



AI WorkShop

金融データを扱った機械学習の演習

フィナンシャル機械学習 第4章 標本の重み付け

2021/10/08

土田晃司

金融データを扱った機械学習の演習

フィナンシャル機械学習 第4章 標本の重み付け

目次

1. 重複した結果とは
2. ラベルの独自性
3. 分類器のバギングと独自性
4. 標準的と逐次のブートストラップ法の比較
5. ラベルの重み付け
6. 演習問題

1. 重複した結果とは

金融のラベリングは、個々のラベルが従う確率分布(周辺確率)がどれも同じで、且つそれらが独立している独立同一分布に従うようなものではないため、重複することがある。

もし、ラベルが重複しないように、ラベリングのタイミングを制御した場合に、もっと良い結果が得られる可能性を制限することになる。

重複した結果が得られた場合にラベルに対して、どのように修正と重み付けを行ったらより良い結果が得られるか、その方法の例を出して問題に取り組む

2. ラベルの独自性

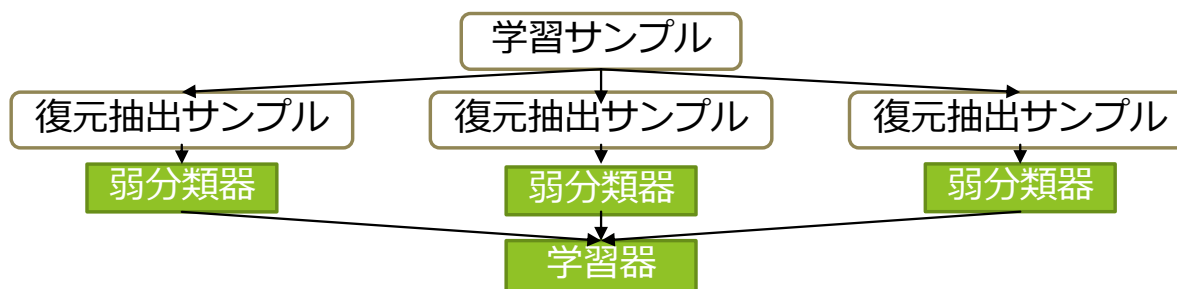
ある区間にラベルが重複した場合に、重複したラベルの件数を数えてラベルの独自性を出す。

ラベルの独自性の平均することで、ラベルを重複しないデータとして修正し、独自性を 0 から 1 の値を割り当てることができるようになる



3. 分類器のバギングと独自性

- ▶ バギング：ブートストラップ法によって、弱分類器を選別し、最終的な結果を多数決する方法



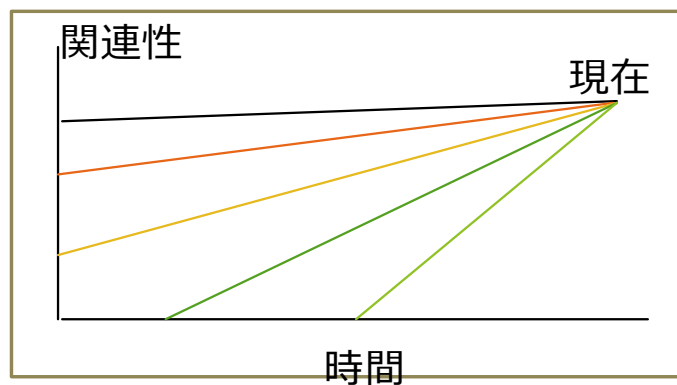
- ▶ 標準的なブートストラップ法では、サンプリングの量が多くなるとバギングを繰り返しても、予想の確率分布がどれも同じだと扱い、重複した結果を除外できず過学習になりやすい
- ▶ 逐次ブートストラップでは、バギングを繰り返す時、2回目移行は重複度の高い結果を抽出する確率を独自性の平均を使って入れ替えることで、重複しなようなラベルを抽出しようという方法

4. 標準的と逐次のブートストラップ法の比較

- ▶ 標準的ブートストラップ法から平均独自性を導く場合と逐次ブートストラップから平均独自性を導く場合を頻度と密度と平均独自性のモンテカル法で、平均独自性の比較を行う。

5. ラベルの重み付け リターン・時間減衰・クラスの重み付け

- ▶ リターンの重み付け：ラベルに対するリターンの大小によっても重み付けを行う必要があり、ある区間のリターンの合計を求めて、その合計によって重み付けを行う。
- ▶ 時間減衰：マーケットで時間が経過するに連れて、古い事例はより新しいものに比べて、関連性を失っていくこと



- ▶ クラスの重み付け：発生頻度は低い重要なイベントに対し重み付けすることで、バギングした分類器の重みの設定を変える

6. 練習問題

- ▶ 3章で、最初のバリアに到達するタイムスタンプの PandasSeriesをt1として、そのインデックスは観測値のタイムスタンプだった。これは、getEvents関数の出力だった。
- 1. E-min S&P500先物のティックデータから身びかれたとるバーについて t1を計算せよ。
- 2. 関数mpNumCoEventsを適用して各時点で重複している結果の数を計算せよ。
- 3. 同時発生的なラベルの数の時系列とリターンの指数加重移動標準偏差の時系列つをプロットせよ。
- ▶ 関数mpSampleTWを使って、各ラベルの平均独自性を計算せよ。この時系列の次数 1 の系列相関AR(1)はいくらだろうか

6. 練習問題

▶ 金融データセットにランダムフォレストを適合させよ

1. アウトオブバックの平均正解率はいくらだろうか
2. K-分割交差検証法の平均正解率はいくらだろうか
3. なぜ交差検証の正解率よりも、アウトオブバックの正解率が非常に高い

▶ 指数時間減衰ファクターを適用せよ

▶ トレンドフォローモデルによって決定される自傷にメタラベルを適用したとしよう。ここで、ラベルのうち3分2は0、3分1は1

1. クラスの重み付けをバランスさせずに分類器に適合させたらどうなるだろうか
2. ラベル1は真陽性をラベル0は偽陽性を意味する、バランスさせたクラスウェイトを適用することで、分類器は真陽性により注意を払い、偽陽性により注意を払わなくなる