10.11 Paper Presentation

Federated Learning

Yibin Xiong

October 11, 2021

Table of Contents

1 SecureBoost: A Lossless Federated Learning Framework

2 Missing Value Handling in Tree-based Models

Federated Learning

Motivations:

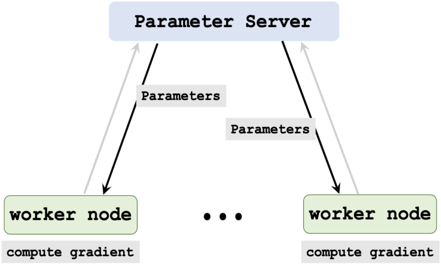
- Protect users' data privacy: keep users' data on their *local* devices rather than collected by servers of Tech companies
- Collaborations among institutions: it is difficult to share data among different institutions but we need a *complete* dataset to train an accurate model

Federated Learning

Motivations:

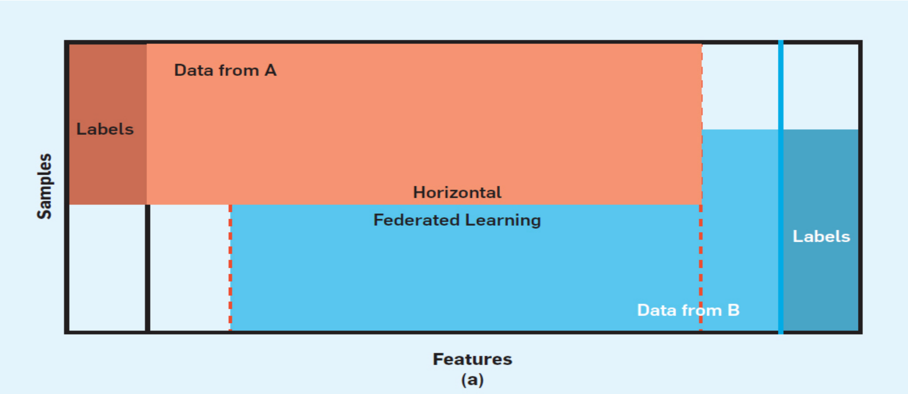
- Protect users' data privacy: keep users' data on their *local* devices rather than collected by servers of Tech companies
- Collaborations among institutions: it is difficult to share data among different institutions but we need a *complete* dataset to train an accurate model

Scheme (of distributed learning):



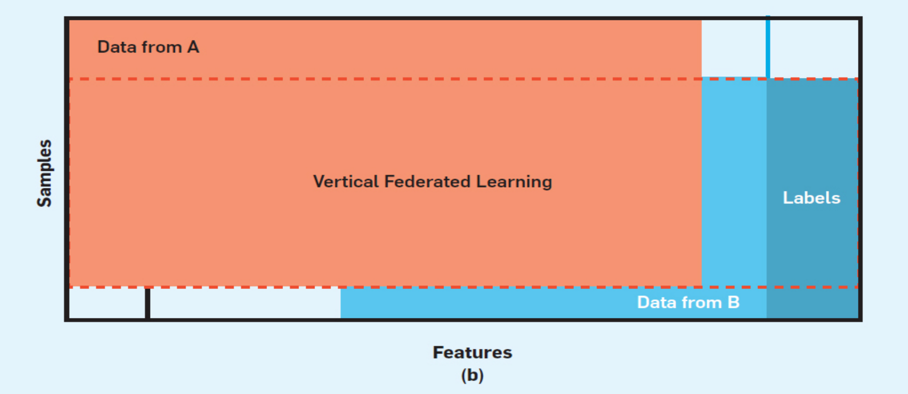
Categories: Horizontal & Vertial Federated Learning

- Horizontal/Sample Partitioned FL: The data at all parties have a large overlap in their feature sets, but they are collected from different groups of users.



Categories: Horizontal & Vertial Federated Learning

- Vertical/Feature Partitioned FL: The data at all parties have a large overlap of the users from which they are sampled, but they have different feature sets.



SecureBoost [2]: Federated Version of XGBoost

Review of XGBoost [1]:

◦ Loss function

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[l\left(y_i, \hat{y_i}^{(t-1)}\right) + g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\mathbf{x}_i\right)\right] + \Omega(f_t) \tag{2}$$

where $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \left\|w\right\|^2$, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$.

◦ Based on this loss, we can find the optimal *weight* for each leaf and the score for evaluating a split

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

SecureBoost: Federated Version of XGBoost

$$\mathcal{L}_{sp} = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \quad (3)$$

Note:

- The evaluation of w_j^* and $\mathcal{L}_{sp}$ depends only on the gradient statistics g_i and h_i, so it is easy to adapt to federated learning scenario.
- g_i and h_i can reveal the information about the true label, so they cannot be directly passed to other parties.
  E.g. When using l2 loss, $g_i = \hat{y}_i^{(t-1)} - y_i$

SecureBoost: Federated Version of XGBoost

◦ Additive Homomorphic Encryption:
$< \cdot >: \mathbb{R} \to \mathbb{R}$ such that $< u + v > = < u > \cdot < v >$

◦ Communication: Every time the passive parties passes **encrypted** versions of the gradient statistics. It is a sum over gradients for samples whose feature values fall into 2 percentile thresholds.

Algorithm 1 Aggregate Encrypted Gradient Statistics

Input: I, instance space of current node
Input: d, feature dimension
Input: $\{\langle g_i \rangle, \langle h_i \rangle\}_{i \in I}$
Output: $\mathbf{G} \in \mathbb{R}^{d \times l}$, $\mathbf{H} \in \mathbb{R}^{d \times l}$

for $k = 0 \to d$ **do**
  Propose $S_k = \{s_{k1}, s_{k2}, ..., s_{kl}\}$ by percentiles on feature k
end for
for $k = 0 \to d$ **do**
  $\mathbf{G}_{kv} = \sum_{i \in \{i | s_{k,v} \geq x_{i,k} > s_{k,v-1}\}} \langle g_i \rangle$
  $\mathbf{H}_{kv} = \sum_{i \in \{i | s_{k,v} \geq x_{i,k} > s_{k,v-1}\}} \langle h_i \rangle$
end for

Why this is still sum rather than product?

SecureBoost: Federated Version of XGBoost

Algorithm 2 Split Finding

Input: I, instance space of current node
Input: $\{\mathbf{G}^i, \mathbf{H}^i\}_{i=1}^m$, aggregated encrypted gradient statistics from m parties
Output: Partition current instance space according to the selected attribute's value

/*Conduct on Active Party*/
$g \leftarrow \sum_{i \in I} g_i, h \leftarrow \sum_{i \in I} h_i$
for $i = 0$ to m **do**
  for $k = 0$ to d_i **do**
    $g_l \leftarrow 0, h_l \leftarrow 0$
    //enumerate all threshold value
    for $v = 0$ to l_k **do**
      get decrypted values $D(\mathbf{G}^i_{kv})$ and $D(\mathbf{H}^i_{kv})$
      $g_l \leftarrow g_l + D(\mathbf{G}^i_{kv}), h_l \leftarrow h_l + D(\mathbf{H}^i_{kv})$
      $g_r \leftarrow g - g_l, h_r \leftarrow h - h_l$
      $score \leftarrow \max(score, \frac{g_l^2}{h_l+\lambda} + \frac{g_r^2}{h_r+\lambda} - \frac{g^2}{h+\lambda})$
    end for
  end for
end for
Return k_{opt} and v_{opt} to the passive party i_{opt} when we obtain the max score.
/*Conduct on Passive Party i_{opt}*/
Determine the selected attribute's value according to k_{opt} and v_{opt} and partition current instance space.
Record the selected attribute's value and return [record id, I_L] back to the active party.
/*Conduct on Active Party*/
Split current node according to I_L and associate current node with [party id, record id].

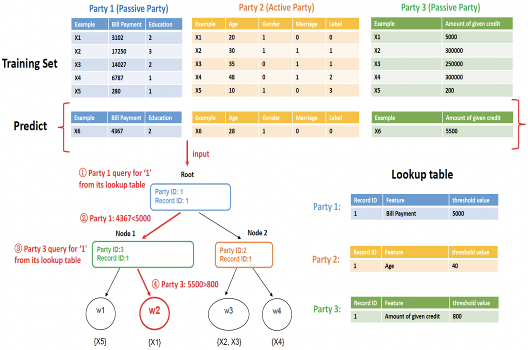


Fig. 3: An illustration of Federated Inference

Theoretical Analysis

Lossless means our FL model can achieve as good performances as a model trained on a centralized server that has access to all the data.

As long as the encryption is *additive homomorphic*, then we can decrypt $< g_l >, < h_l >$ to get g_l, h_l, which are the same as in the centralized learning model.

Paillier cryptosystem

$< m > := \ g^m r^n \mod n^2$ for some random $r \in \{0, \ldots, n-1\}$

$$
\begin{aligned}
< m1 > \cdot < m2 > &= (g^{m1} r_1^n \mod n^2) \cdot (g^{m2} r_2^n \mod n^2) \\
&= g^{m1} r_1^n \cdot g^{m2} r_2^n \mod n^2 \\
&= g^{m1+m2} (r_1 r_2)^n \mod n^2 \\
&= < m1 + m2 >
\end{aligned}
$$

Theoretical Analysis: Some Results

- Sensitive info leakage w.r.t *passive parties*:
  Given a learned SecureBoost model, its first tree's leaf purity can be inferred from the weight of the leaves.

$$\theta_j = a - (a-1)w_j^*, \text{ where } a = \hat{y}_i^{(0)}$$

- As the purity in the first tree increases, the residual information decreased.
- Improvement: Reduced-Leakage SecureBoost

  The first split rule uses one of the features in the *active party* rather than those in the passive parties.

Table of Contents

1 SecureBoost: A Lossless Federated Learning Framework

2 Missing Value Handling in Tree-based Models

Default Direction

Predictive Value Imputation

Probabilistic Imputation [3]

- Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
- Kewei Cheng et al. "Secureboost: A lossless federated learning framework". In: *IEEE Intelligent Systems* (2021).
- Pasha Khosravi et al. "Handling missing data in decision trees: A probabilistic approach". In: *arXiv preprint arXiv:2006.16341* (2020).