前回の分類と同じように、回帰のテストを行った。

データ内容は、

http://archive.ics.uci.edu/ml/index.php

から、落として来た、Physicochemical Properties of Protein Tertiary Structure Data Set(タンパク質の三次構造データセットの物理化学的性質)

というものを使って見る。

特徴量は9個で、

F1 - 総表面積。

F2 - 非極性露出領域。

F3 - 露出した非極性残留物の部分面積。

F4 - 残渣の露出した非極性部分の小面積。

F5 - 分子量加重露出面積。

F6 - 残留物の標準露出面積からの平均偏差。

F7 - ユークリッド距離。

F8 - 二次構造ペナルティ。

F9 - 空間分布制約（N、K値）。

求めたいのは、

RMSD-残渣のサイズ。(Size of the residue.)(値は0~21)

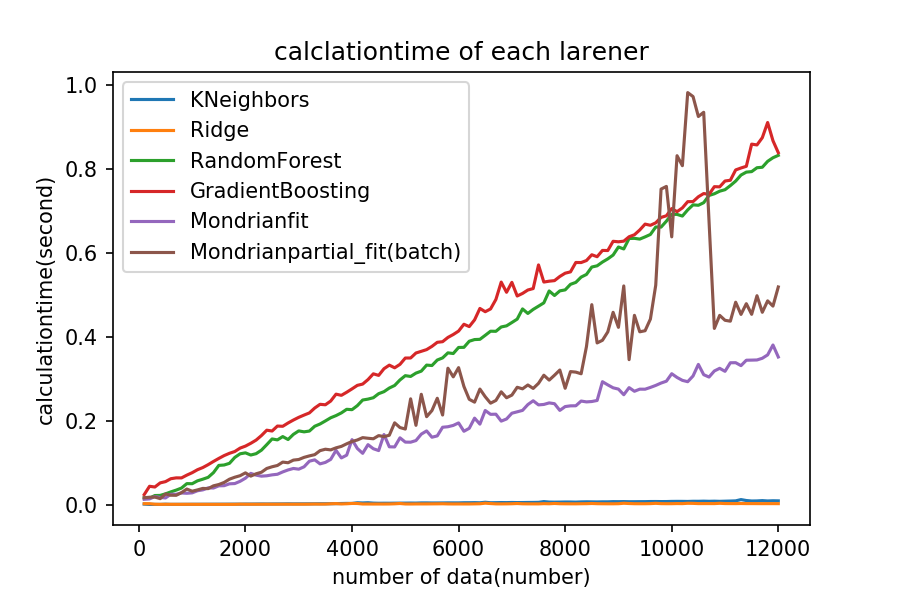
欠損値はなし、F8は整数値で他は少数値。

データは4万個くらいあったが、重くなるので15000個を使用。

これをtrain\_test\_splitで学習用データ12000個とテストデータ3000個に分けてテストを行う。

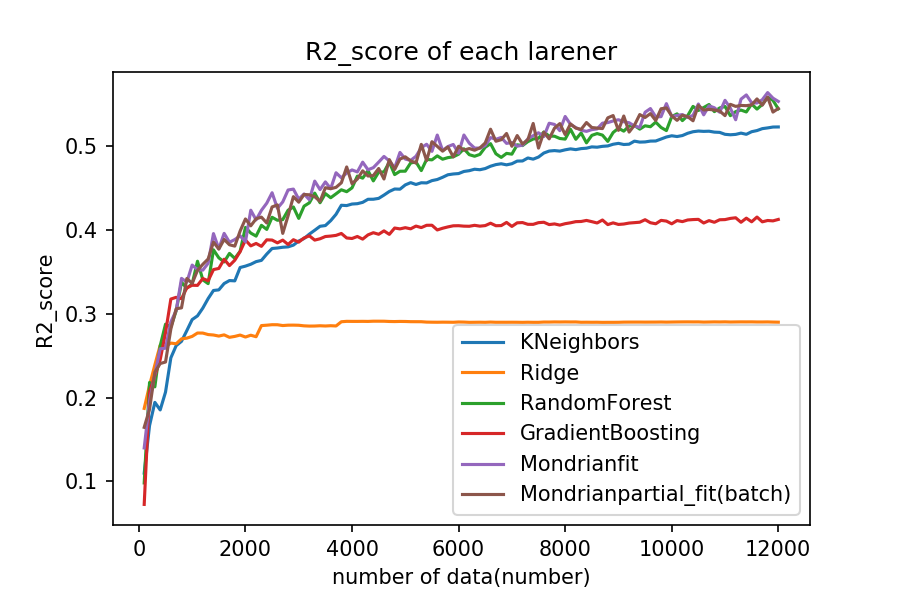
まず、前回と同じようにそもそもの性能をみる。

Mondrianpartial\_fitは一回一回リセットし、データ数ゼロからのスピード(batch)



Mondrianのpartial\_fitは前もそうだったが、計算時間に乱数性がある。

また、この時のR2スコア。

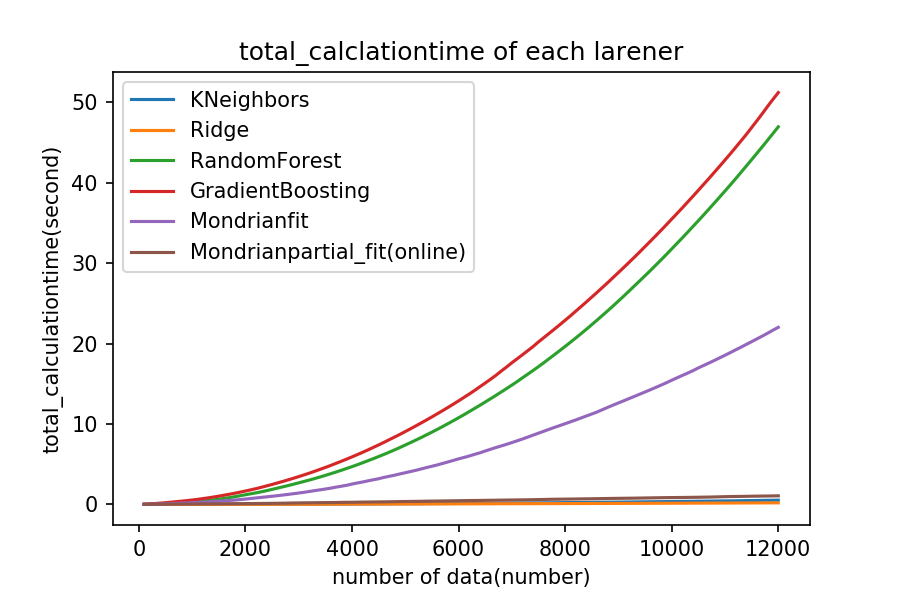


これに関してはかなり良いスコアで、RandomForestを若干上回っている。

おそらく、GradientBoostingは過学習している。

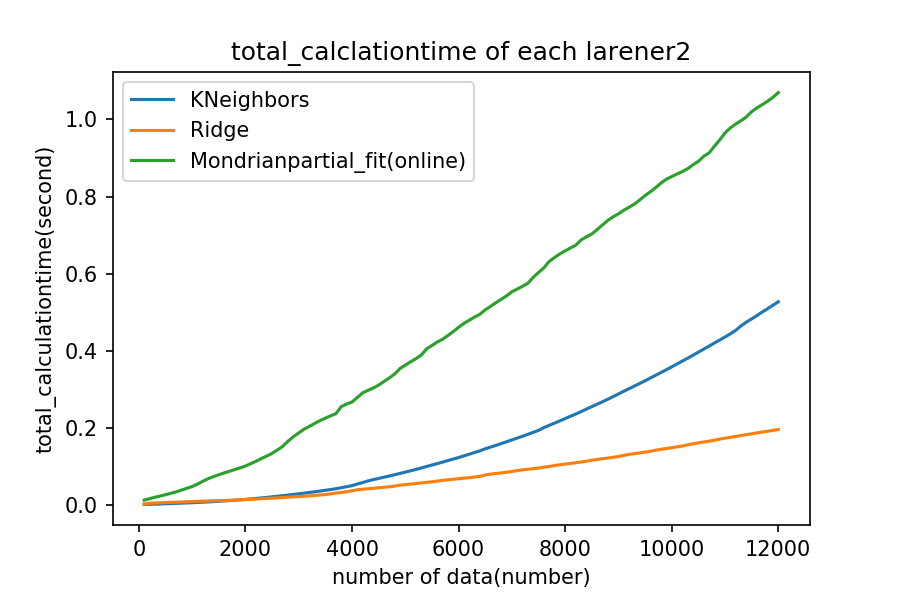
また、データが100個づつ来たと仮定した場合の計算時間。

この時、Mondrianpartial\_fit以外は100個来るたびに全て再学習する



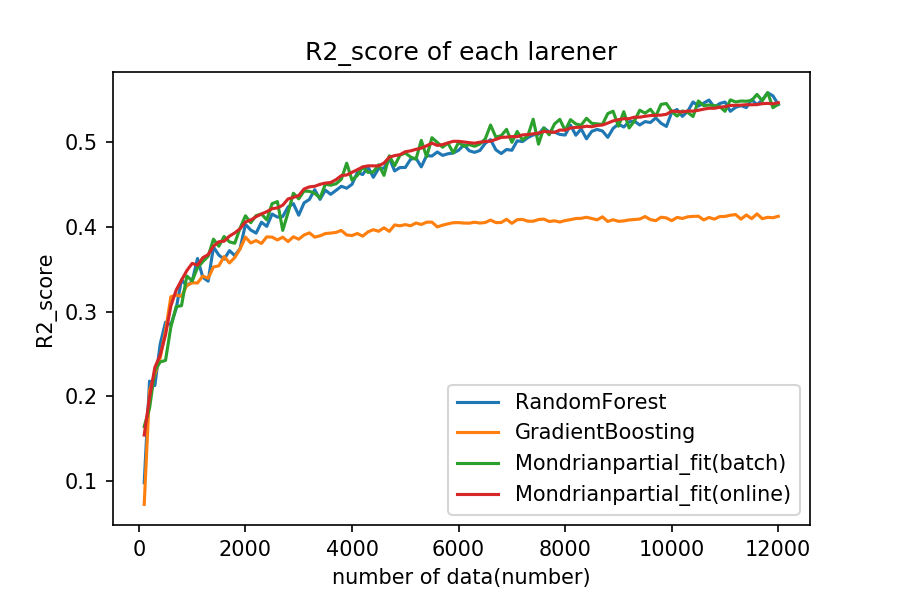
もちろんだが、この状況では最高にMondrianが強い。

全く見えないのででかいのは省く。



リッジ回帰とk近傍回帰がすごい早かったがMondrianはそれに匹敵する。

Onlineでのスコアはどうか



Onlineでもいいbatchと同じ値のスコアが出ている。

きちんとOnline学習できている。