Finding Influential Training Samples for Gradient Boosted Decision Trees

Sharchilev, B.; Ustinovsky, Y.; Serdyukov, P.; de Rijke, M.

Publication date
2018

Document Version
Final published version

Published in
Proceedings of Machine Learning Research

License
Other

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Finding Influential Training Samples for Gradient Boosted Decision Trees

Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, Maarten de Rijke

Abstract

We address the problem of finding influential training samples for a particular case of tree ensemble-based models, e.g., Random Forest (RF) or Gradient Boosted Decision Trees (GBDT). A natural way of formalizing this problem is studying how the model’s predictions change upon leave-one-out retraining, leaving out each individual training sample. Recent work has shown that, for parametric models, this analysis can be conducted in a computationally efficient way. We propose several ways of extending this framework to non-parametric GBDT ensembles under the assumption that tree structures remain fixed. Furthermore, we introduce a general scheme of obtaining further approximations to our method that balance the trade-off between performance and computational complexity. We evaluate our approaches on various experimental setups and use-case scenarios and demonstrate both the quality of our approach to finding influential training samples in comparison to the baselines and its computational efficiency.

1. Introduction and Background

As machine learning-based models become more widespread and grow in both scale and complexity, methods of interpreting their predictions are increasingly attracting attention from the machine learning community. Some of the applications and benefits of employing these methods outlined in previous work include (1) “debugging” the model to expose ways of model failures not discoverable via conventional test set performance measuring (e.g., data or target leakages); (2) boosting developer’s trust in the model’s performance in scenarios when on-line evaluation is not available before deployment; and (3) increasing user satisfaction and/or confidence in provided predictions, etc. Various problem setups (Palczewska et al., 2013; Tolomei et al., 2017; Fong & Vedaldi, 2017) and interpretation methods, both model-agnostic (Ribeiro et al., 2016; Lundberg & Lee, 2017) and model-specific (Shrikumar et al., 2017; Tolomei et al., 2017; Sundararajan et al., 2017), have recently been proposed in the literature.

A common trait shared by the majority of these methods is that they do not provide a way of automatically improving the model, since the model is fixed; the main use-case thus becomes manual analytics by the user or the developer, which is both time and resource-consuming. It is thus desirable to derive a framework for obtaining actionable insights into the model’s behavior allowing us to automatically improve a model’s performance.

One such framework has recently been introduced by Koh & Liang (2017); it deals with finding the most influential training objects. They formalize the notion of “influence” via an infinitesimal approximation to leave-one-out retraining: the core question that this work aims to answer is “how would the model’s performance on a test object $x_{test}$ change if the weight of a training object $x_{train}$ is perturbed?” Assuming a smooth parametric model family (e.g., linear models or neural networks), the authors employ the Influence Functions framework from classical statistics (Cook & Weisberg, 1980; also see Koh & Liang (2017) for a literature review on the topic) to show that this quantity can be estimated much faster than via straightforward model retraining, which makes their method tractable in a real-world scenario. A natural use-case of such a framework is to consider individual test objects (or groups of them) on which the model performs poorly and either remove the most “harmful” training objects or prioritize a batch of new objects for labeling based on which ones are expected to be the most “helpful,” akin to active learning.

Unfortunately, the method suggested by Koh & Liang (2017) heavily relies on the smooth parametric nature of the model family. While this is a large class of machine learning models, it is by far not the only one. In particular, decision tree ensembles such as Random Forests (Ho, 1995; RF) and Gradient Boosted Decision Trees (Friedman, 2001; GBDT) are probably the most widely used model family in industry, largely due to their state-of-the-art performance on struc-
tured and/or multimodal data. Thus, it is important to extend the aforementioned Influence Functions framework to tree ensembles.

In this paper, we propose a way of doing so, while focusing specifically on GBDT. We consider two proxy metrics for the informal notion of influence. For the first one, leave-one-out retraining, we utilize the inner mechanics of fitting decision trees (in particular, assuming that a small training sample perturbation does not change the trees’ structures) to derive LeafRefit and FastLeafRefit, a well-founded family of approximations to leave-one-out retraining that trade off approximation accuracy for computational complexity. For the second, analogously to the Influence Functions framework, we consider infinitesimal training sample weight perturbations and derive LeafInfluence and FastLeafInfluence, methods for estimating gradients of the model’s predictions with respect to training objects’ weights. From a theoretical perspective, LeafInfluence and FastLeafInfluence allow us to deal with the discontinuous dependency of tree structure on training sample perturbations; from a practical one, they allow us to further reduce computational complexity due to the possibility of precomputing certain derivatives.

In our experiments we (1) study the conditions under which our methods, FastLeafRefit and FastLeafInfluence, successfully approximate their proxy metrics, (2) demonstrate our methods’ ability to target training objects which are influential for specific test objects, and (3) show that our algorithms run much faster than straightforward retraining, which makes them applicable in practical scenarios.

2. Problem Definition

First, we formally define the problem setup. We consider standard supervised training of a GBDT ensemble
\[ F(x; w) := \sum_{t=1}^{T} f_t P(x), (A^{t-1}) \]
on a training sample \( X_{train} \). Learning consists of two separate stages: model structure selection and picking the optimal leaf values. The way of choosing the model structure is not important for our work; we refer the interested reader to existing implementations, e.g., [Chen & Gu Estrin 2016, Dorogush et al. 2017]. For picking optimal leaf values, we consider two most commonly used formulas:

Gradient: At leaf \( l \) at step \( t \), output negative average gradients (calculated at current predictions) over the leaf objects:
\[ f_{G}^{t} \]

This is equivalent to minimizing the empirical loss function w.r.t. the current leaf value by doing a single Newton step in function space [Chen & Gu Estrin 2016].

Newton: At leaf \( l \) at step \( t \), output the negative total gradient divided by the total second derivative over the leaf objects:
\[ f_{H}^{t} := \frac{-G_{l}^{t}(A^{t-1})}{H_{l}^{t}(A^{t-1})}. \]

This is equivalent to minimizing the empirical loss function w.r.t. the current leaf value by doing a single Newton step in function space [Chen & Gu Estrin 2016].

3. Approach

In this section, we describe our approach to efficiently calculating the influence of the current leaves value. Since the notion of “influence” is not rigorously defined and partly intuitive, we need to introduce a well-defined, measurable quantity that aims to capture the desired intuition; we refer to it as a proxy for influence. In this work, we follow the general framework of [Koh & Liang 2017] and quantify influence through train set perturbations. We consider two proxies that reflect two natural variations of this approach. First, we describe an algorithm for faster exact leave-one-out retraining of GBDT under the assumption that the model structure remains fixed, and explain how to use that framework for estimating the influence of training points on specific test samples. Secondly, we derive an iterative algorithm to compute gradients of GBDT predictions w.r.t. the weights of training sample and analyze the resulting expressions.
3.1. Leave-One-Out Retraining

For the first proxy, following Koh & Liang (2017), we quantify the (negative) influence of a training sample \( x_{\text{train}} \) on a model’s prediction on a test sample \( F(x_{\text{test}}; w) \) as the change of loss on \( x_{\text{test}} \) after retraining the model without \( x_{\text{train}} \).

**Proxy 1.** In\( \text{f} \text{grad}(x_{\text{train}}, x_{\text{test}}) := L(y_{\text{test}}, F(x_{\text{test}})) - L(y_{\text{test}}, F_{\hat{x}_{\text{train}}}(x_{\text{test}})) \), where \( F_{\hat{x}_{\text{train}}} \) is the model retrained without \( x_{\text{train}} \).

Since, in order to rank the training points according to In\( \text{f} \text{grad}(x_{\text{train}}, x_{\text{test}}) \), we would have to compute Proxy 1 for each \( x_{\text{train}} \), straightforward leave-one-out retraining would be prohibitively expensive even for moderately-sized datasets. Moreover, as mentioned in Section 1, the parametric model framework of Koh & Liang (2017) is not directly applicable here. Thus, a solution tailored specifically for tree ensembles is required.

3.1.1. LeafRefit

In the problem definition (Section 3) we noted that training each tree requires picking its structure and leaf values. Moreover, these two operations respond to small training set perturbations differently: the tree structure is piecewise constant (i.e., it either stays the same or changes abruptly), whereas leaf values change more smoothly. Thus, a natural assumption to make is:

**Assumption 1.** The effect of removing a single training point can be estimated while treating each tree’s structure as fixed.

Under Assumption 1, it is thus sufficient to estimate how the leaf values of each tree are going to change. Since selecting optimal feature splits, e.g., via CART (Quinlan 1986) or C4.5 (Quinlan 2014) algorithms, is often the computational bottleneck in fitting decision trees, this observation already yields a significant complexity reduction.

Thus, our first algorithm for approximate leave-one-out retraining, LeafRefit, is equivalent to fixing the structure of every tree and fitting leaf values without the removed point. A formal listing of the resulting algorithm is given in Algorithm 1.

Note that the effect of removing a training object \( x_i \) is twofold: on the first step, we have to (1) remove \( x_i \) from its leaf (Algorithm 1, line 7) and (2) recalculate the leaf values and record the resulting changes of intermediate predictions for each training object (line 14). Thus, despite improving upon straightforward retraining by not having to search for the optimal tree splits, LeafRefit is still an expensive algorithm. Running it for each training sample has an asymptotic complexity of \( O(T n^2) \); moreover, in practice, for each training step \( t \) it involves an expensive routine of recalculating derivatives for each training point.

Algorithm 1 LeafRefit

1. **Input:** training point index to remove \( i_t \), sample-to-leaf assignments \( \{I_{t,L}\}_{t=1,L=1} \), leaf formula type \( \text{formula} \)
2. **Output:** new leaf values \( \{\hat{f}_l\}_{t=1,L} \)
3. Initialize \( \Delta_i^0 \leftarrow 0, A_t^0 \leftarrow 0 \), \( i = 1 \ldots n \)
4. for \( t = 1 \) to \( T \)
5. \( A_t^{l-1} \leftarrow A_t^{l-1} + \Delta_i^t, i = 1 \ldots n \)
6. for \( l = 1 \) to \( L \)
7. \( I_l^t \leftarrow I_l^t \setminus \{i_t\} \)
8. if formula == Gradient then
9. \( \hat{f}_l^t \leftarrow f_l^t((\hat{A}_t^{l-1})_{i \in I_l^t}) \)
10. else
11. \( \Delta_i^t \leftarrow \hat{f}_l^t((\hat{A}_t^{l-1})_{i \in I_l^t}) \)
12. end if
13. \( \Delta_i^t \leftarrow \Delta_i^{l-1} + \Delta_i^t, i \in I_l^t \)
14. end for
15. end for
16. return \( \{\hat{f}_l\}_{t=1,L} \)

3.1.2. FastLeafRefit

We seek to limit the number of calculations at each step of LeafRefit. Note that, in LeafRefit, we generally cannot make any use of caching the original first and/or second derivatives, since any \( \Delta_i^{l-1} \) (Algorithm 1, line 14) can be nonzero, which forces us to recompute the derivatives for each object. We build on the intuition that, in practice, a lot of \( \Delta_i^{l-1} \) may be negligible; an extreme example is when training samples can be separated in disjoint cliques, i.e., \( I_{t_1}^{l_1} = I_{t_2}^{l_2} \forall t_1, t_2 = 1, \ldots, T, l = 1 \ldots L \). In this case, removing each training point only affects its clique \( I_{i_t}^{l_0} := I_{i_t}^{l_1} \). Thus, at each training step \( t \), we may select a subset of training samples \( U_t \) whose deltas we take into account, and suppose \( \hat{A}_i^{l_t} = A_i^{l_t} \forall i \notin U_t \). We refer to \( U_t \) as the *update set*. Combining this with caching the original \( A_i^{l_t} \) and sums of derivatives in each leaf, we reduce the asymptotic complexity to \( O(T n C) \), where \( C = \max_{i} |U_t| \), which is a significant reduction if \( C \ll n \). A formal listing of the resulting algorithm, FastLeafRefit, is given in Algorithm 2.

3.1.3. Selecting the Update Set

In Section 3.1.2 we introduced FastLeafRefit, an approximate algorithm potentially achieving lower complexity than LeafRefit. Its definition, however, allowed for an arbitrary choice of the update set \( U_t \) telling us which training points’ prediction changes to take into account at boosting step \( t \). It is intuitively clear that different strategies of selecting \( U_t \) allow us to optimize the trade-off between computational complexity and quality of approximating leave-one-out retraining; thus, FastLeafRefit provides a principled way of obtaining approximations of different rigor to LeafRefit. Natural strategies for selecting the update set include:

**SinglePoint:** don’t update any points’ predictions and only
Algorithm 2 FastLeafRefit

Input: \( i_0, \{I^t\}_{t=1}^{T,L}, \{g^t_i(A^{-1})\}_{t=1}^{T,n}, \{h^t_i(A^{-1})\}_{t=1}^{T,n}, \{G^t_i(A^{-1})\}_{t=1}^{T,n}, \{H^t_i(A^{-1})\}_{t=1}^{T,n} \), leaf formula type formula

Output: New leaf values \( \{\hat{f}^t_i\}_{t=1}^{T,L} \)

Initialize \( \Delta^0_i \leftarrow 0, i = 1 \ldots n 

for \( t = 1 \) to \( T \) do
  \( U^t \leftarrow \text{UpdateSet}(t) \)
  for \( l = 1 \) to \( L \) do
    \( U^t_l \leftarrow U^t_l \cap I^t_l \)
    \( \hat{f}^t_l \leftarrow \text{LeafRecalc}(t, l, \{I^t\}_{t=1}^{T,L}, \{g^t_i(A^{-1})\}_{t=1}^{T,n}, \{h^t_i(A^{-1})\}_{t=1}^{T,n}, \{G^t_i(A^{-1})\}_{t=1}^{T,n}, \{H^t_i(A^{-1})\}_{t=1}^{T,n}, \text{leaf formula type formula} \)
    \( \Delta \hat{f}^t_l \leftarrow \hat{f}^t_l - f^t_l \)
    \( \Delta^t_i \leftarrow \Delta^t_i + \Delta \hat{f}^t_i, i \in I^t_l \)
  end for
end for
return \( \{\hat{f}^t_i\}_{t=1}^{T,L} \)

Algorithm 3 LeafRecalc

Input: boosting step \( t \), leaf index \( l \), \( \{I^t\}_{t=1}^{T,L} \), \( \{g^t_i(A^{-1})\}_{t=1}^{T,n}, \{h^t_i(A^{-1})\}_{t=1}^{T,n}, \{G^t_i(A^{-1})\}_{t=1}^{T,n}, \{H^t_i(A^{-1})\}_{t=1}^{T,n} \), leaf formula type formula

Output: New leaf value \( \hat{f}^t_l \)

\( I \leftarrow \{i_0 \in I^t_l \} \)
\( \Delta g^t_j \leftarrow g^t_j(A^{-1}) + \Delta g^t_{j-1} - g^t_j(A^{-1}), j \in U^t_l \)
\( G^t_l \leftarrow G^t_l(A^{-1}) + \sum_{j \in U^t_l} w_j \Delta g^t_j - Iw_{i_0}g^t_{i_0}(A^{-1}) \)
if formula == Gradient then
  \( H^t_l \leftarrow \sum_{j \in U^t_l \{i_0 \}} w_j \)
else
  \( \Delta h^t_j \leftarrow h^t_j(A^{-1}) + \Delta h^t_{j-1} - h^t_j(A^{-1}), j \in U^t_l \)
  \( H^t_l \leftarrow H^t_l(A^{-1}) + \sum_{j \in U^t_l} w_j \Delta h^t_j - Iw_{i_0}h^t_{i_0}(A^{-1}) \)
end if
return \( \frac{G^t_l}{H^t_l} \)

3.2. Prediction gradients

In the previous sections we introduced LeafRefit and FastLeafRefit, fast methods of estimating the effect of a training sample on the GBDT ensemble, which can then be used to rank training points, e.g., by their influence on a test point of interest. Under Assumption[1] these methods are valid approximations of leave-one-out retraining, which gives them theoretical grounding. However, perhaps surprisingly, our experiments show that Assumption[1] fails quite often: in particular, for the Adult dataset (see Section 4.2), 58% of training points change the structures of at least one tree in the ensemble. Moreover, as shown in Section 4.3 when Assumption[1] is violated, LeafRefit and FastLeafRefit are no longer valid approximations to Proxy[1].

The intuition underlying Assumption[1] however, still holds: for a small enough perturbation to the training data, the structure will remain fixed, whereas leaf values will still be changing smoothly. Note that retraining the model without a sample \( i \) is equivalent to setting \( w_i = w_i^{old} + \Delta w_i; \Delta w_i = -w_i^{old} \). This change may be large enough to trigger structural shifts in the ensemble; thus, we need a tool to study a model’s response to smaller (arbitrarily small) perturbations.

An obvious choice for such a tool is the derivative of a model’s prediction w.r.t. a sample’s weight, which was also a crucial tool in the Influence Functions framework from classical statistics (Cook & Weisberg [1980]):

\[ \text{Proxy 2.} \quad \text{Inf}_{\text{grad}}(\mathbf{x}_{\text{train}}, \mathbf{x}_{\text{test}}) := \frac{\partial L(y_{\text{test}}, F(x_{\text{test}}))}{\partial w_i(\mathbf{x}_{\text{train}})} \]

where \( i(x_{\text{train}}) \) is the index of \( x_{\text{train}} \) in \( \mathbf{X}_{\text{train}} \).

3.2.1. LeafInfluence

As mentioned above, in the setup of Proxy[2] the statement of Assumption[1] is now guaranteed to hold and is no longer an assumption; we may consider the tree structures to be fixed and only study perturbations of leaf values, which smoothly
depend on the weights. Using the chain rule
\[
\frac{\partial L(y, F(x; w))}{\partial w_i} = \frac{\partial L(y, z)}{\partial z} \bigg|_{z=F(x; w)}, \quad \frac{\partial F(x; w)}{\partial w_i},
\]
we can then derive various counterfactuals (e.g., “how would the loss on a test point change if we overweight a training point?”), similarly to [Koh & Liang 2017]. Since we have
\[
\frac{\partial F(x; w)}{\partial w_i} = \sum_{t=1}^{T} \frac{\partial f_t^{i}(x)}{\partial w_i} (\mathbf{A}^{t-1}(\mathbf{w}))
\]
for applying Eq. 5 to arbitrary x (for a fixed i) it is necessary and sufficient to calculate \(\frac{\partial f_t^{i}(\mathbf{A}^{t-1}(\mathbf{w}))}{\partial w_i}\), \(t = 1 \ldots T, l = 1 \ldots L\). Applying Eq. 4 then can be done by running x though a new tree ensemble having \(\{\frac{\partial f_t^{i}(\mathbf{A}^{t-1}(\mathbf{w}))}{\partial w_i}\}_{t=1, t=1}\) as leaf values.

Expressions for leaf value derivatives depend on the type of leaf formula.

**Proposition 1.** Leaf value derivatives are given by:
\[
\begin{align*}
\frac{\partial f_t^{i} l_i(x)}{\partial w_i} &= -I_t^{i}(i) \sum_{j \in I_t^{i}} \sum_{l=1}^{L} w_i h_j^{i}(A_{j}^{t-1})_i, \\
\frac{\partial h_t^{i} l_i(x)}{\partial w_i} &= -I_t^{i}(i) (k_t^{i} f_{H,t}^{i} + g_t^{i}) \sum_{j \in I_t^{i}} w_i h_j^{i}(A_{j}^{t-1})_i,
\end{align*}
\]
where \(I_t^{i}(i) := I[i \in I_t^{i}]\) and \(J(A^{t}) := \frac{\partial A^{t}(w)}{\partial w_i}\).

**Proof.** First, let us derive the desired expression\(^3\) for \(f_t^{i} l_{G,t}(A^{t-1}(w), w)\):
\[
\begin{align*}
\frac{\partial f_t^{i} l_{G,t}(A^{t-1}(w), w)}{\partial w_i} &= -\frac{\partial G_t^{i}(A^{t-1}(w), w)}{H_t^{i}(A^{t-1}(w), w)} \\
&= \frac{-\frac{\partial G_t^{i}(A^{t-1}(w), w)}{\partial w_i} H_t^{i}(A^{t-1}(w), w) + \frac{\partial H_t^{i}(A^{t-1}(w), w)}{\partial w_i} G_t^{i}(A^{t-1}(w), w)}{H_t^{i}(A^{t-1}(w), w)^2}.
\end{align*}
\]

Let us calculate the derivatives of \(G_t^{i}(A^{t-1}(w), w)\) and \(H_t^{i}(A^{t-1}(w), w)\) separately:
\[
\begin{align*}
\frac{\partial G_t^{i}(A^{t-1}(w), w)}{\partial w_i} &= \sum_{j \in I_t^{i}} \left[ \delta_{ij} j_t^{i}(A_{j}^{t-1}(w)) + w_j h_j^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij} \right] \\
&= I_t^{i}(i) j_t^{i}(A_{i}^{t-1}(w)) + \sum_{j \in I_t^{i}} w_j h_j^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij}; \\
\frac{\partial H_t^{i}(A^{t-1}(w), w)}{\partial w_i} &= I_t^{i}(i)
\end{align*}
\]
Plugging this back into Equations 7 and grouping terms with and without \(I_t^{i}(i)\) separately, we get:
\[
\begin{align*}
\frac{\partial f_t^{i} l_{G,t}(A^{t-1}(w), w)}{\partial w_i} &= -I_t^{i}(i) f_t^{i} l_{H,t}(A^{t-1}(w), w) + \frac{\partial G_t^{i}(A^{t-1}(w), w)}{\partial w_i} \\
&= \sum_{j \in I_t^{i}} w_j h_j^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij},
\end{align*}
\]
which proves the first part of Proposition 1.

For the second part, all we have to change is to substitute \(H_t^{i} l_{H,t}(A^{t-1}(w), w)\) for \(H_t^{i} l_{G,t}(A^{t-1}(w), w)\). Its derivative is given by
\[
\begin{align*}
\frac{\partial h_t^{i} l_{G,t}(A^{t-1}(w), w)}{\partial w_i} &= \sum_{j \in I_t^{i}} \left[ \delta_{ij} j_t^{i}(A_{j}^{t-1}(w)) + w_j k_t^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij} \right] \\
&= I_t^{i}(i) h_t^{i}(A_{i}^{t-1}(w)) + \sum_{j \in I_t^{i}} w_j k_t^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij}.
\end{align*}
\]
Just like before, plugging it back into Equations 7 and grouping terms containing and not containing \(I_t^{i}(i)\) separately, we get:
\[
\begin{align*}
\frac{\partial h_t^{i} l_{H,t}(A^{t-1}(w), w)}{\partial w_i} &= -I_t^{i}(i) h_t^{i}(A_{i}^{t-1}(w)) f_t^{i} l_{H,t}(A^{t-1}(w), w) + \frac{\partial h_t^{i}(A^{t-1}(w), w)}{\partial w_i} \\
&= \sum_{j \in I_t^{i}} w_j k_t^{i}(A_{j}^{t-1}(w)) J(A^{t-1})_{ij}.
\end{align*}
\]
This concludes the proof of Proposition 1.

It can be seen from Eq. 6 that leaf value derivatives at step \(t\) depend on the Jacobi matrix \(J(A^{t-1})_{ij}\). These values, in turn, are connected by a recursive relationship:
\[
J(A^{t})_{ij} = J(A^{t-1})_{ij} + \frac{\partial f_t^{i l_{G,t}}(x)}{\partial w_i}.
\]
Thus, we can calculate leaf value derivatives in an iterative fashion similar to LeafRefit. A formal listing of the resulting algorithm, LeafInfluence, can be found in Algorithm 4.

Besides providing means for analyzing small weight perturbations, two important traits yielding complexity reductions can be seen from Eq. 6.

**A.** Using Eq. 6 we can write out \(\nabla w f_t^{i} l_{G,t} = \left(\frac{\partial f_t^{i l_{G,t}}}{\partial w_i}\right)_{i=1}\) in vector form; since computing each \(\frac{\partial f_t^{i l_{G,t}}}{\partial w_i}\) involves addition and a vector dot product, \(\nabla w f_t^{i}\) can then be expressed via vector addition and matrix/vector product for easy parallelization/vectorization.

**B.** The derivatives \(\{g_t^{i}, h_t^{i, k_t^{i}}\}_{t=1}^{T, n}\) used in Eq. 6 can now be precomputed only once during GBDT training and not for
We can select an update set $U$.

Algorithm 4 \textit{LeafInfluence}

\textbf{Inputs:} training point index $i_0$, sample-to-leaf assignments \( \{h_l^{T,L}\}_{l=1}^{T}, \{g_l^{T,L}\}_{l=1}^{T}, \{h_l^{0,I}\}_{l=1}^{T}, \{g_l^{0,I}\}_{l=1}^{T}, \{h_l^{I,L}\}_{l=1}^{T}, \{g_l^{I,L}\}_{l=1}^{T} \), leaf formula type formula

\textbf{Outputs:} leaf value derivatives \( \{\frac{\partial f_l^{I,L}(A^{t-1})}{\partial w_j}\}_{l=1}^{T},j=1...n \)

\begin{enumerate}
\item for $t=1$ to $T$
\item $\frac{\partial f_l^{I,L}(A^{t-1})}{\partial w_j}\leftarrow$l/According to Eq. 6, $l = 1 \ldots L$
\item $J(A^{t})_{ij}\leftarrow$l/According to Eq. 8, $i = 1 \ldots n, j = 1 \ldots n$
\end{enumerate}

return \( \{\frac{\partial f_l^{I,L}(A^{t-1})}{\partial w_j}\}_{l=1}^{T},j=1...n \)

each training object $i$ whose influence we want to compute. This contrasts \textit{LeafInfluence} with \textit{LeafRefit} and \textit{FastLeafRefit}, where these derivatives had to be recalculated for each $i$ depending on the values of $\Delta_j^{t-1}$, which change for different $i$.

3.2.2. \textsc{FastLeafInfluence}

The final step to be made is analogous to the transition from \textit{LeafRefit} to \textit{FastLeafRefit}: \textit{LeafInfluence} is, again, $O(Tn^2)$ because it has to compute matrix/vector products with the matrix $J(A^{t-1})_{ij}$ for every $t$. The same approximation that powers \textit{FastLeafRefit} can be made here as well: at each step, we can select an update set $U^t$ and only take into account the influences of a subset of training objects on $A^{t-1}$. This is equivalent to assuming $J(A^{t-1})_{ij} = 0 \forall j \notin U^t$, making $J(A^{t-1})_{ij}$ a sparse matrix with the number of nonzero elements in each row bounded by $C = \max_i |U^t|$. Strategies of selecting $U^t$ and the resulting asymptotics become the same as described in Section 3.1.3 with the additional benefit of being able to compute the derivatives “off-line.”

4. Experiments

4.1. Research Questions

The experiments that we conduct can be broadly categorized as serving two purposes: (1) studying the fundamentals of our framework and (2) evaluating its quality in two applied problem setups. For the first part, the research questions that we seek to answer are as follows:

\textbf{RQ1.} How well do the different methods introduced in Sections 3.1 and 3.2 approximate their respective influence proxies?

\textbf{RQ2.} Do smaller update sets significantly reduce the runtimes of \textit{FastLeafRefit} and \textit{FastLeafInfluence}? Does \textit{FastLeafInfluence} yield a notable runtime speedup over \textit{FastLeafRefit}?

For the second part, we proceed by considering two applied scenarios: (1) classification in the presence of label noise, and (2) classification with train/test domain mismatch. Specifically, the research questions for this part are:

\textbf{RQ3.} For Scenario 1\footnote{We use the “Plain” mode which disables CatBoost’s conceptual modifications to the standard GBDT scheme.}, do our methods allow to detect noise in general and, more specifically, to identify training objects most harmful for specific test points?

\textbf{RQ4.} For Scenario 1\footnote{We use the “Plain” mode which disables CatBoost’s conceptual modifications to the standard GBDT scheme.}, how do the proxies and their respective approximations compare in terms of quality?

\textbf{RQ5.} For Scenario 2\footnote{We use the “Plain” mode which disables CatBoost’s conceptual modifications to the standard GBDT scheme.}, are our methods capable of detecting domain mismatch and, moreover, providing recommendations on how to fix it?

4.2. Datasets and Framework

For our experiments with GBDT, we use CatBoost\textsuperscript{cat} 2018 an open-source implementation of GBDT by Yandex\textsuperscript{yandex}. The datasets used for evaluation are: (1) Adult Data Set (\textit{Adult}, \textit{dat}, 1996)), (2) Amazon Employee Access Challenge dataset (\textit{Amazon}, \textit{dat}, 2013)), (3) the KDD Cup 2009 Upselling dataset (\textit{Upselling}, \textit{dat}, 2009) and, for the domain mismatch experiment, (4) the Hospital Readmission dataset (\textit{Strack et al.}, 2014). Dataset statistics and corresponding CatBoost parameters can be found in the supplementary material. Since we approach the problem as a search (for influential examples) problem, the main metrics we will be using are ranking metrics—specifically, DCG and NDCG\textsuperscript{dcg} 2018 with linear gains.

4.3. Proxy Approximation Quality

Here, we evaluate how well do variations of \textit{FastLeafRefit} and \textit{FastLeafInfluence} match their respective Proxies\textsuperscript{proxies} and\textsuperscript{proxies}. For that, we use the Adult Data Set. For \textit{LeafRefit}, its validity heavily depends on whether Assumption\textsuperscript{assumption} holds; thus, we split the training points into two disjoint sets based on whether they violate Assumption\textsuperscript{assumption} (\textit{Changed} in Table 2) or not. We then randomly sample $n = 2000$ points from both groups to ensure that they are equal in size and, in both of them, for each test object, we rank the training points with respect to their influence on this test object. We then measure NDCG@100 with respect to the relevance labels produced by ground-truth rankings induced by the respective proxies for \textit{LeafRefit} and \textit{FastLeafRefit}, Proxy\textsuperscript{proxy} and Proxy\textsuperscript{proxy}. Finally, we average the results over the test objects. Results are given in Table 2.
Table 2: NDCG@100 of proxy ranking approximation.

<table>
<thead>
<tr>
<th>Method</th>
<th>FastLeafRefit</th>
<th>FastLeafInfluence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Same</td>
<td>Changed</td>
</tr>
<tr>
<td>SinglePoint</td>
<td>0.38</td>
<td>0.10</td>
</tr>
<tr>
<td>Top1Leaves</td>
<td>0.41</td>
<td>0.10</td>
</tr>
<tr>
<td>Top2Leaves</td>
<td>0.53</td>
<td>0.10</td>
</tr>
<tr>
<td>Top8Leaves</td>
<td>0.87</td>
<td>0.10</td>
</tr>
<tr>
<td>Top22Leaves</td>
<td>0.96</td>
<td>0.10</td>
</tr>
<tr>
<td>Top64Leaves</td>
<td>1.00</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Figure 1: Wall times elapsed per training object for different variations of FastLeafRefit and FastLeafInfluence. Top k denotes the TopKLeaves(k) update set.

analogous to that of FastLeafRefit. This shows that LeafInfluence is more robust in approximating its corresponding proxy than LeafRefit.

4.4. Runtime Comparison

In this section, we compare different variations (update set choices) of FastLeafRefit and FastLeafInfluence in terms of their runtimes. For each dataset used in the study, we randomly pick $k = 100$ training objects for influence evaluation, calculate the resulting change in the model (new leaf values for FastLeafRefit and leaf value derivatives for FastLeafInfluence), measure the total elapsed wall time and divide the result by $k$ to obtain the average time elapsed per one training object. The results are given in Fig. 1.

Firstly, as expected, we observe that smaller update sets considerably reduce the runtimes of our algorithms, with the most radical speedup yielded by SinglePoint due to not having to recalculate any derivatives at all. Secondly, quite naturally, FastLeafInfluence performs much faster than FastLeafRefit, presumably due to vectorization and gradient precomputation (see end of Section 3.2.1). These observations confirm RQ2.

4.5. Harmful Object Removal

In this experiment, we consider a particular use-case scenario, classification in the presence of label noise, and evaluate whether our methods are able to identify training objects that are (1) noisy, (2) harmful for specific test objects. In order to do that, we randomly select $k$ training samples, flip their labels, and obtain GBDT’s predictions on test data before and after noise injection. We then conduct two experiments:

A. We sort the training points in ascending order of average influence on test objects and measure ROC-AUC of noise detection. In addition to variations of FastLeafRefit and FastLeafInfluence, we also compare against (1) A noise detection method exploiting the problem structure, which scores the training points using GBDT’s prediction in favor of the class opposite to its observed label (Detector), (2) actual loss changes after leave-one-out retraining (Leave-One-Out), and (3) ground-truth binary labels of the train object being noisy (Oracle). The results are given in Fig. 2.

B. We select $n = 50$ test points that suffered the largest Logloss increase, thus simulating problematic test objects. For each of these objects, we sort the training points in ascending order of influence and incrementally remove them from the training set in batches of $m = 50$ objects; on each iteration we measure the relative Logloss reduction both on this given test object and on the whole test set $X_{test}$ and, similarly to ranking, calculate DCG using these reductions as gains. Finally, we average these metrics over the $n$ test points. The results are given in Fig. 3a and 3b.

Firstly, from Fig. 2 we note that all variations of FastLeafRefit and FastLeafInfluence perform strongly on the over-
Finding Influential Training Samples for Gradient Boosted Decision Trees

4.6. Debugging Domain Mismatch

A common issue in the supervised machine learning is domain mismatch. This is a situation, when the joint distribution of points in the test dataset \( X_{\text{test}} \) differs from the one in the labeled training dataset \( X_{\text{train}} \). Often in such scenarios, a model fine-tuned on the training dataset fails to produce accurate predictions on the test data. A standard way to cope with this problem is re-weighting \( X_{\text{train}} \).

In the following experiment we demonstrate how our methods allow to detect domain mismatch and get a hint on how the distribution of points in \( X_{\text{train}} \) should be modified in order to better match the distribution of points in \( X_{\text{test}} \). The design of this experiment is a modification of the corresponding use-case of Koh & Liang (2017). We use the same Hospital dataset (see Section 4.2), with each point being a hospital patient represented by 127 features and the goal is to predict the readmission. To introduce domain mismatch we bias the distribution in the training dataset by filtering out a subsample of patients with age \( \in [40; 50) \) and label \( y = 1 \). Originally we had 169/1853 readmitted patients in this group and 2140/20000 overall; after we get 17/1601 in the age \( \in [40; 50) \) group and 1988/19848 overall. Clearly, the distribution of labels in this specific age group becomes highly biased, while the proportion of positive labels in the whole dataset changes slightly (from 10.7% to 10.0%).

Training set \( X_{\text{train}} \) is naturally split into four parts \( \{ X_{\text{train}}^i \}_{i=1}^4 \) depending on the value of \( y \) and whether age \( \in [40; 50) \). One would expect that in the modified training dataset, samples with age \( \in [40; 50) \) and \( y = 1 \) are the most (positively) influential, so their removal will be the most harmful for the performance on the test dataset, while the removal of the samples with age \( \in [40; 50) \) and \( y = 0 \) might even be beneficial, since it is the most straightforward way to align the distributions in the test and train datasets. Below we confirm this expectation.

Let us focus on the subset \( X_{\text{test}}^0 := \{ x \in X_{\text{test}} \mid \text{age}(x) \in [40; 50) \} \), since its elements are expected to be the most affected by the introduced domain mismatch. We sample 100 points from every part \( \{ X_{\text{train}}^i \}_{i=1}^4 \) (or take the whole part, if it has < 100 points). For each of the methods \( \text{FastLeafRefit} \) and \( \text{FastLeafInfluence} \) with various update sets we compute the influence of the training samples on \( X_{\text{test}}^0 \). Specifically, (a) with \( \text{FastLeafRefit} \), for an element \( x \in X_{\text{train}} \) we find the average Logloss reduction on \( X_{\text{test}}^0 \), introduced by removing \( x \); (b) with \( \text{FastLeafInfluence} \), for an element \( x \in X_{\text{train}} \) we find the derivative of the average Logloss on \( X_{\text{test}}^0 \) with respect to the weight \( w \) of \( x \) at \( w = 1 \).

Table 3 provides the average influence among the sampled train points with the fixed label \( y \in \{0, 1\} \) and the fixed indicator \( I(\text{age} \in [40; 50)) \in \{0, 1\} \). As expected, with all methods, the train samples of the same type are consistently the most influential, which corresponds to maximal loss increase upon addition. In all cases removal of elements with \( y = 0 \) and age \( \in [40; 50) \) is estimated to be profitable, also confirming the initial expectations. These results allow us to answer RQ5 in the positive.

5. Conclusion

In this work, we addressed the problem of finding train objects that exerted the largest influence on the GBDT’s prediction on a particular test object. Building on the Influence Function framework for parametric models, we derived \( \text{LeafRefit} \) and \( \text{LeafInfluence} \), methods for estimating influences based on their respective proxy metrics, Proxies [1] and [2]. By utilizing the structure of tree ensembles, we also derived computationally efficient approximations to these methods, \( \text{FastLeafRefit} \) and \( \text{FastLeafInfluence} \). In our experiments, through considering several applied scenarios, we showed the practical applicability of these approaches, as well as their ability to produce actionable insights allowing to improve the existing model.

<table>
<thead>
<tr>
<th>Method</th>
<th>age ( \in [40; 50) )</th>
<th>age ( \not\in [40; 50) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR SinglePoint</td>
<td>-0.525</td>
<td>0.151</td>
</tr>
<tr>
<td>LR Top1Leaves</td>
<td>-0.515</td>
<td>0.150</td>
</tr>
<tr>
<td>LR Top2Leaves</td>
<td>-0.489</td>
<td>0.150</td>
</tr>
<tr>
<td>LR Top8Leaves</td>
<td>-0.397</td>
<td>0.147</td>
</tr>
<tr>
<td>LR Top22Leaves</td>
<td>-0.385</td>
<td>0.146</td>
</tr>
<tr>
<td>LR Top64Leaves</td>
<td>-0.384</td>
<td>0.146</td>
</tr>
<tr>
<td>LI SinglePoint</td>
<td>-0.652</td>
<td>0.015</td>
</tr>
<tr>
<td>LI Top1Leaves</td>
<td>-0.642</td>
<td>0.014</td>
</tr>
<tr>
<td>LI Top2Leaves</td>
<td>-0.616</td>
<td>-0.033</td>
</tr>
<tr>
<td>LI Top8Leaves</td>
<td>-0.524</td>
<td>-0.015</td>
</tr>
<tr>
<td>LI Top22Leaves</td>
<td>-0.512</td>
<td>-0.012</td>
</tr>
<tr>
<td>LI Top64Leaves</td>
<td>-0.511</td>
<td>-0.011</td>
</tr>
</tbody>
</table>
Acknowledgments
We would like to thank Anna Veronika Dorogush for valuable commentary and discussions, as well as technical assistance with the CatBoost library.

References


