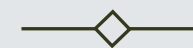




# LIGHTGBM とは



2023/22 岡田 隆之

LightGBM | 用語解説 | 野村総合  
研究所(NRI)をまとめた

# 勾配ブースティング

「**ブースティング**」・・・与えられたデータから**決定木分析**を行った後に、予測が正しくできなかった**データに重みをつけて**、再度、決定木分析を行い、これを繰り返すことで精度を高める方法。

上のブースティングからさらに、予測値と実績値の**誤差**を計算して**誤差を決定木に反映**させる方法が「勾配ブースティング」。ブースティングと同様に、誤差に対する学習を繰り返すことで精度を高めていく。

LightGBMも勾配ブースティングを用いたアルゴリズムであり、そのほかとしては、XGBoost、Catboostなどがある。

\* このスライドでは、**赤**：とても重要なところ **青**：注意しておきたいところ とする。

# LightGBMの特徴

**LightGBM** (Light Gradient Boosting Machine) は、その名の通り、決定木の「Gradient Boosting (勾配ブースティング)」を用いた手法で、独自のアルゴリズムによって「Light (軽い、高速)」なことが特徴。

ほかの一般的な勾配ブースティングの場合は、**誤差**を最小化するように“分割”の要素、基準を見つけるため、データ量に応じて計算量が増えてしまう (bad)。

1つ1つの決定木の精度をなるべく落とさずに、高速に構築できるようにしたことがLightGBMの最大の特徴だという (good)。工夫としては以下の**4**点がある。

- ① Leaf-wise tree growth    ② Histogram based
- ③ Gradient-based One-Side Sampling (GOSS)
- ④ Exclusive Feature Bundling (EFB)    ⇒次ページから説明。

## ① Leaf-wise tree growth

一般的な決定木の場合は、決定木の階層ごとに計算（Level wise）するため、1つの階層の分岐がすべて終わってから次の階層を計算するが、分岐がなくなった要素（＝葉、leaf）については、それ以上は計算しない。

## ② Histogram based

決定木の分岐をする際に、すべての値をみるのではなく、ヒストグラムをつかって、数値をまとめて分岐させる。

### ③ Gradient-based One-Side Sampling (GOSS)

学習できていない要素を学ぶことを優先するため、誤差が小さいデータは減らし、誤差の大きいデータだけを残すことで学習データの量を減らす。

## ④ Exclusive Feature Bundling (EFB)

異なる特徴量の中でも、まとめても問題がなさそうな特徴量を1つにすることで計算量を減らす。

## 注意事項

ハイパーパラメータと言われる変数を設定する必要がある。決定木の「葉の数（`num_leaves`）」や「1つの葉に含まれる最小データ数（`min_data_in_leaf`）」、「階層の深さ（`max_depth`）」などを、モデルの精度と過学習のバランスを考えながらチューニングすることが求められる。



## 終わりに

データサイエンスをやるならだれもが知っておきたいLightGBMについてのテンプレート資料を作成したので、そちらも説明する。

今回の資料作りはほとんどNRIの文章を使ったもので、大変勉強になった。参考リンクを載せておく：

[https://www.nri.com/jp/knowledge/glossary/1st/alphabet/light\\_gbm](https://www.nri.com/jp/knowledge/glossary/1st/alphabet/light_gbm)