

アンサンブル学習

7章

アジェンダ

- ~~アンサンブル学習の概要~~
- バギング
 - 仕組み
 - 特徴
 - 実装
- ブースティング
 - アダブーストの仕組み
 - アダブーストの特徴
 - アダブーストの実装
- 使用例

自己紹介

- 川本将人
- 26才
- 大阪府四条畷市出身
- 音楽が好き
- 借金50万

2023年2月19日

15:40

A dynamic illustration of a horse race in progress. Three horses and their jockeys are shown in mid-stride, kicking up dust. The jockey in the foreground wears a blue and white silks with a white helmet. The horse behind them has a blue saddle cloth with the number 13 and the name 'レモンボニー' (Lemon Bonny). The background is a warm, golden-brown color with a textured, dusty appearance. The text 'FEBRUARY STAKES' is prominently displayed in the center in a large, white, distressed font. Below it, the Japanese text 'フェブラリース特集' (February Special) is written in white, followed by a blue circle containing the white text 'GI'. At the bottom, the date and time '2023年2月19日 (日) 15:40' and the location '東京競馬場 ダ1600m' are listed in white. The 'netkeiba' logo is in the top right corner.

netkeiba

FEBRUARY STAKES

フェブラリース特集 GI

2023年2月19日 (日) 15:40
東京競馬場 ダ1600m

したいこと

アンサンブル学習を使って、勝つ馬を調べる



したいこと

有識者に聞いてみた



mot Today at 1:57 PM

競馬の勝つ馬予想するとき、何の情報見たらええん？



たわし Today at 1:59 PM

初心者みたいな質問やな



mot Today at 2:03 PM

なんなん

重要なやつ



たわし Today at 2:15 PM

最近の成績、上り、好走パターン(競馬場、距離、展開)



mot Today at 2:21 PM

あざー

本題

バギングとは？

バギング

バギングとは？



バギング (Bagging) は、Bootstrap Aggregatingの略称です。これは、複数の学習モデルを生成し、これらを組み合わせて、より良い予測結果を生成するためのアンサンブル学習アルゴリズムです。

バギングは、各学習モデルが独立してトレーニングデータから学習を行い、最終的な予測結果は各モデルの予測結果の平均や投票などによって決定されます。このため、バギングは、学習モデルの汎化性能を向上させることができます。

バギングは、分類や回帰など、様々なタスクに使用することができます。また、弱い学習アルゴリズムを組み合わせることができるため、高い汎化性能を実現することができます。

バギングの仕組み①

元データからランダムにデータセットをn回サンプルする

元データ

名前	性別	年齢	斤量	入賞
ディーブ	牡	4	57	1
オグリ	牡	6	56	0
ウオッカ	牝	5	55	0
オルフェ	牡	5	57	1
ナリタ	牡	6	57	1

ブートストラップ

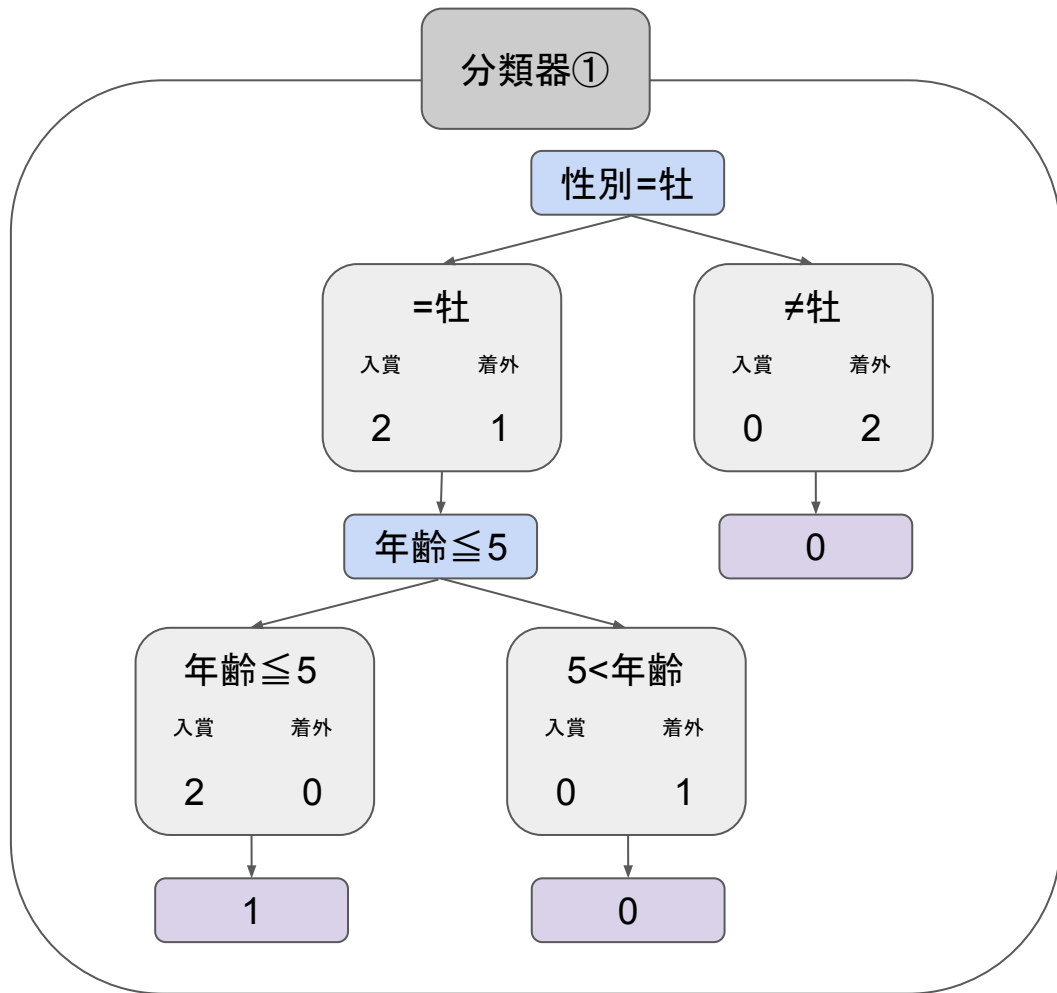
名前	性別	年齢	斤量	入賞
オルフェ	牡	5	57	1
オグリ	牡	6	56	0
ウオッカ	牝	5	55	0
ウオッカ	牝	5	55	0
ディーブ	牡	4	57	1

バギングの仕組み②

分類器(決定木)をつくる

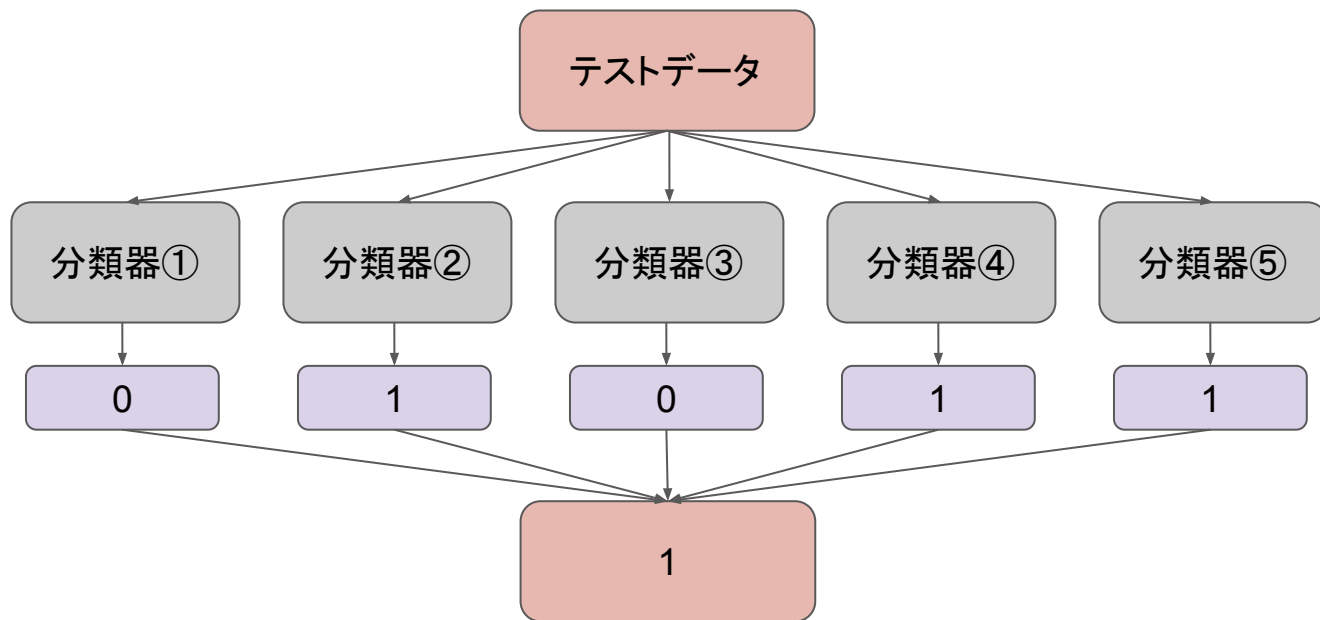
ブートストラップ

名前	性別	年齢	斤量	入賞
オルフェ	牡	5	57	1
オグリ	牡	6	56	0
ウォッカ	牝	5	55	0
ウォッカ	牝	5	55	0
ディーブ	牡	4	57	1



バギングの仕組み③

投票(or 平均)で結果を予測する



バギング

特徴

- **複数のモデル**を作成して、最終的な予測結果を平均化することで**精度向上**を図る
- 各モデルはトレーニングサンプルを**独立してランダム**に選び、**同じ特徴量**から作成する
 - これにより**過学習**を防ぐ
- 同じサンプルが**複数回**選ばれる可能性がある

メリット

- 複数のモデルを作成することで精度向上
- 過学習を防ぐ
- 簡単な実装

デメリット

- 計算量が多いので、時間がかかる
- 個々のモデルの精度は低い可能性

バギング

実装

jupiterへ

本題

ブースティング(アダブースト)とは？

ブースティング

ブースティングとは



ブースティング (Boosting) は、複数の学習モデルを生成し、これらを組み合わせて、より良い予測結果を生成するためのアンサンブル学習アルゴリズムです。

ブースティングは、各学習モデルが前回の予測結果を反映して学習を行い、最終的な予測結果は各モデルの予測結果を加重平均することで決定されます。このため、ブースティングは、分類や回帰など、様々なタスクに使用することができます。また、弱い学習アルゴリズムを強い学習モデルに改善することができるため、高い汎化性能を実現することができます。

アダブーストの仕組み①

等しく重みをつける

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5

アダブーストの仕組み②

エントロピーを計算する

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5

$$H = -\sum_{i=1} P(x_i) \log_2 P(x_i)$$

性別	
=牡	≠牡
確率	4/5 1/5

$$-\left(\frac{1}{5} \times \log_2 \frac{1}{5} + \frac{4}{5} \times \log_2 \frac{4}{5}\right) = 0.72$$

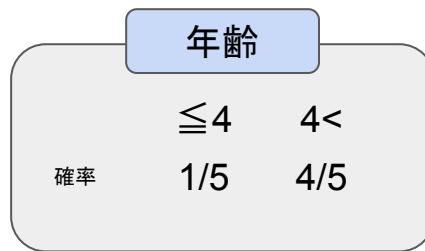
アダブーストの仕組み②

エントロピーを計算する

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5

$$H = -\sum_{i=1} P(x_i) \log_2 P(x_i)$$



$$-\left(\frac{1}{5} \times \log_2 \frac{1}{5} + \frac{4}{5} \times \log_2 \frac{4}{5}\right)$$

$$= 0.72$$



$$-\left(\frac{3}{5} \times \log_2 \frac{3}{5} + \frac{2}{5} \times \log_2 \frac{2}{5}\right)$$

$$= 0.97$$

アダブーストの仕組み②

エントロピーを計算する

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5

$$H = -\sum_{i=1} P(x_i) \log_2 P(x_i)$$



$$-\left(\frac{1}{5} \times \log_2 \frac{1}{5} + \frac{4}{5} \times \log_2 \frac{4}{5}\right)$$

$$= 0.72$$



$$-\left(\frac{2}{5} \times \log_2 \frac{2}{5} + \frac{3}{5} \times \log_2 \frac{3}{5}\right)$$

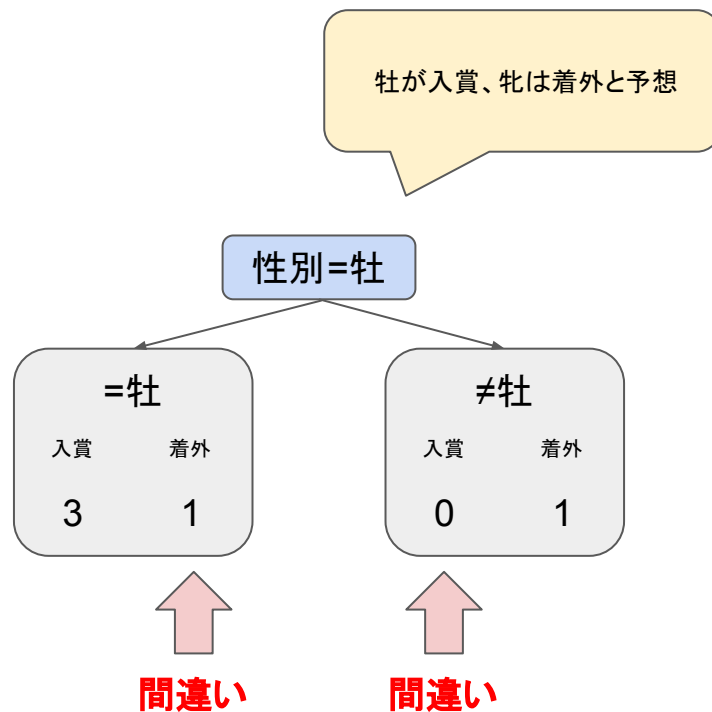
$$= 0.97$$

アダブーストの仕組み③

分類器(決定木)をつくる

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5



間違ったデータの重みの合計

アダブーストの仕組み④

影響度 (Amount of Say) を計算する

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オルフェ	牡	5	57	1	1/5
ナリタ	牡	6	57	1	1/5

$$Alpha = \frac{1}{2} \log \frac{1 - TotalError}{TotalError}$$

=牡

入賞 着外

3

1

≠牡

入賞 着外

0

1

$$Alpha = \frac{1}{2} \log \frac{1 - \frac{1}{5}}{\frac{1}{5}} = \frac{1}{2} \log \frac{\frac{4}{5}}{\frac{1}{5}} = \frac{1}{2} \log(4) = 0.69$$

アダブーストの仕組み⑤

新しい重みを計算する

元データ

名前	性別	年齢	斤量	入賞	重み	重み(仮)
ディーブ	牡	4	57	1	0.2	
オグリ	牡	6	56	0	0.2	0.4
ウオッカ	牝	5	55	0	0.2	
オルフェ	牡	5	57	1	0.2	
ナリタ	牡	6	57	1	0.2	

新しい重み(不正解) = 重み $\times e^{Alpha}$

$$\frac{1}{5} \times e^{0.69} = \frac{1}{5} \times 1.99 = 0.4$$

アダブーストの仕組み⑤

新しい重みを計算する

元データ

名前	性別	年齢	斤量	入賞	重み	重み(仮)
ディーブ	牡	4	57	1	0.2	0.1
オグリ	牡	6	56	0	0.2	0.4
ウオッカ	牝	5	55	0	0.2	0.1
オルフェ	牡	5	57	1	0.2	0.1
ナリタ	牡	6	57	1	0.2	0.1

新しい重み(正解) = 重み $\times e^{-Alpha}$

$$\frac{1}{5} \times e^{-0.69} = \frac{1}{5} \times 0.5 = 0.1$$


アダブーストの仕組み⑥

新しい重みを正規化する

元データ

名前	性別	年齢	斤量	入賞	重み	重み(仮)
ディーブ	牡	4	57	1	0.2	0.1
オグリ	牡	6	56	0	0.2	0.4
ウオッカ	牝	5	55	0	0.2	0.1
オルフェ	牡	5	57	1	0.2	0.1
ナリタ	牡	6	57	1	0.2	0.1
					計	0.8

$\div 0.8$



NEW重み
0.125
0.5
0.125
0.125
0.125
1

アダブーストの仕組み⑦

使用したデータからサンプルする

元データ

名前	性別	年齢	斤量	入賞	重み
ディーブ	牡	4	57	1	0.125
オグリ	牡	6	56	0	0.5
ウオッカ	牝	5	55	0	0.125
オルフェ	牡	5	57	1	0.125
ナリタ	牡	6	57	1	0.125

重い方が予測が難しい

ブートストラップ

名前	性別	年齢	斤量	入賞
オルフェ	牡	5	57	1
オグリ	牡	6	56	0
ウオッカ	牝	5	55	0
オグリ	牡	6	56	0
オグリ	牡	6	56	0

アダブーストの仕組み⑧～

重みを更新し、エントロピーを計算する

ブートストラップ

名前	性別	年齢	斤量	入賞	重み
オルフェ	牡	5	57	1	1/5
オグリ	牡	6	56	0	1/5
ウオッカ	牝	5	55	0	1/5
オグリ	牡	6	56	0	1/5
オグリ	牡	6	56	0	1/5

性別

=牡 ≠牡

確率

4/5 1/5

= 0.72

年齢

≤5 5<

確率

2/5 3/5

= 0.97

斤量

≤55 55<

確率

1/5 4/5

= 0.72

斤量

≤56 56<

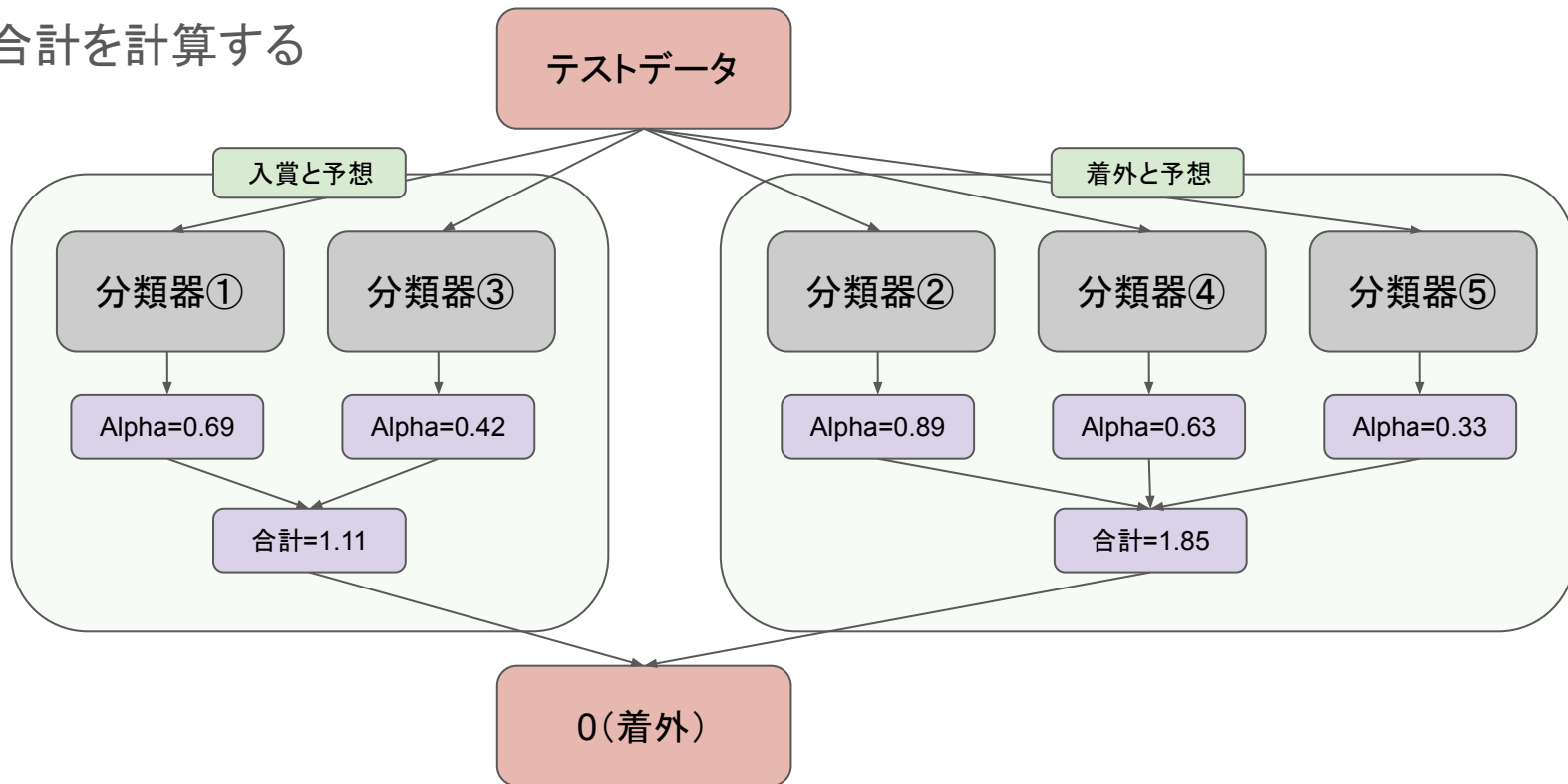
確率

4/5 1/5

= 0.72

アダブーストの仕組み(最後)

Alphaの合計を計算する



アダブースト

特徴

- **複数の弱学習器**を組み合わせて強い予測モデルを構築する手法
- 毎回のイテレーションで前回の予測とは異なる標本を重視することで、改善を図る
- 適切な**ハイパーパラメータチューニング**などによって結果をさらに改善できる可能性あり

メリット

- 弱い分類器を組み合わせることによる、高い予測性能
- 分類の難しいサンプルをより学習するため、不均衡なデータセットに強い

デメリット

- 計算量が多いので、時間がかかる
- 弱い分類器を繰り返し適用するため、過学習しやすい

ブースティング

実装

jupiterへ

レース

分析

- [Netkeiba.com](https://netkeiba.com)を使用
- 3位以内を1に、4位以下を0に変換しこれを予測する
- 問題数点

まとめ

改善点

- データ数？
- 特徴量？
- 次元削減？
- モデル？

➡続報に期待

参考

ADABOOST:https://www.youtube.com/watch?v=NLRO1-jp5F8&ab_channel=KrishNaik

SCRAPING:<https://github.com/unonao/race-predict>

DECISION

TREE:https://www.youtube.com/watch?v=qQa9Emh0pZE&ab_channel=K_DM%E3%80%90%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92xPython%E3%80%91

KEIBA:https://race.netkeiba.com/race/shutuba.html?race_id=202305010811&rf=race_list