DART Booster

- XGB combines a large number of regression trees with a small learning rate
    - Trees added early are significant and trees added late are unimportant
- Vinayak and Gilad-Bachrach proposed a new method to add dropout techniques, this is dart

Features

- Drops trees to prevent over-fitting
    - Trivial trees (trivial errors) are prevented
- Differences
    - Can be slower than gbtree because random dropout rate
    - Early stop may be unstable due to randomness

How it Works

- In m-th training round, suppose k trees are selected to be dropped.
- Let $D = \sum_{i \in \mathbf{K}} F_i$ be the leaf scores of dropped trees and $F_m = \eta \tilde{F}_m$ be the leaf scores of a new tree.
- The objective function is as follows:

$$\text{Obj} = \sum_{j=1}^{n} L\left(y_j, \hat{y}_j^{m-1} - D_j + \tilde{F}_m\right) + \Omega\left(\tilde{F}_m\right).$$

- D and F_m are overshooting, so using scale factor

$$\hat{y}_j^m = \sum_{i \notin \mathbf{K}} F_i + a\left(\sum_{i \in \mathbf{K}} F_i + bF_m\right).$$

Parameters

The booster `dart` inherits `gbtree` booster, so it supports all parameters that `gbtree` does, such as `eta`, `gamma`, `max_depth` etc.

Additional parameters are noted below:

- `sample_type`: type of sampling algorithm.
  - `uniform`: (default) dropped trees are selected uniformly.
  - `weighted`: dropped trees are selected in proportion to weight.
- `normalize_type`: type of normalization algorithm.
  - `tree`: (default) New trees have the same weight of each of dropped trees.

$$
\begin{aligned}
a\left(\sum_{i\in\mathbf{K}} F_i + \frac{1}{k}F_m\right) &= a\left(\sum_{i\in\mathbf{K}} F_i + \frac{\eta}{k}\tilde{F}_m\right) \\
&\sim a\left(1+\frac{\eta}{k}\right)D \\
&= a\frac{k+\eta}{k}D = D, \\
a &= \frac{k}{k+\eta}
\end{aligned}
$$

  - `forest`: New trees have the same weight of sum of dropped trees (forest).

$$
\begin{aligned}
a\left(\sum_{i\in\mathbf{K}} F_i + F_m\right) &= a\left(\sum_{i\in\mathbf{K}} F_i + \eta\tilde{F}_m\right) \\
&\sim a\left(1+\eta\right)D \\
&= a(1+\eta)D = D, \\
a &= \frac{1}{1+\eta}.
\end{aligned}
$$

- `rate_drop`: dropout rate.
  - range: [0.0, 1.0]
- `skip_drop`: probability of skipping dropout.
  - If a dropout is skipped, new trees are added in the same manner as gbtree.
  - range: [0.0, 1.0]