Quickly Boosting Decision Trees – Supplementary Material

Ron Appel Thomas Fuchs Piotr Dollár Pietro Perona

In the main text of the paper, we prove a bound on the misclassification error given a preliminary m-subset of the data, given as Proposition 1:

$$m \le n \quad \Rightarrow \quad Z_m \varepsilon_m \le Z_n \varepsilon_n$$

Using this bound, we are able to prune features early and speed up the training of decision trees using the classification error criterion. In this document, we prove the same bound on other common types of stump splitting criteria: information gain, Gini purity, and variance minimization, extending our method for use with regression trees as well as binary or multi-class classification trees.

In any decision tree, an input propagates down the tree until it reaches a single leaf node. Recall that an n-subset is the set of n heaviest datapoints - the n samples with the largest weights. Let ρ be the set of elements in the n-subset that are assigned to a particular leaf. The sum of the weights of the elements that belong to that leaf is Z_ρ and the sum of the weights of the elements in ρ with class y is Z_ρ^y

$$Z_n \equiv \sum_{i \le n} w_i \qquad Z_\rho \equiv \sum_{i \in \rho} w_i \qquad Z_\rho^y \equiv \sum_{i \in \rho} w_i \, \mathbf{1}_{\{y_i = y\}}$$

In regression, elements have values $\mathbf{y}_i \in \mathbb{R}^d$, and we define the weighted average regression value of a leaf as:

$$\tilde{\mathbf{y}}_\rho \equiv \frac{1}{Z_\rho} \sum_{i \in \rho} w_i \, \mathbf{y}_i$$

Given an n-subset of data (or possibly all of it), the total error is computed by summing the error in each leaf, proportionally weighted by the total mass of samples belonging to that leaf:

$$\varepsilon_n = \sum_j \frac{Z_\rho}{Z_n} \varepsilon_\rho \quad \text{which we reformulate as:} \;\; Z_n \varepsilon_n = \sum_j Z_\rho \varepsilon_\rho$$

[for brevity, we have omitted the subscript j in what should be ρ_j]

Hence, we only need to show that for each leaf ρ, the product of total error and subset mass always exceeds that of a smaller subset. Let u be the set of elements in ρ that are also in an m-subset and $\bar{u}$ be the set of elements in ρ that are not in the m-subset, where $m \le n$:

$$u \equiv \{i \,|\, i \in \rho, i \le m\}, \;\; \bar{u} \equiv \{i \,|\, i \in \rho, i > m\} \qquad \text{Note that:} \;\; u \cup \bar{u} = \rho$$

$$Z_u \equiv \sum_{i \in u} w_i \quad Z_{\bar{u}} \equiv \sum_{i \in \bar{u}} w_i \quad Z_u^y \equiv \sum_{i \in u} w_i \, \mathbf{1}_{\{y_i = y\}} \quad Z_{\bar{u}}^y \equiv \sum_{i \in \bar{u}} w_i \, \mathbf{1}_{\{y_i = y\}}$$

Accordingly, in the following sections, it is enough to show that $Z_\rho \varepsilon_\rho \ge Z_u \varepsilon_u$ since:

$$Z_\rho \varepsilon_\rho \ge Z_u \varepsilon_u \quad \Rightarrow \quad Z_n \varepsilon_n = \sum_j Z_\rho \varepsilon_\rho \ge \sum_j Z_u \varepsilon_u = Z_m \varepsilon_m$$

Each of the following proofs is based on an inequality. All inequalities are proven at the very end. We also note that these proofs apply to trees of any depth, not just stumps.

Information Gain

$$\varepsilon_n \equiv \sum_j \frac{Z_\rho}{Z_n}\Big(-\sum_y \frac{Z_\rho^y}{Z_\rho}\ln\Big(\frac{Z_\rho^y}{Z_\rho}\Big)\Big) \qquad \Rightarrow \quad Z_n\varepsilon_n = \sum_j \overbrace{\Big(-\sum_y Z_\rho^y \ln\Big(\frac{Z_\rho^y}{Z_\rho}\Big)\Big)}^{Z_\rho\varepsilon_\rho}$$

$$Z_\rho\varepsilon_\rho = -\sum_y Z_\rho^y \ln\Big(\frac{Z_\rho^y}{Z_\rho}\Big) = -\sum_y (Z_u^y + Z_{\bar{u}}^y)\ln\Big(\frac{Z_u^y + Z_{\bar{u}}^y}{Z_u + Z_{\bar{u}}}\Big) \geq \underbrace{\Big(-\sum_y Z_u^y \ln\Big(\frac{Z_u^y}{Z_u}\Big)\Big)}_{Z_u\varepsilon_u} + \underbrace{\Big(-\sum_y Z_{\bar{u}}^y \ln\Big(\frac{Z_{\bar{u}}^y}{Z_{\bar{u}}}\Big)\Big)}_{Z_{\bar{u}}\varepsilon_{\bar{u}} \geq 0}$$

The proof for Information Gain Ratio is a trivial adaptation of the proof above.

Gini Impurity

$$\varepsilon_n \equiv \sum_j \frac{Z_\rho}{Z_n}\sum_y \frac{Z_\rho^y}{Z_\rho}\left(1 - \frac{Z_\rho^y}{Z_\rho}\right) \qquad \Rightarrow \quad Z_n\varepsilon_n = \sum_j \overbrace{\left(Z_\rho - \sum_y \frac{(Z_\rho^y)^2}{Z_\rho}\right)}^{Z_\rho\varepsilon_\rho}$$

$$Z_\rho\varepsilon_\rho = Z_\rho - \sum_y \frac{(Z_\rho^y)^2}{Z_\rho} = (Z_u + Z_{\bar{u}}) - \sum_y \frac{(Z_u^y + Z_{\bar{u}}^y)^2}{Z_u + Z_{\bar{u}}} \geq \underbrace{\left(Z_u - \sum_y \frac{(Z_u^y)^2}{Z_u}\right)}_{Z_u\varepsilon_u} + \underbrace{\left(Z_{\bar{u}} - \sum_y \frac{(Z_{\bar{u}}^y)^2}{Z_{\bar{u}}}\right)}_{Z_{\bar{u}}\varepsilon_{\bar{u}} \geq 0}$$

Variance Minimization

$$\varepsilon_n \equiv \sum_j \frac{Z_\rho}{Z_n}\sum_{i\in\rho}\frac{w_i}{Z_\rho}|\mathbf{y}_i - \tilde{\mathbf{y}}_\rho|^2 \qquad \Rightarrow \quad Z_n\varepsilon_n = \sum_j \overbrace{\Big(\sum_{i\in\rho} w_i|\mathbf{y}_i|^2 - \frac{|Z_\rho\,\tilde{\mathbf{y}}_\rho|^2}{Z_\rho}\Big)}^{Z_\rho\varepsilon_\rho}$$

$$\begin{aligned} Z_\rho\varepsilon_\rho = \sum_{i\in\rho} w_i|\mathbf{y}_i|^2 - \frac{|Z_\rho\,\tilde{\mathbf{y}}_\rho|^2}{Z_\rho} &= \sum_{i\in u} w_i|\mathbf{y}_i|^2 + \sum_{i\in\bar{u}} w_i|\mathbf{y}_i|^2 - \frac{|Z_u\,\tilde{\mathbf{y}}_u + Z_{\bar{u}}\,\tilde{\mathbf{y}}_{\bar{u}}|^2}{(Z_u + Z_{\bar{u}})} \\ &\geq \underbrace{\left(\sum_{i\in u} w_i|\mathbf{y}_i|^2 - \frac{|Z_u\,\tilde{\mathbf{y}}_u|^2}{Z_u}\right)}_{Z_u\varepsilon_u} + \underbrace{\left(\sum_{i\in\bar{u}} w_i|\mathbf{y}_i|^2 - \frac{|Z_{\bar{u}}\,\tilde{\mathbf{y}}_{\bar{u}}|^2}{Z_{\bar{u}}}\right)}_{Z_{\bar{u}}\varepsilon_{\bar{u}} \geq 0} \end{aligned}$$

Inequalities

For positive scalars $a, b \geq 0$ and $\alpha, \beta > 0$, the following inequality holds:

$$\begin{aligned} (a+b)\ln\Big(\frac{a+b}{\alpha+\beta}\Big) \leq a\ln\Big(\frac{a}{\alpha+\beta}\Big) + b\ln\Big(\frac{b}{\alpha+\beta}\Big) &= a\ln\Big(\frac{a}{\alpha}\cdot\frac{\alpha}{\alpha+\beta}\Big) + b\ln\Big(\frac{b}{\beta}\cdot\frac{\beta}{\alpha+\beta}\Big) \\ &= a\ln\Big(\frac{a}{\alpha}\Big) + b\ln\Big(\frac{b}{\beta}\Big) - \underbrace{\Big(a\ln\Big(1+\frac{\beta}{\alpha}\Big) + b\ln\Big(1+\frac{\alpha}{\beta}\Big)\Big)}_{\geq 0} \end{aligned}$$

$$\Rightarrow \quad (a+b)\ln\Big(\frac{a+b}{\alpha+\beta}\Big) \leq a\ln\Big(\frac{a}{\alpha}\Big) + b\ln\Big(\frac{b}{\beta}\Big)$$

For any vectors (or scalars) $\mathbf{a}, \mathbf{b}$ and positive scalars $\alpha, \beta > 0$, the following inequality holds:

$$\frac{|\mathbf{a}+\mathbf{b}|^2}{\alpha+\beta} = \frac{|\mathbf{a}|^2 + |\mathbf{b}|^2}{\alpha+\beta} + \frac{2\langle\mathbf{a},\mathbf{b}\rangle}{\alpha+\beta} = \frac{|\mathbf{a}|^2}{\alpha}\Big(1 - \frac{\beta}{\alpha+\beta}\Big) + \frac{|\mathbf{b}|^2}{\beta}\Big(1 - \frac{\alpha}{\alpha+\beta}\Big) + \frac{2\langle\mathbf{a},\mathbf{b}\rangle}{\alpha+\beta} = \frac{|\mathbf{a}|^2}{\alpha} + \frac{|\mathbf{b}|^2}{\beta} - \underbrace{\frac{|\beta\,\mathbf{a} - \alpha\,\mathbf{b}|^2}{\alpha\beta(\alpha+\beta)}}_{\geq 0}$$

$$\Rightarrow \quad \frac{|\mathbf{a}+\mathbf{b}|^2}{\alpha+\beta} \leq \frac{|\mathbf{a}|^2}{\alpha} + \frac{|\mathbf{b}|^2}{\beta}$$