Mondrian_test4_Regression

前回の分類と同じように、回帰のテストを行った。

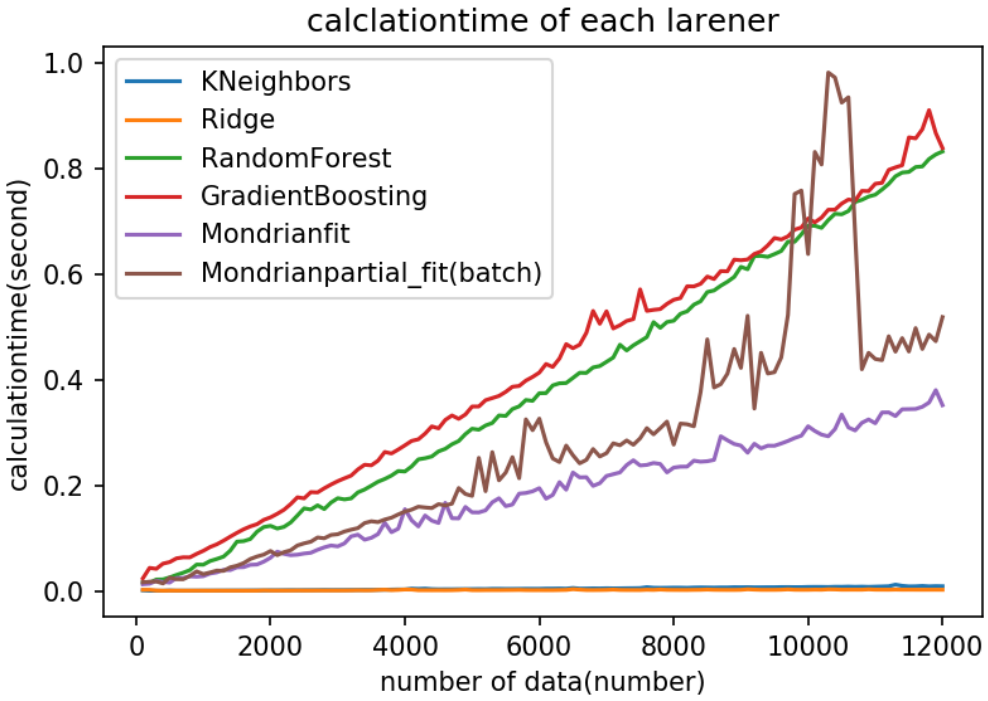
データ内容は、
http://archive.ics.uci.edu/ml/index.php
から、落として来た、Physicochemical Properties of Protein Tertiary Structure Data Set(タンパク質の三次構造データセットの物理化学的性質)
というものを使って見る。
特徴量は 9 個で、
F1 - 総表面積。
F2 - 非極性露出領域。
F3 - 露出した非極性残留物の部分面積。
F4 - 残渣の露出した非極性部分の小面積。
F5 - 分子量加重露出面積。
F6 - 残留物の標準露出面積からの平均偏差。
F7 - ユークリッド距離。
F8 - 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。

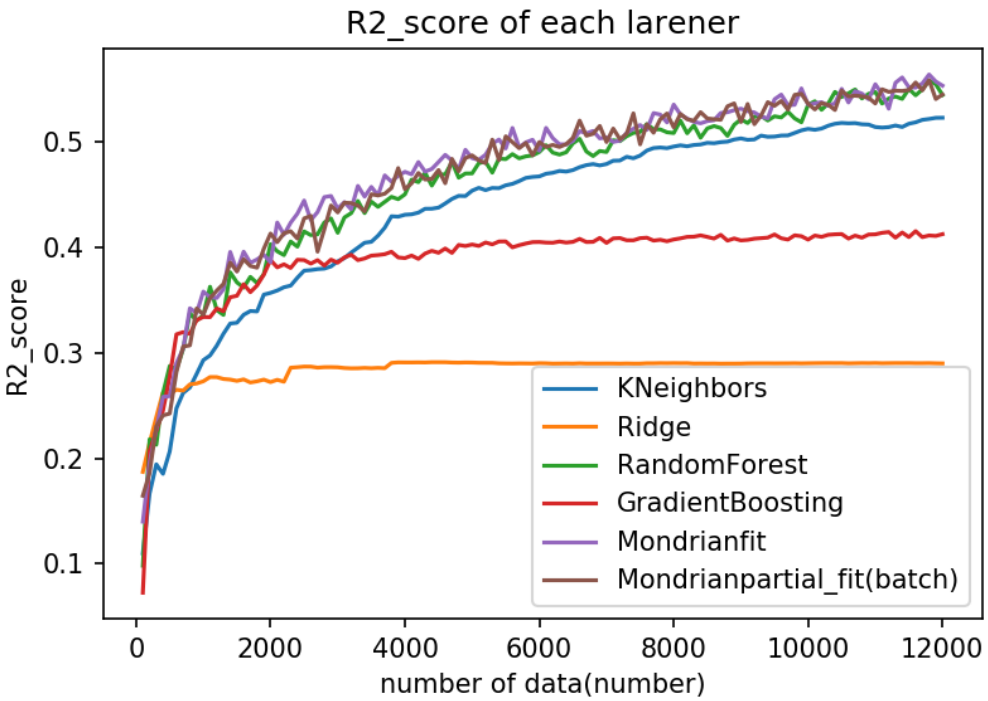
まず、前回と同じようにそもそもの性能をみる。
Mondrianpartial_fit は一回一回リセットし、データ数ゼロからのスピード(batch)

Mondrian_test4_Regression



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。

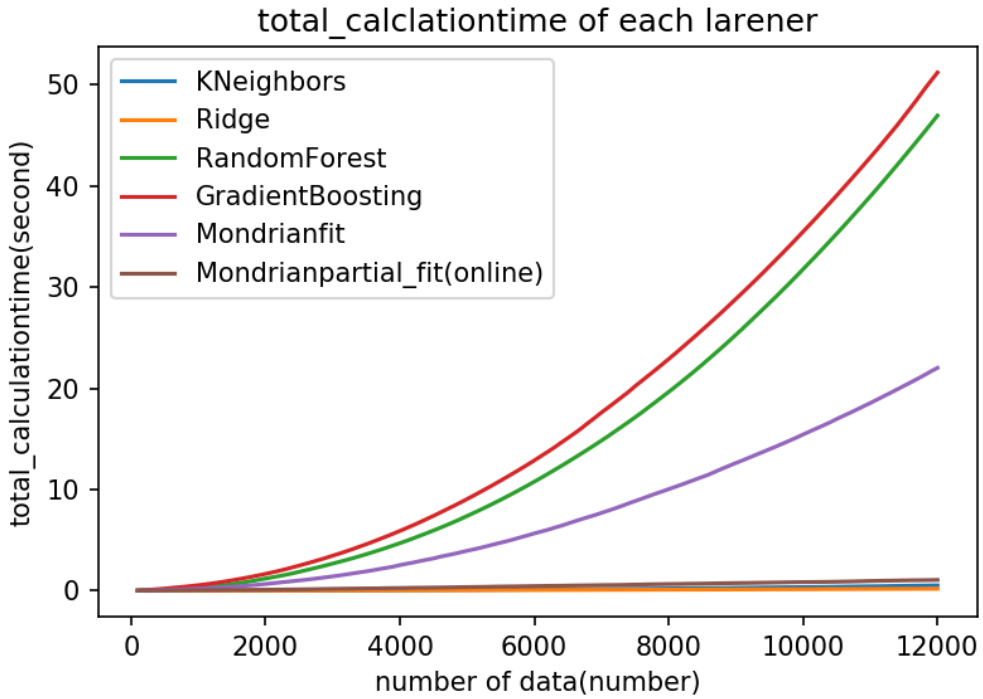


これに関してはかなり良いスコアで、RandomForest を若干上回っている。

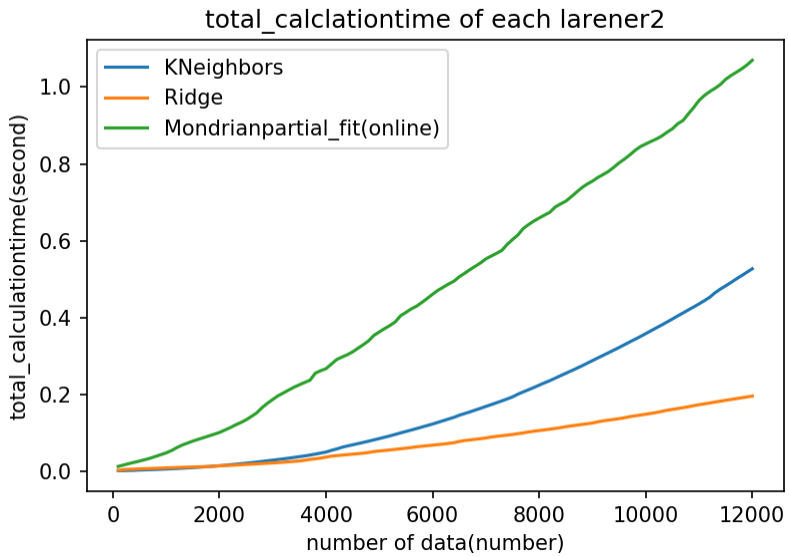
Mondrian_test4_Regression

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



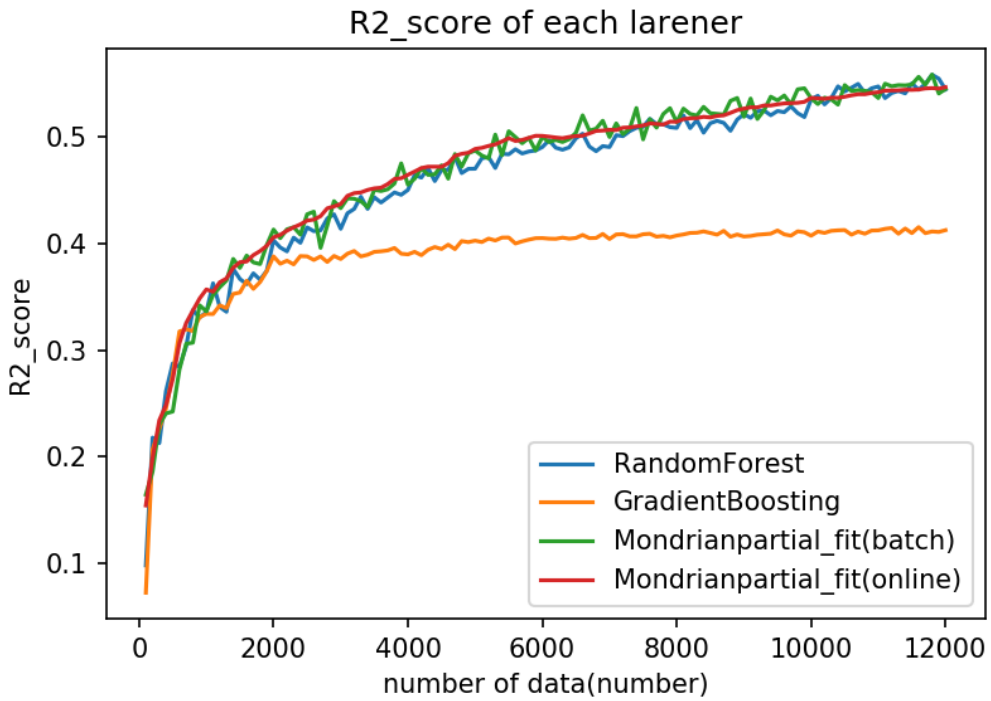
もちろんだが、この状況では最高に Mondrian が強い。
全く見えないのででかいのは省く。



Mondrian_test4_Regression

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。