Model Stacking/Ensemble

川田恵介

2025-06-17

1 Stacking

1.1 モデルの集計

- OLS, LASSO, Random Forest 等で推定した予測値のうち、どれを使用するのか?
  - テストデータを用いた予測値の集計が有力
    - Random Forest 等と同じアイディア
- 線型モデル $= \omega_{OLS} \times OLSの予測 + \omega_{RF} \times RandomForest予測 + ...$
  - 各予測値を”X”として用いた、線型モデル
- 詳細は Causal ML Chap 8.5 参照

1.2 サンプル分割による推定方法

1. データをサブデータ (Train/Test) にランダム分割
2. Training データを用いて、予測モデルを推定し、Test データに予測値を入力
3. Test データを用いて、予測対象 Y に対して、各予測値で回帰してω を推定

1.3 数値例: Step 1.

```
# A tibble: 9 × 3
  StationDistance    Price Group
            <int>    <dbl> <fct>
1               9  6.05    2
2               4  3.94    1
3               7 31.0     2
4               1  8.64    1
5               2 -5.99    1
6               7 -4.48    2
7               2 -0.895   1
8               3  0.00785 2
9               1 -3.12    2
```

1.4 数値例: Step 2.

```
# A tibble: 9 × 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    2     -1.67         1.29
2               4  3.94    1     NA           NA
3               7 31.0     2     -0.751        1.29
4               1  8.64    1     NA           NA
5               2 -5.99    1     NA           NA
6               7 -4.48    2     -0.751        1.29
7               2 -0.895   1     NA           NA
8               3  0.00785 2      1.08         1.29
9               1 -3.12    2      2.00         1.29
```

1.5 数値例: Step 3

- 交差推定から Stacking も可能

```
lm(Price ~ 0 + OLS + RandomForest, PopData, subset = Group == 2)
```

```

Call:
lm(formula = Price ~ 0 + OLS + RandomForest, data = PopData,
    subset = Group == 2)

Coefficients:
         OLS  RandomForest
      -3.943         4.520
```

- ω を非負、総和を 1 に基準化することも有効 (Laan, Polley and Hubbard, 2007)

2 交差推定の活用

2.1 交差推定; Cross estimation

- Train/Test への分割は、予測モデルやωの推定に、データの一部しか利用できない
  - 事例数が限られている場合、“もったいない”
- 交差推定を用いれば、全事例が活用できる
  - モデルの性能評価 (交差検証; Cross-validation)や LASSO 等におけるλ の選択にも利用される
- Causal ML 3.B 参照

2.2 推定方法

1. データをサブデータ $(1, .., G)$ にランダム分割
2. 第 1 サブデータ**以外**で予測モデルを推定し、第 1 サブデータを予測
3. 第 2 サブデータ**以外**で予測モデルを推定し、第 2 サブデータを予測
4. 以上を全てのデータについて繰り返す
5. 予測対象 Y に対して、各予測値で回帰してω を推定

2.3 数値例: 3 分割

```
# A tibble: 9 × 3
  StationDistance    Price Group
            <int>    <dbl> <fct>
1               9  6.05    3
2               4  3.94    2
3               7 31.0     3
4               1  8.64    1
5               2 -5.99    3
6               7 -4.48    1
7               2 -0.895   1
8               3  0.00785 2
9               1 -3.12    2
```

2.4 数値例: Step 1

```
# A tibble: 9 × 5
  StationDistance    Price Group   OLS RandomForest
            <int>    <dbl> <fct> <dbl>        <dbl>
1               9  6.05    3     NA           NA
2               4  3.94    2     NA           NA
3               7 31.0     3     NA           NA
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3     NA           NA
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2     NA           NA
9               1 -3.12    2     NA           NA
```

- Group 2,3 を Training データとして活用

2.5 数値例: Step 2

```
# A tibble: 9 × 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
```

```
1                9  6.05    3     NA           NA
2                4  3.94    2      4.86        -0.189
3                7 31.0     3     NA           NA
4                1  8.64    1     -4.12        -1.89
5                2 -5.99    3     NA           NA
6                7 -4.48    1     12.9         16.7
7                2 -0.895   1     -1.29        -1.91
8                3  0.00785 2      3.55        -0.189
9                1 -3.12    2      0.938        1.91
```

- Group 1,3 を Training データとして活用

2.6 数値例: Step 3

```
# A tibble: 9 × 5
  StationDistance    Price Group    OLS RandomForest
            <int>    <dbl> <fct>  <dbl>        <dbl>
1               9  6.05    3     -4.88        -1.84
2               4  3.94    2      4.86        -0.189
3               7 31.0     3     -3.03        -1.84
4               1  8.64    1     -4.12        -1.89
5               2 -5.99    3      1.61         0.945
6               7 -4.48    1     12.9         16.7
7               2 -0.895   1     -1.29        -1.91
8               3  0.00785 2      3.55        -0.189
9               1 -3.12    2      0.938        1.91
```

- Group 1,2 を Training データとして活用

2.7 数値例: Stacking

- 交差推定から Stacking も可能

```
lm(Price ~ 0 + OLS + RandomForest, PopData)
```

```

Call:
lm(formula = Price ~ 0 + OLS + RandomForest, data = PopData)

Coefficients:
         OLS  RandomForest
     -1.1627        0.3721
```

- ω を非負、総和を 1 に基準化することも有効 (Laan, Polley and Hubbard, 2007)

2.8 モデルの評価

- Training/Test の分割し、Training データのみで(交差推定)を用いて Stacking モデルを推定し、Test データで評価できる
- 用いるアルゴリズムの数が、Training データの事例数に比べて少数であれば、最終段階での OLS における平均二乗誤差などを用いて、近似的な評価ができる
  - X が事例数に比べて少ないケースと同じ
  - 経済学の多くの応用において、成立
- 大量のアルゴリズムを集計する場合は、過剰適合が生じ、評価できない
  - 最終段階で LASSO などを用いることも検討すべき

2.9 実践

- 実装用パッケージは、大量存在
  - R: SuperLearner, mlr3verse + mlr3pipelines, tidymodels
- アルゴリズムの多様性が重要
  - 少なくとも線型モデルと回帰木系統のモデル、ベースとなるモデル(単純平均や単純な OLS)は含めるべき
    - より詳細な議論は、Phillips et al. (2023) などを参照

2.10 Reference

Bibliography

Laan, M.J. Van der, Polley, E.C. and Hubbard, A.E. (2007) “Super learner,” Statistical applications in genetics and molecular biology, 6(1).

Phillips, R.V. et al. (2023) “Practical considerations for specifying a super learner,” International Journal of Epidemiology, 52(4), pp. 1276–1285.