罰則付き回帰モデル

川田恵介

罰則付き回帰モデル

- X の数が多い場合、線形予測モデルの推定は困難
- 決定木 (RandomForest) は有力な代替案だが、X の数が極めて多くなると機能しなくなる
- 有力な選択肢は、線形予測モデル推定の改良

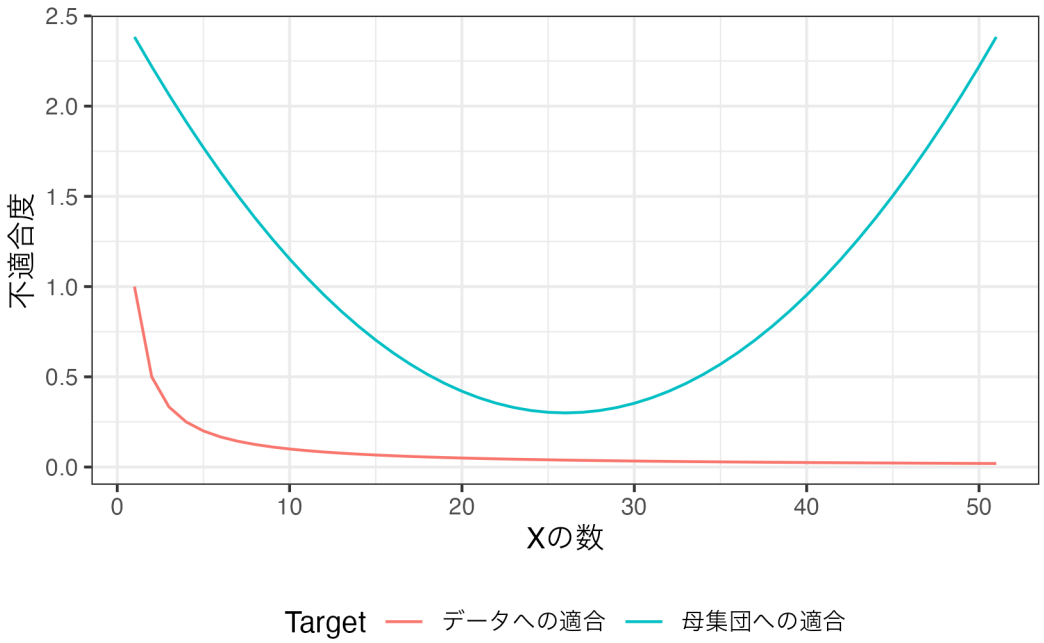
復習

- 線形予測モデル

$$g(X) = \beta_0 + .. + \beta_L X_L$$

- データに当てはめるように推定
  - β の数が多くなると、予測性能が悪化
  - 過剰適合

イメージ



対応

- 環境税などと同じアイディア
- 自動車は便利な道具であるが、同時に排気ガス/渋滞など負の外部性が存在
  - 何も対応しないと保有台数が過大になりうる
- 適切な水準に誘導するために、自動車税を貸す
- **何も対応しないと複雑なモデルになりすぎるので、複雑性に課税する**

罰則付き回帰

- 線形モデル $g(X)$ を、以下を最小化するように推定する

$$\text{データへの当てはまり} + \underbrace{\lambda \times \text{複雑性}}_{\text{複雑性への課税}}$$

- λ：課税額
  - 交差推定で決定
  - 母集団の当てはまり最大化を目指す

複雑性の指標

- Ridge: $\beta_1^2 + .. + \beta_L^2$
- LASSO: $|\beta_1| + .. + |\beta_L|$
- OLS: “0”

LASSO の利点

- 予測において重要ではない β を、厳密に 0 にできる
  - 重要ではない変数をモデルから除外する
- OLS や Ridge では、厳密に 0 にはできない

テキスト分析への有効性

- 単語数が多い $\rightarrow$ X が多い $\rightarrow$ 重要ではない単語も多いかも?
- LASSO が有効な場面も多い

例

```
75 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)        0.416060017
新型コロナ          .
対策                0.219799986
政策                .
評価                0.004301000
日本                .
現状                .
緊急事態宣言        .
行動               -0.061143788
経済                0.134167999
効果               -0.101701950
分析               -0.025663484
解除                0.221239332
危機                .
感染               -0.011528803
```

データ	-0.078310301
マクロ経済	0.040566237
コロナ	.
ショック	.
ウイルス	.
感染拡大	.
影響	.
対応	.
ウィルス感染	.
社会	-0.001772270
考察	-0.165931588
コロナショック	.
pos	0.399362564
みる	.
購買動向	.
地域	.
サービス	.
消費	.
金融	.
財政	.
体制	.
ワクチン接種	.
ウイルス感染拡大	.
医療	.
支援	0.008889229
covid-19	-0.005062291
ワクチン	.
教育	.
格差	0.350948735
比較	.
及ぼす	0.535560140
拡大	.
流行	.
2020	.
企業	-0.086510087
雇用	.
ウイルス感染	-0.003114238
調査	-0.037371384
被害	.
集中	0.263074731

男女	-0.016724458
与える	.
倒産	.
歴史	.
変化	.
リスク	-0.169136243
モデル	-0.226852226
外出	-0.172451078
関係	-0.127511887
伴う	-0.110177684
家計	-0.139945887
役割	-0.134354955
実証	-0.004942870
第一	-0.063500591
たか	-0.170915385
状況	-0.189837039
結果	.
在宅	-0.183642752
組織	-0.185439026
対応緊急調査	.

まとめ

- 事例数を大きく超える X から、予測モデルを構築することは困難なチャレンジ
  - 一つのアプローチは、LASSO
  - 発展: DeepLearning の重要な応用分野