LIGHTGBM とは

2023/4/3 岡田 隆之

LightGBM | 用語解説 | 野村総合研究所(NRI)を中心にまとめた

(事前知識)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**ことで効率的に精度を高めていく。

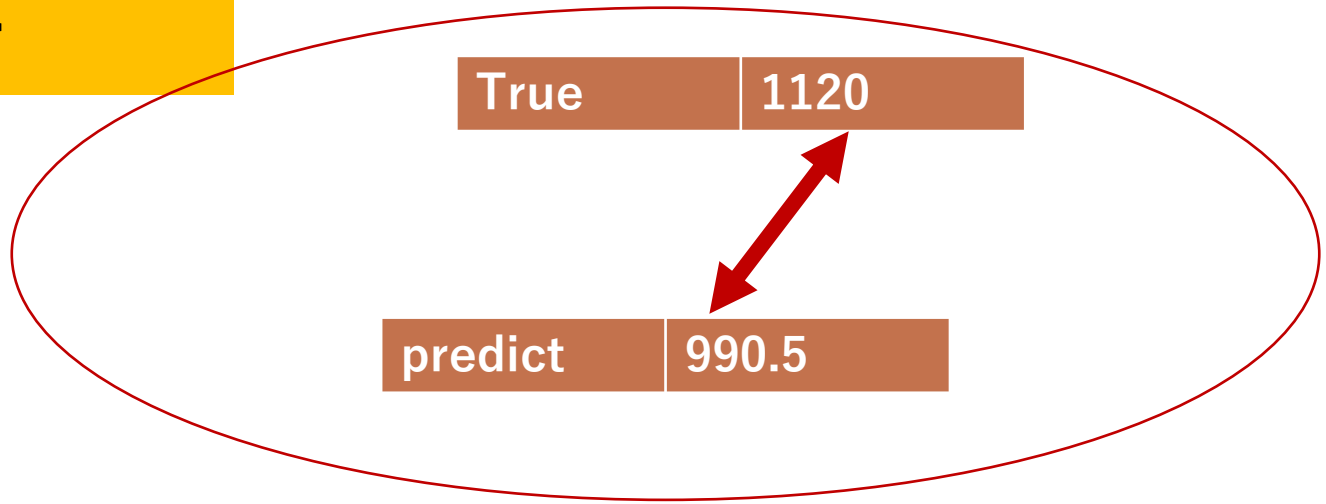
LightGBMも勾配ブースティングを用いたアルゴリズムであり、そのほかとしては、XGBoost、Catboostなどがある。

＊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

緑：応用事項　とする。

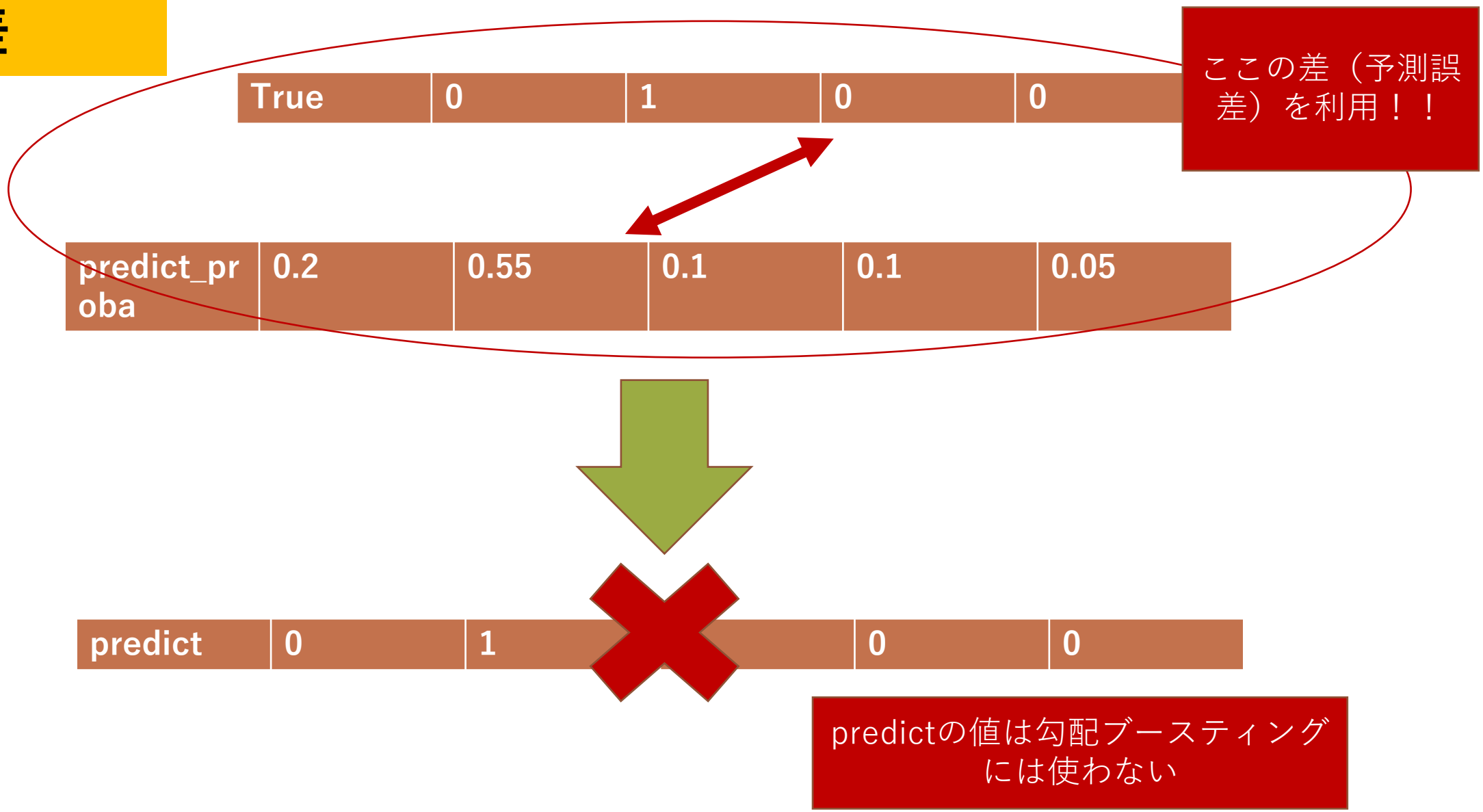
ここで、

回帰の場合の誤差



単純にpredictとの差（予測誤差）を利用！！
簡単。
大きい誤差が出たデータから優先的に再学習！！

分類の場合の誤差


True
0
1
0
0
ここの差（予測誤差）を利用！！
predict_proba
0.2
0.55
0.1
0.1
0.05
predict
0
1
0
0
predictの値は勾配ブースティングには使わない

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、総当たりで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。工夫としては以下の4点がある。

①Leaf-wise tree growth　②Histogram based

③Gradient-based One-Side Sampling (GOSS)

④Exclusive Feature Bundling (EFB)　⇒次ページから説明。

①Leaf-wise tree growth

一般的な決定木の場合は、~~決定木の階層ごとに上から計算（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、**分岐が必要なくなった要素（＝葉、leaf）**については、**それ以上は計算しない。**

leaf wiseだと何が嬉しいの？

うさぎ

leaf wiseを採用することで，ある程度精度がでる決定木をlevel wiseよりも早く構築することができます．(損失を下げるノードから優先的に分割していくので，理解しやすいかと思います．)

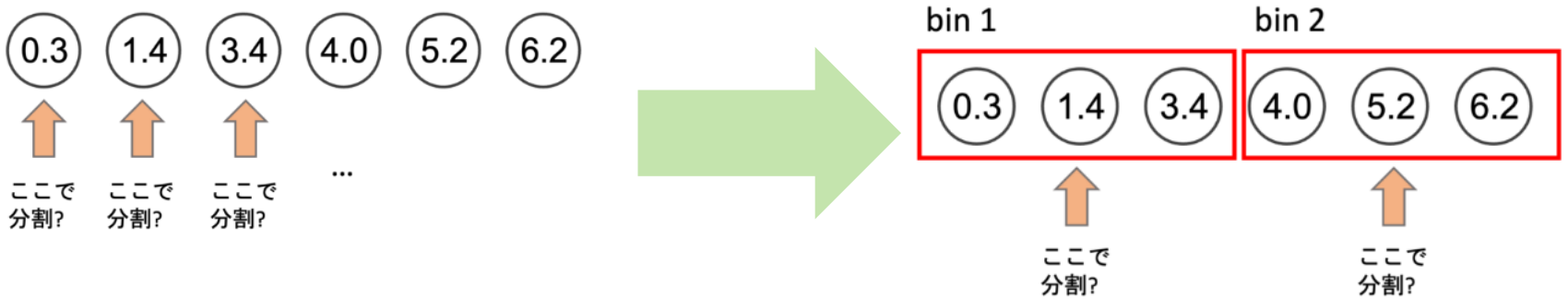
ほかの決定木系モデル（level wise）と比べて、作成/分割の順番が変わるだけ

参考：https://datawokagaku.com/lightgbm/#histogram_based

②Histogram based

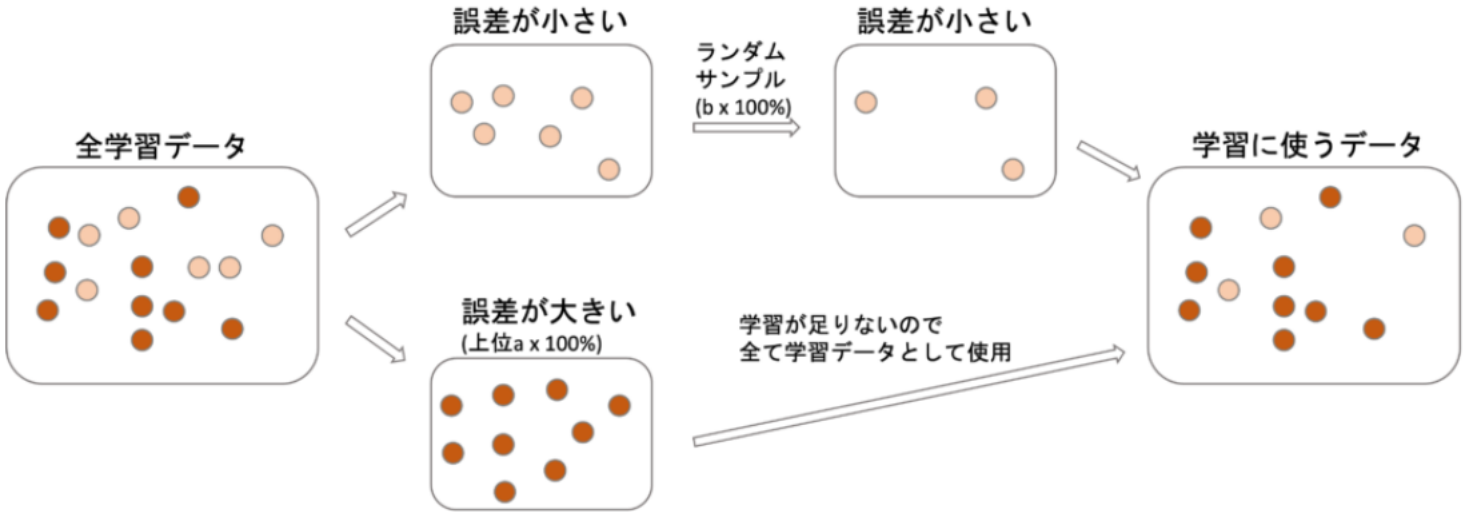
決定木の分岐をする際に、~~すべての値をみる~~のではなく、**ヒストグラム**をつくって、**数値をまとめて分岐**させる。

これを緩和するために，一つ一つの値を分岐点の候補にするのではなく，値をヒストグラム化し，いくつかの値を一つのbinにして**そのbinを分岐点の候補にする**ことによって高速化を狙います.これをhistogram based algorithmと呼びます.



③Gradient-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。



うまくいっていない部分(教科、単元)を優先的に学びなおすので、不得意部分が改善しやすい。

④Exclusive 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。

特徴量A	特徴量B
1	0
2	0
1	0
2	0
0	0
0	1
0	2
0	3
0	4

特徴量バンドル
1
2
1
2
0
3
4
5
6

アルゴリズム詳細は不明だが、同時に有効な値を取らないような特徴量をまとめて特徴量の削減を行っている。
⇒特徴量が減れば、計算も早く！！

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**データ分析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**NRIの資料と下記の2資料**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/lst/alphabet/light_gbm

https://datawokagaku.com/lightgbm/#histogram_based

https://threecourse.hatenablog.com/entry/2019/10/31/141921