Self-Terminating Induction of Multiclass Trees (Abstract)

Sally Goldman
Google Research
1600 Ampitheatre Parkway
Mountain View, CA 94043
sgoldman@google.com

Yoram Singer
Google Reserach
1600 Ampitheatre Parkway
Mountain View, CA 94043
singer@google.com

February 9, 2011

Last year at Snowbird, we introduced a self-terminating generalization of a decision tree in which each node s has associated with it both a predicate π_s used for branching, as in a standard decision tree, and a real value α_s. For a prediction tree T we define the *norm variation complexity* $V^p(T)$ as $V^p(T) := \sum_{s \in T} \lambda(s) \left\| \alpha_{C(s)} \right\|_p$ where $C(s)$ is the set of children of s and $\lambda(s)$ is a penalty for node s (e.g. its depth). By convention $\alpha = 0$ for null children. We consider both $p = 1$ and $p = \infty$ to encourage small decision trees. When using the ℓ_1 regularizer, we get edge-based termination, in which only some of the edges are expanded. The motivation for the ℓ_∞ regularizer is that we wish to induce a sparse solution in which children $C(s)$ of a node s are zero. However, if the optimal solution is such that at least some $\alpha_{s'}$ for $s' \in C(s)$ are non-zero, then the rest of the children can be non-zero as well without incurring further penalty. Let f_T be the function represented by the prediction tree T, where for example $\boldsymbol{x}$, $f_T(\boldsymbol{x})$ is simply the sum of the α values on the path from the root of T to the leaf reached by $\boldsymbol{x}$. Let $\widehat{R}_n(L, f, w)$ denote the empirical risk of the function f with loss L weighted by $w \succeq 0$, that is, given examples $\boldsymbol{x}_i$ and labels y_i, $\widehat{R}_n(L, f, w) := \sum_{i=1}^n w_i L(f(\boldsymbol{x}_i), y_i)$. Our goal is to minimize the penalized empirical risk

$$\widehat{R}_n(L, f, w) + V^p(T) = \sum_{i=1}^n w_i L(f(\boldsymbol{x}_i), y_i)) + V^p(T). \tag{1}$$

The end result is a learning algorithm for decision and prediction trees in which growing and termination happen concurrently and are tightly coupled. Also our approach is "backward compatibility" with existing tree learning procedures. Upon omitting the variation penalty, we obtain well known growing criteria such as the information gain and the Gini index (Breiman et al. (1984); Quinlan (1986)).

This year we will present an extension of this work for multiclass problems. For any node s in the prediction tree, let $\mathcal{P}_s(\boldsymbol{x})$ be the nodes along the path from the root to s when evaluating example $\boldsymbol{x}$. In the classification setting, each node s is associated with a bias value $b = \sum_{v \in \mathcal{P}_s} \alpha_v$. When using the log loss as our empirical loss, we can view b as a prior distribution over the target label where the probability of the label being 1 is $u = 1/(1 + e^{-b})$ for all examples that reach the node s. In the multiclass setting we instead need to represent the label distribution as a probability vector, $\boldsymbol{u}$, rather than a single scalar. So we replace the *single* scalar α associated with each node by a vector $\boldsymbol{\alpha}$. The distribution induced over the labels takes the form $p_i \sim e^{b_i + \alpha_i}$. Our goal is to further endow the self-terminating property and promote solutions where the *entire* vector $\boldsymbol{\alpha}$ is zero in the lack of strong empirical evidence. To do so, we use the ℓ_∞ regularization which promotes *group* sparsity (Negahban and Wainwright (2008)).

We overview our optimization algorithm focusing on a single branch from s with prior $\boldsymbol{u}$ for which $\boldsymbol{q}$ is the empirical distribution over the labels following that branch. Let $w_{ij} = w_i$ when example i follows branch j and $w_{ij} = 0$ otherwise. Let $q_k = \frac{1}{\kappa} \sum_{u: y_{i,j} = k} w_{i,j}$, where κ is a normalization constant which ensures that $\boldsymbol{q}$ is a proper distribution. Our goal is to determine the (posterior) distribution $\boldsymbol{p}$ of the

labels for child node residing at the branch. This posterior distribution becomes in turn the prior $\boldsymbol{u}$ as we proceed to perform the growing procedure at the child node. Formally, the multiclass penalized risk minimization for the logistic loss amounts to minimizing $-\sum_i q_i \log p_i + \lambda \|\alpha\|_\infty$ where $p_i \sim e^{\alpha_i + b_i}$. Finding the optimal solution of this problem is not an easy task due to the ℓ_∞ penalty. We solve instead its Legendre dual, which is $\min_{\boldsymbol{\gamma}} \sum_i ((q_i - \gamma_i)\log(q_i - \gamma_i) + \gamma_i \log u_i)$ such $\|\boldsymbol{\gamma}\|_1 \leq \lambda$ and $\sum_i \gamma_i = 0$. To solve the dual form we introduce a Lagrange multiplier $\theta \geq 0$ for the ℓ_1 constraint and δ for the constraint that $\sum_i \gamma_i = 0$, to get $\min_\gamma \sum_i ((q_i - \gamma_i)\log(q_i - \gamma_i) + \gamma_i \log u_i + \theta(\|\boldsymbol{\gamma}\|_1 - \lambda) + \delta \sum_i \gamma_i$. Denoting $s_i = \text{sign}(\gamma_i)$, and using the sub-gradient optimality condition with respect to $\boldsymbol{\gamma}$ yields that,

$$p_i = q_i - \gamma_i = \begin{cases} u_i e^{\theta}/z & \gamma_i > 0 \\ u_i e^{-\theta}/z & \gamma_i < 0 \\ q_i & \gamma_i = 0 \end{cases} , \tag{2}$$

where z is the standard normalization (partition function) which ensures that $\boldsymbol{p}$ is a proper distribution. Eq. (2) underscores the relation between $\boldsymbol{\gamma}$ and $\boldsymbol{p}$. Specifically, Eq. (2) implies that when $\gamma_i > 0$, $u_i \leq p_i < q_i$, and for $\gamma_i < 0$, $u_i \geq p_i > q_i$. In words, the solution $\boldsymbol{p}$ lies between $\boldsymbol{q}$ and $\boldsymbol{u}$ where the lower and upper bounds on each coordinate in $\boldsymbol{p}$ depends on the relation between the corresponding components in $\boldsymbol{q}$ and $\boldsymbol{u}$.

Let I_+ be the set of indices for which $\gamma_i > 0$, I_- be the set of indices for which $\gamma_i < 0$, and I_0 be the set of indices for which $\gamma_i = 0$. Define $Q_+ = \sum_{i \in I^+} q_i$, $Q_- = \sum_{i \in I^-} q_i$, and similarly, $U_+ = \sum_{i \in I^+} u_i$, $U_- = \sum_{i \in I^-} u_i$. Combining Eq. (2) with the constraint that $\sum_i \gamma_i = 0$ (which stems from the requirement $\sum_i p_i = 1$) yields $(e^{\theta} U_+ + e^{-\theta} U_-)/z = Q_+ + Q_-$. Similarly, combining Eq. (2) with the constraint $\sum_i |\gamma_i| = \lambda$ yields $(-e^{\theta} U_+ + e^{-\theta} U_-)/z = \lambda - Q_+ + Q_-$. Combining the last two equalities gives a closed form solution for θ and z,

$$\theta = \frac{1}{2} \log \left(\frac{(Q_+ - \lambda/2)\, U_-)}{(Q_- + \lambda/2)\, U_+} \right) \ , \ z = \frac{e^{\theta} U_+ + e^{-\theta} U_-}{Q_+ + Q_-} \quad . \tag{3}$$

We can further characterize the structure of the correct partition of the components of $\boldsymbol{\gamma}$ into the sets I_+, I_-, I_0. From Eq. (2) it immediately follows that when $\gamma_i > 0$, $\log(p_i/u_i) + \log z = \theta$ and when $\gamma_i < 0$, $\log(p_i/u_i) + \log z = -\theta$. Furthermore, by applying the KKT conditions for optimality, the following property holds,

$$|\log(q_i/u_i) + \log z| < \theta \Rightarrow \gamma_i = 0 \quad . \tag{4}$$

We can combine these properties to obtain an efficient algorithm for finding this optimal partition. First we sort the components according to the ratios q_i/u_i. Without loss of generality, assume that $q_1/u_1 \leq q_2/u_2 \leq \cdots \leq q_n/u_n$, where n is the number of different labels. From Eq. (4) we know that there must exist two indices r and s such that $1 \leq r < s \leq n$ and $q_r/u_r < 1$ and $q_s/u_s > 1$. In turn, these ratio properties imply that that for $j \leq r$, $\gamma_j < 0$, $\gamma_{r+1} = \ldots = \gamma_{s-1} = 0$, and for $j \geq s$, $\gamma_j > 0$. Further, for a given proposed partition, we can efficiently compute the solution corresponding to that partition by combining Eq. (2) and Eq. (3). Finally, from Eq. (4), it is clear that a candidate partition is optimal iff $\theta > 0$ and for all i such that $|\log(q_i/u_i) + \log z| < \theta$, the value of γ_i is zero.

References

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, 1984.

S. Negahban and M. Wainwright. Phase transitions for high-dimensional joint support recovery. In *Advances in Neural Information Processing Systems 22*, 2008.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.