Appendix To Evasion and Hardening of Tree Ensemble Classifiers

Alex Kantchelian AKANT@CS.BERKELEY.EDU
J. D. Tygar TYGAR@CS.BERKELEY.EDU
Anthony D. Joseph ADJ@CS.BERKELEY.EDU
University of California, Berkeley

1. Proof that the feasibility subproblem of (1) is NP-Complete

First, given an instance x, computing the sign of $f(x)$ can be done in time at most proportional to the model size. Thus the feasibility problem is in NP. It is further NP-complete by a linear time reduction from 3-SAT as follows. We encode in x the assignment of values to the variables of the 3-SAT instance S. By convention, we choose $x_i > 0.5$ if and only if variable i is set to true in S. Next, we construct f by arranging each clause of S as a binary regression tree. Each regression tree has exactly one internal node per level, one for each variable appearing in the clause. Each internal node holds a predicate of the form $x_i > 0.5$ where i is a clause variable. The nodes are arranged such that there exists a unique prediction path corresponding to the falseness of the clause. For this path, the prediction value of the leaf is set to the opposite of the number of clauses in S, which is also the number of trees in the reduction. The remaining leaves predictions are set to 1. Figure 1 illustrates this construction on an example.

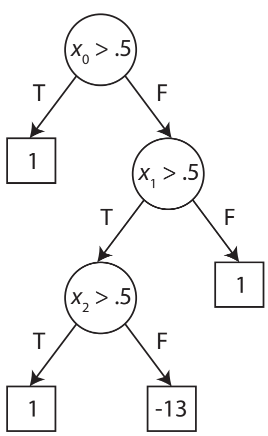

Figure 1. Regression tree for the clause $x_0 \vee \neg x_1 \vee x_2$. In this example, S has 13 clauses.

It is easy to see that S is satisfiable if and only if there exists x such that $f(x) > 0$. Indeed, a satisfying assignment for S corresponds to x such that $f(x) = |\mathcal{T}| > 0$ and any non-satisfying assignment for S corresponds to x such that $f(x) \le -1 < 0$ because there is at least one false clause which corresponds to a regression tree which output is $-|\mathcal{T}|$.

2. Objective weights

Recall that for each feature dimension $1 \le k \le n$, we have a collection of predicate variables $(\boldsymbol{p}_i)_{i=1..K}$ associated with predicates $x'_k < \tau_1, \ldots, x'_k < \tau_K$ where the thresholds are sorted $\tau_1 < \cdots < \tau_K$. Thus, the $\boldsymbol{p}$ variables effectively encode the interval to which x'_k belongs to, and any feature value within the interval will lead to the same prediction $f(x')$. There are exactly $K+1$ distinct possible valuations for the binary variables $\boldsymbol{p}_1 \le \boldsymbol{p}_2 \le \cdots \le \boldsymbol{p}_K$ and the value domain mapping $\phi : \boldsymbol{p} \to (\mathbf{R} \cup \{-\infty; \infty\})^2$ is:

$$x'_k \in \phi(\boldsymbol{p}) = [\tau_i, \tau_{i+1})$$
$$i = \max\{k | \boldsymbol{p}_k = 0, 0 \le k \le K+1\}$$

where by convention $\boldsymbol{p}_0 = 0$, $\boldsymbol{p}_{K+1} = 1$ and $\tau_0 = -\infty$, $\tau_{K+1} = \infty$. Setting aside the L_∞ case for now, consider $\rho \in \mathbb{N}$ the norm we are interested in for d. Instead of directly minimizing $\|x - x'\|_\rho$, our formulation equivalently minimizes $\|x - x'\|_\rho^\rho$. By minimizing the latter, we are able to consider the contributions of each feature dimension independently:

$$\|x - x'\|_\rho^\rho = \sum_{k=1}^{n} |x_k - x'_k|^\rho$$

We take $0^0 = 0$ by convention. At the optimal solution, $|x_k - x'_k|^\rho$ can only take $K+1$ distinct values. Indeed, if x'_k and x_k belong to the same interval, then $x'_k = x_k$ minimizes the distance along feature k, and this distance is zero. If x'_k and x_k do not belong to same interval, then setting x'_k at the border of $\phi(\boldsymbol{p})$ that is closest to x_k minimizes the distance along k. If $\phi(\boldsymbol{p}) = [\tau_i, \tau_{i+1})$, this distance is simply equal to $\min\{|x_k - \tau_i|^\rho, |x_k - \tau_{i+1}|^\rho\}$. Note that because of the right-open interval, the minimum distance is actually an infimum. In our implementation, we simply use a guard value $\epsilon = 10^{-4}$ of the same magnitude order than the numerical tolerance of the MILP solver.

Hence, we can express the minimization objective of problem (1) as a weighted sum of $\boldsymbol{p}$ variables without loss of generality. Let $0 \leq j \leq K+1$ be the indices such that $x_k \in [\tau_j, \tau_{j+1})$. Let $(w_i)_{i=0..K+1}$ such that for any valid valuation of $\boldsymbol{p}$ we have $\sum_{i=0}^{K+1} w_i \boldsymbol{p}_i = \inf_{x'_k \in \phi(\boldsymbol{p})} |x_k - x'_k|^\rho$. By the discussion above and exhaustively enumerating the $K+1$ valuations of $\boldsymbol{p}$, w is the solution to the following $K+1$ equations:

$$\begin{aligned}
w_{K+1} &= |x_k - \tau_K|^\rho \\
w_K + w_{K+1} &= |x_k - \tau_{K-1}|^\rho \\
&\dots \\
w_{j+1} + \cdots + w_{K+1} &= |x_k - \tau_{j+1}|^\rho \\
w_j + w_{j+1} + \cdots + w_{K+1} &= 0 \\
w_{j-1} + w_j + w_{j+1} + \cdots + w_{K+1} &= |x_k - \tau_j - \epsilon|^\rho \\
&\dots \\
w_1 + w_2 + w_3 + \cdots + w_{K+1} &= |x_k - \tau_2 - \epsilon|^\rho \\
w_0 + w_1 + w_2 + w_3 + \cdots + w_{K+1} &= |x_k - \tau_1 - \epsilon|^\rho
\end{aligned}$$

Note that this system of linear equations is already in triangular form and obtaining the w values is immediate. To obtain the full MILP objective, we repeat this process for every feature $1 \leq k \leq n$ and take the sum of all weighted sums of subsets of $\boldsymbol{p}$.

Finally, for the L_∞ case, we use 1 continuous variable $\boldsymbol{b}$. We introduce n additional constraints to the formulation, one for each feature dimension k. As per the previous discussion, we can generate the weights w such that $\sum_{i=0}^{K+1} w_i \boldsymbol{p}_i = \inf_{x'_k \in \phi(\boldsymbol{p})} |x_k - x'_k|$ (this is the $\rho = 1$ case). The additional constraint on dimension k is then:

$$\sum_{i=0}^{K+1} w_i \boldsymbol{p}_i \leq \boldsymbol{b}$$

and the MILP objective is simply the variable $\boldsymbol{b}$ itself.