Learning to Branch

Balcan, Dick, Sandholm, Vitercik

Introduction

- Parameter tuning tedious and time-consuming
- Algorithm configuration using Machine Learning

- Focus on tree search algorithms
  - Branch-and-Bound

Tree Search

- Widely used for solving combinatorial and nonconvex problems
- Systematically partition search space
- Prune infeasible and non-optimal branches
- Partition by adding constraint on some variable

Paritioning strategy is important!

- Tremendous effect on the size of the tree

Example: MIPs

Maximize $c^T x$ subject to $Ax \leq b$

- Some entries of x constrained to be in $\{0, 1\}$.

- Models many NP-hard problems.
- Applications such as Clustering, Linear separators, etc.

(Winner determination)

$$
\begin{array}{lll}
\text{maximize} & \sum_{i=1}^{n} \sum_{b \in B_i} v_i(b) x_{i,b} & \\
\text{s.t.} & \sum_{i=1}^{n} \sum_{b \in B_i, j \ni b} x_{i,b} \leq 1 & \forall j \in [m] \\
& \sum_{b \in B_i} x_{i,b} \leq 1 & \forall i \in [n] \\
& x_{i,b} \in \{0, 1\} & \forall i \in [n], b \in B_i.
\end{array}
$$

Model

- Application domain as distribution over instances
- Unknown underlying distribution but have sample access

Use samples to learn a variable selection policy.

- As small a search tree as possible in expectation over the distribution

Variable selection

Learning algorithm returns empirically optimal parameter (ERM)

- Adaptive nature is necessary
- Small change in parameters can cause drastic change (unconventional, e.g. SCIP)
- Data-driven approach is beneficial

Contribution

Theoretical:

- Use ML to determine optimal weighting of partitioning procedures.
- Possibly exponential reduction in tree size.
- Sample complexity guarantees that ensure empirical performance over samples matches expected performance on the unknown distribution.

Experimental:

- Different partitioning parameters can result in trees of vastly different sizes.
- Data-dependent vs worst-case generalization guarantees.

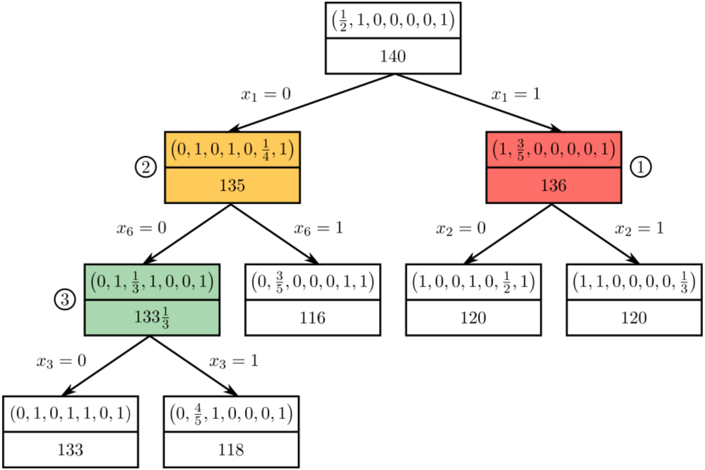
MILP Tree Search

- Usually solved using branch-and-bound.
- Subroutines that compute upper and lower bound of a region.
- Node selection policy.
- Variable selection policy (branch on a fractional var).

Fathom every leaf. A leaf is fathomed if:

- Optimal solution to LP relaxation is feasible.
- Relaxation is infeasible.
- Obj. value of relaxation is worse than current OPT.

MILP B & B example



Variable selection

- Score-based variable selection
- Deterministic function
- Takes partial tree, a leaf and a variable as input and returns a real value

Some common MILP score functions:

- Most fractional
- Linear scoring rule
- Product scoring rule
- Entropic lookahead

Learning to branch

Goal: Learn convex combination of scoring rules that is nearly optimal in expectation.

$$\mu_1 \mathit{score}_1 + \ldots + \mu_d \mathit{score}_d$$

(ϵ, δ)-learnability

Data-independent approaches

Theorem 3.1. *Let*

$$\texttt{score}_1(\mathcal{T}, Q, i) = \min\left\{\breve{c}_Q - \breve{c}_{Q_i^+}, \breve{c}_Q - \breve{c}_{Q_i^-}\right\}, \ \texttt{score}_2(\mathcal{T}, Q, i) = \max\left\{\breve{c}_Q - \breve{c}_{Q_i^+}, \breve{c}_Q - \breve{c}_{Q_i^-}\right\},$$

and $\texttt{cost}(Q, \mu\texttt{score}_1 + (1-\mu)\texttt{score}_2)$ *be the size of the tree produced by B&B. For every* a, b *such that* $\frac{1}{3} < a < b < \frac{1}{2}$ *and for all even* $n \geq 6$*, there exists an infinite family of distributions* $\mathcal{D}$ *over MILP instances with* n *variables such that if* $\mu \in [0,1] \setminus (a,b)$*, then*

$$\mathbb{E}_{Q\sim\mathcal{D}}\left[\texttt{cost}\left(Q, \mu\texttt{score}_1 + (1-\mu)\texttt{score}_2\right)\right] = \Omega\left(2^{(n-9)/4}\right)$$

and if $\mu \in (a,b)$*, then with probability 1,* $\texttt{cost}\left(Q, \mu\texttt{score}_1 + (1-\mu)\texttt{score}_2\right) = O(1)$*. This holds no matter which node selection policy B&B uses.*

- Infinite family of distributions such that the expected tree size is exponential in n.
- Infinite number of parameters such that the tree size is just a constant (with probability 1).

Sample complexity guarantees

Assumes path-wise scoring rules.

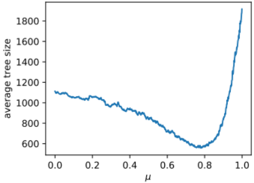
Lemma 3.3. *Let* $\texttt{cost}$ *be a tree-constant cost function, let* $\texttt{score}_1$ *and* $\texttt{score}_2$ *be two path-wise scoring rules, and let* Q *be an arbitrary problem instance over* n *binary variables. There are* $T \leq 2^{n(n-1)/2} n^n$ *intervals* $I_1, \ldots, I_T$ *partitioning* $[0,1]$ *where for any interval* I_j*, across all* $\mu \in I_j$*, the scoring rule* $\mu\texttt{score}_1 + (1-\mu)\texttt{score}_2$ *results in the same search tree.*

- Bound on the intrinsic complexity of the algorithm class defined by range of paremeters.

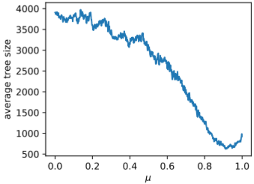
Theorem 3.7. *Let* $\texttt{cost}$ *be a tree-constant cost function, let* $\texttt{score}_1$ *and* $\texttt{score}_2$ *be two path-wise scoring rules, and let* $\mathcal{C}$ *be the set of functions* $\{\texttt{cost}(\cdot, \mu\texttt{score}_1 + (1-\mu)\texttt{score}_2) : \mu \in [0,1]\}$*. Then* $\mathrm{Pdim}(\mathcal{C}) = O\left(n^2\right)$.

- Implies generalization guarantee.

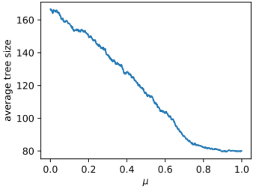
Experiments



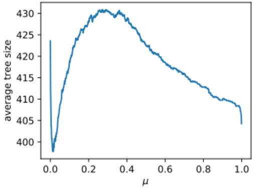
(a) CATS "arbitrary"



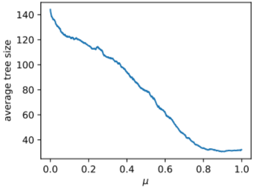
(b) CATS "regions"



(c) Facility location



(d) Linear separators

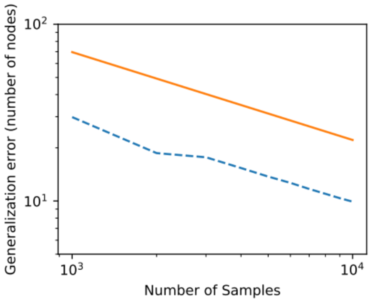


(e) Clustering

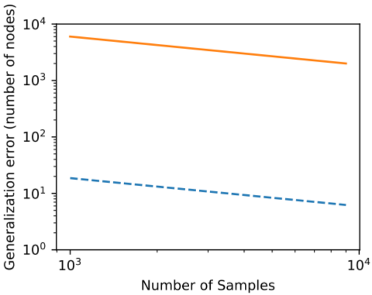
Stronger generalization guarantees

In practice, number of intervals partioning $[0,1] << 2^{n(n-1)/2} n^n$

- Derive stronger generalization guarantees.



(a) Linear separators



(b) Clustering

Related work

- Mostly experimental
- Node selection policy
- Pruning policy

Thank you