

MondrianForest テスト

MondrianForest のテストを行った。

<http://archive.ics.uci.edu/ml/index.php> というなんか色々データがあるサイトから落としてきた、以前学習用に遊んでいた adult というデータ(エクセルで修正)を使ってテストを試みる。

データの内容は Age, Workclass, Fnlwgt, Education, Education_number, Marital_Status, Occupation, Relationship, Race, Sex, Capital_Gain, Capital_Loss, Hours_pre_week, Native_Country の 14 項目から、Income が >50K か、≤50K かを判別するというもの。

データ数は 32561 個で、テストデータは答え付きで 16281 個ある。

(計算時間なのでデータ数が多い方がいいと思ってこうしたのですが、処理が重すぎて死にました。勉強になりました。)

欠損値などあったので、seminar の P04S06_Scoring process-2 に従って get_dummies, RFE による特徴量選択などのデータの前処理をした。

特徴量は 15 個に絞り、残ったのは、Age, Fnlwgt, Education_number, Capital_Gain, Capital_loss, Hours_pre_week, Workclass_Self-emp-not-inc, ..., など。

Partial_fit は pipeline できないため、事前に StanderdSchler による標準化を行い、「fit」など自体の処理を計測した。

まず、そもそもの性能を見るために、一度データ数に伴って各学習器の計算時間とスコアがどうなるかを調べる。ここでは、MondrianForest の Partial_fit は毎回リセットし処理を行っているので、蓄積されたデータ数がゼロの時から、データを入れるとどれくらい時間がかかるかを示している。

この時の accuracy score。

また、f1 score

最初の方でほぼ収束してしまっているのを、軸を変えて。

とりあえず、ここから読み取れることは、

- ・Mondrian の Partial_fit と fit はほぼ精度が変わらない(fit の方が若干高い)上に、partial_fit の方がだいぶ遅い。

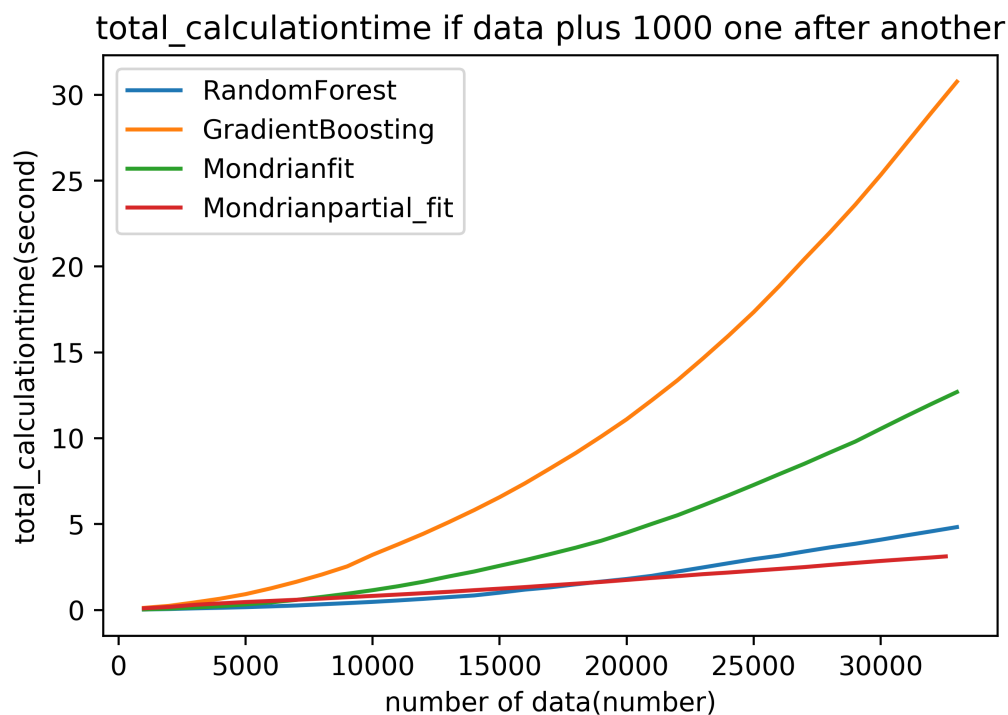
- ・ MondrianForest は RandomForest と比べるとそこまで精度に差はないが、GradientBoosting と比べると有意な差が出てしまっている。

GradientBoosting は α (学習率) など全てデフォルト値でやっているの、スピードより精度重視の設定になってはいるが。

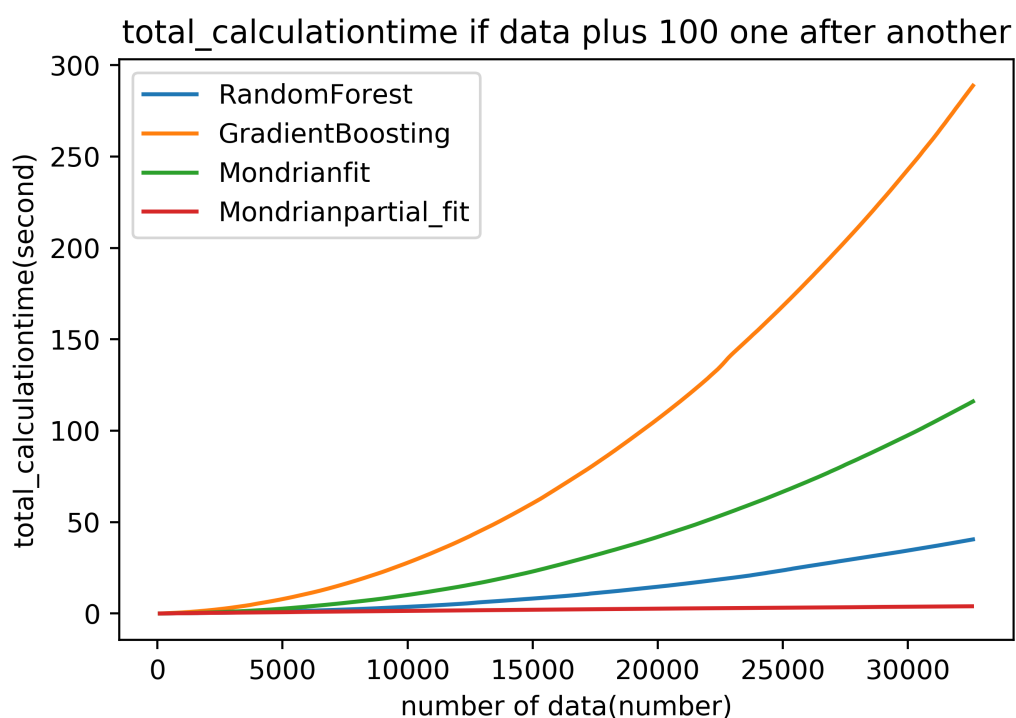
次に、データが 1000 個ずつ次々に到着して来たという状況を仮定する。

その場合の合計の計算時間を調べる。

ここでは、各学習器は(そこまでひどい状況はないと思うが)データが 1000 個追加するたびに再学習させた。Mondrian の partial_fit は 1000 個来るたびに、追加学習させた。



また、データが 100 個ずつ来た場合。



さすがに Mondrian が最も、そして大きく計算時間が早かった。

気になるのはこの時、精度が出ているかということ。

以下で、Mondrianpartial_fit(batch)は、最初に性能を見た時のようにデータ数がゼロから学習させたもの、Mondrianpartial_fit(online_plus100)は 100 個ずつデータを追加学習させたものである。

