УДК 004.8

Резнік Р.Ю., студент 1 курсу спеціальності «Комп'ютерні технології обробки даних (DataScience)»
Нескородєва Т.В., д. т .н., доцент, завідувач кафедри інформаційних технологій

ВІДБІР ТА РЕГУЛЯРИЗАЦІЯ ЛІНІЙНИХ МОДЕЛЕЙ

Донецький національний університет імені Василя Стуса, м. Вінниця

В регресійних задачах для опису зв'язку між відгуком Y та набором змінних $X_1, X_2, X_3, \ldots, X_p$ часто використовують стандартну лінійну модель, зазвичай цю модель підганяють за методом найменших квадратів[1].

Часто відбувається так, що деякі або більшість змінних, включених в регресійну модель, в дійсності не пов'язані з відгуком. Включення таких невпливових змінних приводить до непотрібної складності кінцевої моделі. Видаливши такі змінні, тобто прирівнявши їх коефіцієнти до 0, можна отримати наочнішу модель, яку буде легше інтерпретувати[2]. Існує багато альтернатив використанню метода найменших квадратів для підгонки моделі, як класичних так і сучасних.

Відбір підмножини змінних: цей підхід включає певний набір змінних, які, на нашу думку, пов'язані з відгуком. І далі ми підганяємо модель за методом найменших квадратів з використанням цього зменшеного набору змінних.

Стиснення: цей підхід включає підгонку моделі, яка містить всі p предикторів. Однак, навідміну від коефіцієнтів, отриманих в методі найменших квадратів, тут коефіцієнти «стискаються» в напрямку до 0. Ефектом такого стиснення(відомого як «регуляризація») є зниження дисперсії. В залежності від типу виконуваного стиснення оцінки деяких коефіцієнтів можуть виявитися в точності рівними 0. А отже, методи стиснення можуть виконати відбір змінних.

Щоб виконати відбір оптимальної підмножини, ми підганяємо окрему регресію за методом найменших квадратів для всіх можливих комбінацій із p предикторів. Іншими словами ми підганяємо всі p моделей, які містять лише один предиктор, всі $\binom{p}{2}$ моделей, які містять два предиктора, і т.д. Далі проводиться перевірка всіх отриманих моделей з метою визначення найбільш оптимальної з них. Не дивлячись на те що відбір оптимальної підмножини предикторів є простим і концептуально привабливим, він страждає обмеженнями обчислювального характеру.

Через проблеми обчислювального характеру, при великому p відбір оптимальної підмножини предикторів є неприйнятним. Покрокове включення змінних є ефективною з обчислювальної точки зору альтернативою. Алгоритм покрокового включення починає з моделі яка не містить ніяких змінних, а потім

додає по одному предиктору в цю модель до тих пір, поки не будуть включені всі предиктори. Зокрема, на кожному кроці додається зміна, яка забезпечує найбільший приріст якості моделі. Подібно до покрокового включення, покрокове виключення змінних також забезпечує ефективну альтернативу. Однак, на відміну від покрокового включення, цей метод стартує з повної моделі, і на кожному кроці вилучає найменш корисні предиктори.

Також існують гібридні алгоритми покрокового включення та виключення. Але в них, після додавання кожної нової змінної, будь-які інші, які більше не забезпечують покращення якості моделі можуть бути видалені. Цей метод поєднує в собі всі позитивні риси відбору оптимальної підмножини та покрокового включення і виключення, а саме: швидкодію та якість моделі.

В якості альтернативи попереднім методам ми можемо побудувати модель зі всіма p предикторами за допомогою метода, який виконає обмеження, або регуляризацію, оцінок коефіцієнтів, або, іншими словами, стисне оцінки коефіцієнтів в напрямку до 0. Двома найбільш відомими методами стиснення регресійних коефіцієнтів до 0 є гребнева регресія та ласо.

Гребнева регресія дуже схожа на регресію за методом найменших квадратів, за винятком того, що коефіцієнти оцінюються шляхом мінімізації дещо іншої величини. Зокрема, оцінки коефіцієнтів гребневої регресії β^R представляють собою значення, які мінімізують[1]:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = RSS + \lambda \sum_{j=1}^{p}\beta_j^2 \qquad (1)$$

У гребневої регресії є один явний недолік. Навідміну від попередніх методів, які зазвичай вибирають моделі з деякою підмножиною змінних, гребнева регресія буде включати в кінцеву модель всі p предикторів.

Метод ласо є відносно недавно розробленою альтернативою гребневої регресії, який нівелює цей недолік. Єдина відмінність полягає в тому, що в методі ласо, член в штрафному додатку β_j^2 замінений на $|\beta_j|$. Подібно до гребневої регресії, ласо стискає оцінки коефіцієнтів в напряму до 0. Однак в випадку з ласо при достатньо великому параметрі робить деякі оцінки коефіцієнтів справді рівними 0. В результаті чого моделі, отримані за допомогою метода ласо, зазвичай легше інтерпретувати[3].

Для порівняння даних підходів було створені 3 моделі [4]: відбір оптимальної підмножини (рис.1), на основі гребневої регресії (рис.2) та метода ласо (рис.3). Дослідження проводилося на основі датасету Restaurant Business Rankings 2020 [5], фрейм даних містить 50 записів які характеризуються 9 змінними:

- Rank – рейтинг ресторану в топ 50;
- Restaurant – назва ресторану;
- Location – місце появи першого ресторану;
- Sales – продажі 2019 року в млн.$;
- YOY_Sales – зростання продажів за рік у %
- Units – кількість закладів
- YOY_Units – щорічне збільшення закладів у %
- Unit_Volume – середня вартість одного закладу у 2019 тис.$;
- Franchising – чи надає ресторан свою франшизу.

Моделі синтезувались для прогнозування змінної YOY_Sales.

```
(Intercept)        Rank   YOY_Units
 19.1062345  -0.3048700   0.8149075
```

Рисунок 1 – відібрані змінні та їх оцінки за методом відбору оптимальної підмножини

```
 (Intercept)  (Intercept)        Rank       Sales       Units
24.0998436999 0.0000000000 -0.3589370601 -0.0589067560 0.0560364111
    YOY_Units  Unit_Volume  Franchising
0.6990154040 -0.0001751964 -0.1891396162
```

Рисунок 2 - відібрані змінні та їх оцінки на основі гребневої регресії

```
 (Intercept)  (Intercept)        Rank       Sales       Units     YOY_Units
18.6115042354 0.0000000000 -0.2499263234 0.0000000000 0.0009136236 0.7807299523
  Unit_Volume  Franchising
0.0000000000 0.0000000000
```

Рисунок 3 - відібрані змінні та їх оцінки на основі метода ласо

Таблиця 1 – показники точності отриманих моделей

Назва моделі	Відбір оптимальної підмножини	Гребнева регресія	Метод ласо
MAE(Середня абсолютна похибка)	122,56	130,55	124,94

Як показано на рис.1 оптимальна модель містить дві змінні: Rank та YOY_Units, в даному прикладі наглядно продемонстровано недолік гребневої регресії та перевага метода ласо. А саме: в гребневу модель потрапили всі змінні, деякі з них з малими коефіцієнтами, що значно ускладнює її інтерпретацію, а от в модель ласо потрапило три зміні, це зменшує складність моделі, та покращує її інтерпретацію, також, з табл.1 видно, що вона має кращий показник точності.

Список літератури:

1. *Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten – «An Introductionto Statistical Learning: with Applications in R» – Springer, 2013. 456p.*
2. *В.В. Христиановский, Т.В. Нескородева, Ю.Н. Полшков Экономико-математические методы и модели: практика применения в курсовых и дипломных работах. Донецк, 2012.*
3. *Классификация, регрессия и другие алгоритмы Data Mining с использованием R [Електронний ресурс]. – Режим доступу до ресурсу: https://ranalytics.github.io/data-mining/042-Regularization.html (дата звернення 15.04.2022)*
4. *Хмелівський Ю.С Нескородєва Т.В. Аналіз даних для прогнозування серцевої недостатності засобами мови R. Матеріали II всеукраїнської науково-практичної конференції для студентів, аспірантів та молодих вчених "Компютерні технології обробки даних" (10 грудня 2021 року) - Вінниця: ДонНУ імені Василя Стуса., с.57-60.*
5. *Restaurant Business Rankings 2020 [Електронний ресурс]. – Режим доступу до ресурсу: https://www.kaggle.com/datasets/michau96/restaurant-business-rankings-2020 (дата звернення 13.04.2022)*

УДК 004.82

Резнік Р.Ю., студент 1 курсу спеціальності «Комп'ютерні технології обробки даних (DataScience)»
Штовба С.Д., д-р. техн. наук, професор, професор кафедри інформаційних технологій