \in \arg\min_{T \preceq T_{\max}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}(\mathbf{X}_i))^2 + \alpha |T| \right\},$$
where $T_{\max}$ is fully grown tree, temperature $\alpha = \Theta((d/n)\log(n/d))$, and $|T|$ is # of terminal nodes

Theorem

Suppose $\mathbf{X}$ *is uniformly distributed on* $[0,1]^d$ *and*
$$Y = \sum_j g_j(X_j)$$
is a sparse additive model with $d_0 \ll d$ *smooth component functions* $g_j(\cdot)$*, where each function is not too locally 'flat'. Then,*
$$\limsup_n \frac{Err(\widehat{Y}(\widehat{T}))}{((d/n)\log(n/d))^{\Omega(1/d_0)}} \overset{a.s.}{=} \mathcal{O}(1).$$

Proof idea

- Reduction in impurity $\widehat{\Delta}(\hat{s}, \mathrm{t})$ can be written as
$$\widehat{\Delta}(\mathrm{t}) \times \widehat{\rho}^{\,2}(\widehat{Y}, Y | \mathbf{X} \in \mathrm{t}),$$
where $\widehat{\rho} = \widehat{\rho}(\widehat{Y}, Y | \mathbf{X} \in \mathrm{t})$ is Pearson correlation between response data Y and optimal decision stump
$$\widehat{Y} = \overline{Y}_L \mathbf{1}(X_{\hat{\jmath}} \le \hat{s}) + \overline{Y}_R \mathbf{1}(X_{\hat{\jmath}} > \hat{s})$$
- Training error bound
$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}(\mathbf{X}_i))^2 \le \widehat{\mathrm{Var}}(Y) \exp(-K \times \min_{\mathrm{t}} \widehat{\rho}^{\,2}),$$
where $K = \Theta(\log_2(n))$ is tree depth and $\liminf_n \min_{\mathrm{t}} \widehat{\rho}^{\,2} = \Omega(1/d_0)$ a.s.

Experiments

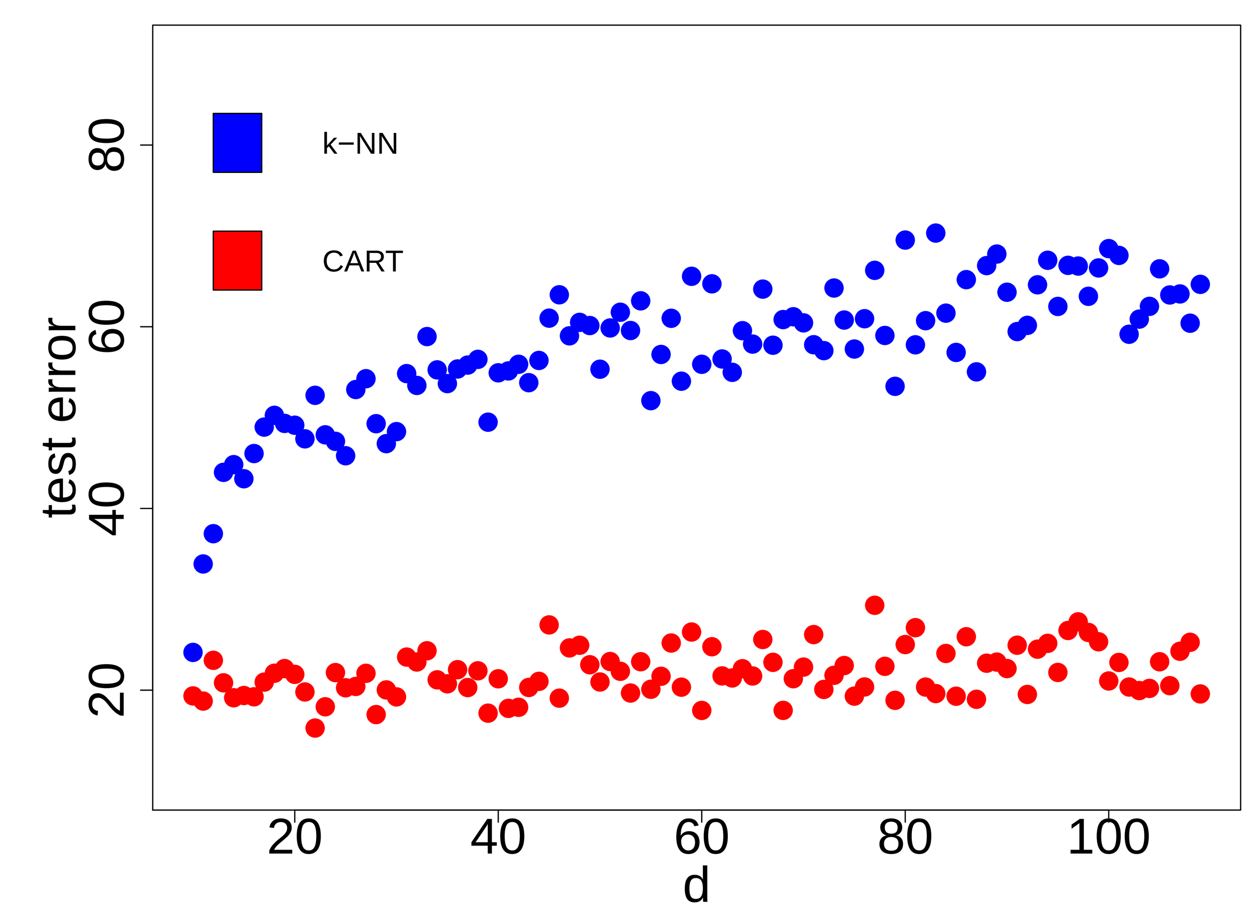

Figure: Boston housing dataset [1] ($d_0 = 10$ and $n = 506$) with $d - d_0$ noisy features added. Plot shows prediction error of pruned CART vs. cross-validated k-NN as d varies.

Conclusion

- CART adapts to underlying sparsity, whereas kernel methods with nonadaptive weights (like k-NN) suffer from curse of dimensionality

References

[1] Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone.
Classification and regression trees.
Chapman and Hall/CRC, 1984.

PRINCETON UNIVERSITY

NSF