УДК 004.8

ПРИМЕНЕНИЕ РЕГУЛЯРИЗАЦИИ ПРИ ПОСТРОЕНИИ ПОЛИНОМИАЛЬНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ НА ПРИМЕРЕ ПРОГНОЗИРОВАНИЯ СТОИМОСТИ БРИЛЛИАНТОВ

Облакова Т.В., Григорян В.М., Зубарев К.М.

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана», Москва, e-mail: zubarev.bmstu@mail.ru

Цель исследования – применение регрессионных моделей при прогнозировании стоимости бриллиантов на основе следующих признаков: вес, размер, цвет, чистота, качество огранки и пр. Для построения полиномиальных моделей авторы используют методы регуляризации L1 и L2, применяемые для улучшения обобщающей способности моделей машинного обучения и уменьшения их склонности к переобучению. Метод L1, также известный как лассо-регуляризация, способствует разреженности модели, что помогает в отборе значимых признаков и упрощении интерпретации результатов. В свою очередь, метод L2, или ридж-регуляризация, наказывает большие веса и тем самым способствует сглаживанию модели, что может улучшить ее стабильность и обобщение. Разработанные алгоритмы протестированы на сгенерированном примере, который наглядно иллюстрирует, как методы регуляризации позволяют решить проблему переобучения в случае большого числа признаков. На основе разработанных алгоритмов была построена модель полиномиальной регрессии, позволяющая спрогнозировать цену на бриллианты, качество проверено с помощью коэффициента детерминации, значение которого показывает, насколько хорошо модель соответствует данным. В конце авторы приводят сравнение различных методов построения регрессии, которое показывает преимущество использования методов регуляризации.

Ключевые слова: регуляризация, коэффициент детерминации, машинное обучение, переобучение, линейная модель, полиномиальная регрессия

APPLICATION OF REGULARIZATION IN CONSTRUCTING POLYNOMIAL REGRESSION MODELS USING THE EXAMPLE OF FORECASTING THE PRICE OF DIAMONDS

Oblakova T.V., Grigoryan V.M., Zubarev K.M.

Bauman Moscow State Technical University, Moscow, e-mail: zubarev.bmstu@mail.ru

The purpose of this work is to apply regression models in predicting the cost of diamonds based on the following features: weight, size, color, clarity, cut quality, etc. To build polynomial models, the authors use L1 and L2 regularization methods, which are used to improve the generalization ability of machine learning models and reduce their tendency to overfitting. The L1 method, also known as lasso regularization, promotes model sparseness, which helps in selecting significant features and simplifying the interpretation of results. In turn, the L2 method, or ridge regularization, penalizes large weights and thereby promotes smoothing of the model, which can improve its stability and generalization. The developed algorithms are tested on a generated example, which clearly illustrates how regularization methods can solve the problem of overfitting in the case of a large number of features. Based on the developed algorithms, a polynomial regression model was built that allows predicting the price of diamonds, the quality was checked using the determination coefficient, the value of which shows how well the model fits the data. At the end, the authors provide a comparison of different methods for constructing regression, which shows the advantage of using regularization methods.

Keywords: regularization, coefficient determination, machine learning, retraining, linear model, polynomial regression

Введение

Линейная регрессия является фундаментальным методом в машинном обучении, широко используемым для предсказания непрерывных результатов [1, 2]. Однако, когда данные большой размерности, линейные регрессионные модели могут стать склонными к переобучению, что приводит к плохой обобщающей способности [3 с. 57]. При прогнозировании цен на бриллианты необходимо учитывать множество независимых друг от друга признаков, таких как вес, цвет, качество огранки, прозрачность и пр. В случае применения модели полиномиальной регрессии количество признаков вырастает кратно степени полинома и вследствие этого модель начинает излишне хорошо ложиться на тренировочный набор данных, что, естественно, приводит к высокому качеству на обучающих данных [4]. Но при этом страдает ее предсказательная способность на тренировочном наборе, то есть модель начинает подстраиваться под данные вместо того, чтобы искать зависимости, это явление и называется переобучением [3, с. 58].

Для улучшения интерпретируемости модели авторы используют такие методы, как L1 и L2 регуляризация [5], в которых для уменьшения переобучения по сути добавля-

ется штрафной член к функции потерь, который препятствует большим весам, способствуя более простым моделям [6], которые лучше обобщаются на новые данные [4].

Целью исследования является реализация алгоритмов L1 и L2 регуляризации и их применение для построения полиномиальной регрессионной модели, позволяющей спрогнозировать цену на бриллианты, в зависимости от их веса, цвета, качества огранки, прозрачность и прочих признаков.

Материалы и методы исследования

Для построения модели, которая будет предсказывать цену на бриллианты, был использован набор данных с Kaggle. В качестве признаков выделяются: вес бриллианта в каратах; качество огранки (удовлетворительное, хорошее, очень хорошее, высшее, идеальное); цвет бриллианта, от J (худший) до D (лучший) [7]; чистота – мера чистоты бриллианта (I1 (худшая), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (наилучшая)); x, y, z – размеры бриллианта; глубина; ширина площадки верха алмаза относительно самой широкой точки. Для категориальных признаков, таких как качество огранки, цвет, чистота проведем перекодировку их в числовые. Пример преобразованных данных представлен на рис. 1.

	carat	cut	color	clarity	depth	table	price	x	y	z	log_price	volume
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43	5.786897	38.202030
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31	5.786897	34.505856
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31	5.789960	38.076885
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63	5.811141	46.724580
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75	5.814131	51.917250

Рис. 1. Первые 4 записи в датасете

Для определения стоимости бриллиантов, в зависимости от выбранных признаков была использована модель полиномиальной регрессии [8, 9], где для аппроксимации будем искать многочлен, наилучшим образом описывающий данные. То есть входная переменная одна и находятся оптимальные коэффициенты многочлена степени k, описывающие данные из D = $\{(x_i, y_i)\}_{i=1}^{N}$, обозначим $\bar{\boldsymbol{x}} = (x, x^2, \ldots, x^k)$, тогда отклик или моделируемую переменную можно выразить следующим образом:

$$\hat{y}(x) = w_0 + \sum_{i=0}^{k} w_i \bar{\boldsymbol{x}}_i = (1 x x^2 \ldots x^k)^T w. \quad (1)$$

В случае регуляризации L1, также называемой Лассо, в уравнение (1) добавляется дополнительное слагаемое, налагающее «штраф» за сложность модели, то есть высокие веса (ограничивая сумму абсолютных значений параметров модели $|w_i|$):

$$L_1 = \sum_{i=1}^{N} \left(y_i - \hat{y}(\boldsymbol{x}_i, \boldsymbol{w}) \right)^2 + \alpha \sum_{i=1}^{N} |w_i|. \quad (2)$$

Такой подход помогает решить задачу отбора признаков, так как в процессе построения модели с использованием L1 регуляризации некоторые весовые значения могут оказаться нулевыми. Это позволяет исключить признаки, которые слабо влияют на целевую переменную, и тем самым способствует упрощению модели и улучшению обобщающей способности

Для сравнения, в работе будет рассмотрена регуляризация L2, также называемая Ridge регрессией, которая добавляет к целевой функции штрафную функцию

$$L_2 = \sum_{i=1}^{N} \left(y_i - \hat{y}(\boldsymbol{x}_i, \boldsymbol{w}) \right)^2 + \alpha \left\| \boldsymbol{w} \right\|^2. \quad (3)$$

В отличие от регуляризации L1, построение Ridge регрессии не приводит к обнулению ни одного из параметров модели, но также позволяет повысить устойчивость модели.

Результаты исследования и их обсуждение

Реализованные алгоритмы были протестированы на следующем примере: для обучающей выборки были заданы 60 точек (x, y), распределенных равномерно на промежутке $[-5,5] \times [-5,5]$, посчитаем значение функции в этих точках согласно уравнению (4) и добавим к результату шум, имеющий нормальное стандартное распределение.

$$f(x,y) = 0.1x^2 + 0.1y^2 - 4\cos(x) - 4\cos(y),$$
$$(x,y) \in [-5,5] \times [-5,5], \quad (4)$$

На рис. 2 изображен график поверхности, описанной уравнением (4), и смоделированные значения.

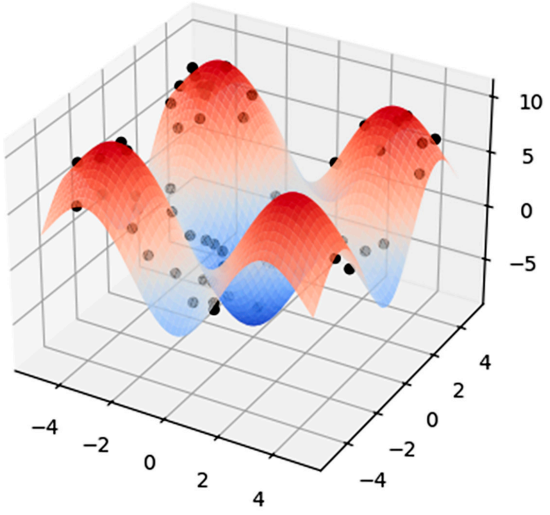

Рис. 2. График функции

$$f(x,y) = 0.1x^2 + 0.1y^2 - 4\cos(x) - 4\cos(y)$$

Для тестовой выборки было сгенерировано 500 значений, и в качестве базовых функций для регрессионной модели были выбрана модель третьей степени из элементов *x, y,* cos(*x*), cos(*y*), sin(*x*), sin(*y*).

В результате получили явно переобученную модель, поверхность проходит очень близко ко всем сгенерированным точкам, что можно видеть на рис. 3, значение среднеквадратичной ошибки (MSE) для тестового набора данных оказалось равным 1051.755.

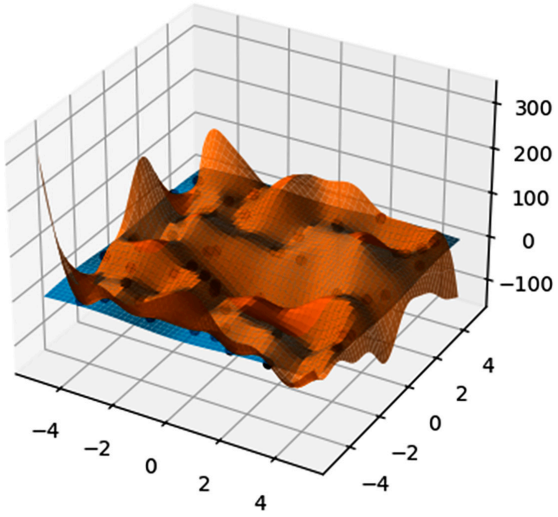

Рис. 3. Результат переобучения

При построении модели с использованием регуляризации L1 наилучшее значение MSE [10] получилось равным 0.661, при коэффициенте регуляризации равным α = 0.071.

Также для данного набора данных была построена модель с использованием L2 регуляризации, наилучшее значение среднеквадратической ошибки, равное 1.2, было получено при коэффициенте регуляризации α = 3.162.

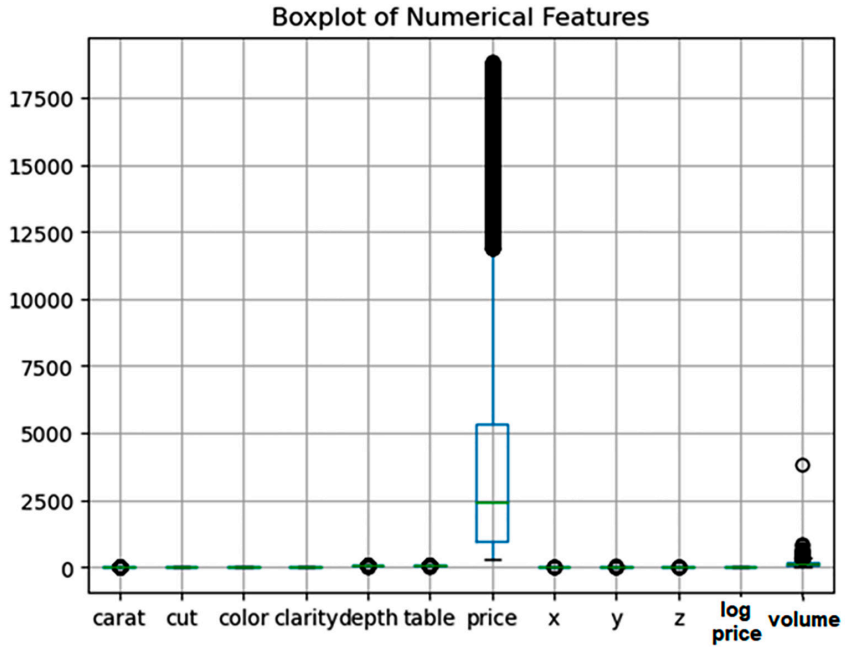

Рис. 4. Проверка на наличие шумов в данных

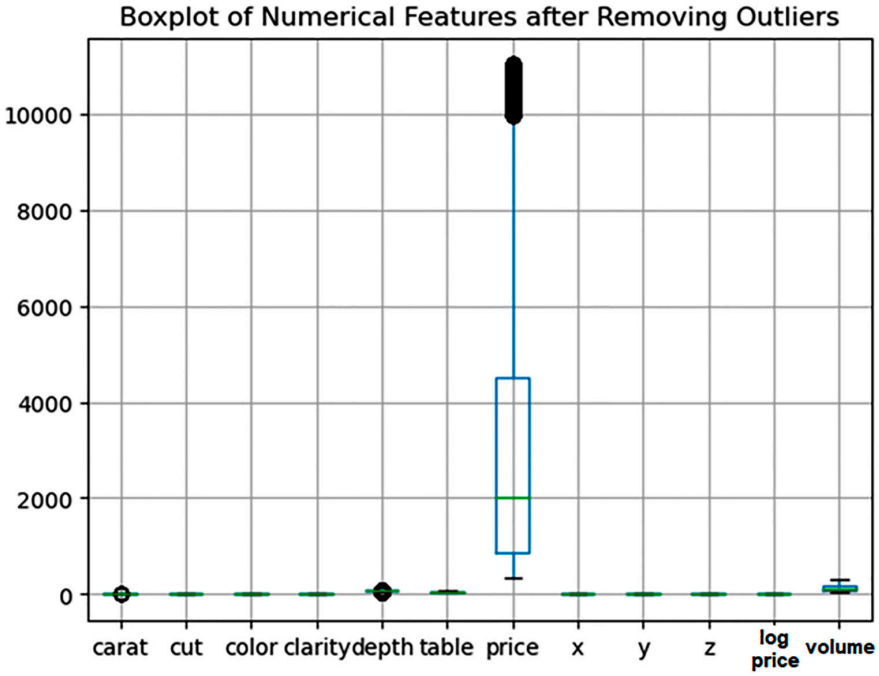

Рис. 5. Данные после очистки от шумов

```
Mean Squared Error train: 564659.0794387351
Mean Squared Error test: 554184.568040858
Mean Absolute Error train: 522.7534634175115
Mean Absolute Error test: 517.0462566327137
R^2 train: 0.9167169465262452
R^2 test: 0.9166134196922929
```

Рис. 6. Метрики для линейной модели

Можно заключить, что разработанные алгоритмы действительно позволяют строить модели с большим количеством признаков, исключая явление переобучения. Для построения модели, которая будет определять стоимость бриллиантов также была использована модель полиномиальной регрессии с использованием регуляризации, дальше будет показано ее преимущество перед моделью линейной регрессии.

В качестве первоначальной обработки очистим данные от шумов с помощью boxplot.

Как видим на рис. 4, в выборке присутствует достаточно много шумов. В результате очистки данных было удаленно 7000 записей, результат представлен на рис. 5.

Для начала рассмотрим модель простейшей линейной регрессии для прогнозирования цены на бриллианты. Исключим из выборки значения log_price – логарифм цены, так как очевидно он явным образом зависит от цены. В качестве метрик качества модели были выбраны стандартные MSE – среднеквадратичная ошибка, MAE – Средняя абсолютная ошибка, а также коэффициент детерминации R^2 [10], который показывает насколько хорошо модель соответствует данным. На рис. 6 представлены значения метрик для тестовой и обучающей выборки при построении линейной модели.

Значения для тестовой и обучающей выборок получились примерно одинаковые, следовательно, переобучения в данном случае не происходит, и применять регуляризацию в данном случае не имеет смысла. Также видим, что MAE и MSE не очень показательны в данном случае, поэтому будем ориентироваться на R^2, значение которого для линейной модели получилось очень неплохим, попробуем улучшить качество модели за счет использование полиномиальной модели.

Построим на данных полином 7 степени, а потом для него линейную регрессионную модель. Значение метрик качества для обучающей и тестовой выборки отображены на рис. 7 и 8.

Коэффициент регуляризации	Среднеквадратичная ошибка	Средняя абсолютная ошибка	R2 score
0	1.22577785550419e+19	39152768.16268585	-1810725885012.6782
31.623	257928.6739671999	235.06548932775198	0.9618986324200188
74.989	230297.5356980278	231.64625687853393	0.9659803195766041
177.828	208022.39999091107	229.2810283024707	0.9692708150473742
421.697	194116.29648295703	229.28959688676545	0.9713250324137971
1000.0	187314.38759476	233.38251968957684	0.9723298141885744

Рис. 7. Метрики для обучающей выборки

```
Mean Squared Error train: 23134.512353332666
Mean Absolute Error train: 93.68023637196356
R^2 train: 0.9965608129300068
```

Рис. 8. Метрики для тестовой выборки

```
Mean Squared Error test: 1.22577785550419e+19
Mean Absolute Error test: 39152768.16268585
R^2 test: -1810725885012.6782
```

Рис. 9. Метрики при различных коэффициентах регуляризации Лассо

Коэффициент регуляризации	Среднеквадратичная ошибка	Средняя абсолютная ошибка	R2 score
0	1.22577785550419e+19	39152768.16268585	-1810725885012.6782
0.1	154574.00081704164	225.56334942578096	0.9771662423845618
0.316	152990.0943252507	224.60221893677502	0.9774002179349642
1.0	150957.432089295	226.03546431627333	0.9777004839341914
3.162	150779.99027274727	229.8815657411848	0.9777266957383013
10.0	158012.70868868745	238.9770372859401	0.9766582745397986

Рис. 10. Метрики при различных коэффициентах гребневой регуляризации

В данном случае получили явное переобучение, почти идеальное совпадение результата для обучающей выборки и огромная погрешность на тестовых данных. Попробуем регуляризировать полученную модель.

Применим сначала регуляризацию Лассо с различными коэффициентами, чтобы выявить лучший результат. Возьмем следующие коэффициенты регуляризации: $\alpha = [0.1,0.316,1.0,3.162,10.0]$. На рис. 9 изображены значения MSE, MAE и коэффициента детерминации в зависимости от значения коэффициента регуляризации

Как видно, модель стала намного лучше работать, пропал эффект переобучения. Наилучший результат получился при $\alpha = 3.162$, $R^2 \approx 0.97773$, что также дает преимущество перед моделью линейной регрессии (тот же вывод можно сделать, опираясь на другие метрики качества)

Применим для тех же данных регуляризацию L2 и сравним качества построенных моделей.

Возьмем следующие коэффициенты: $\alpha = [31.623,74.989,177.828,421.697,1000.0]$.

Значение метрик качества модели с использование Ridge регрессии приведено на рис. 10.

Как и в предыдущем случае, можно отметить существенное улучшение качества модели по сравнению с линейной. Максимальное значение метрики $R^2 \approx 0.97773$ при $\alpha = 1000$.

Сравнивая значения ключевых метрик при применении различных методов регуляризации, можно заключить, что серьезного различия в данном случае нет, но применение регуляризации является необходимой мерой при построении полиномиальной модели для прогнозирования цен на бриллианты.

Заключение

Как видно из приведенных исследований, при решении задачи регрессии с большим количеством признаков методы регуляризации L1 и L2, существенно улучшают как точность предсказаний, так и интерпретируемость моделей и оказывают положительное влияние на обобщаю способность. Эти методы играют ключевую роль в предотвращении переобучения, что особенно важно при работе с большими и сложными наборами данных, где модели могут чрезмерно подстраиваться под обучающие данные.

Авторами была получена робастная модель для прогнозирования стоимости бриллиантов в зависимости от веса, прозрачности, качества обработки и других признаков. Перспективные направления дальнейших исследований включают изучение и применение других методов регуляризации, таких как эластичная сеть (Elastic Net), которая комбинирует преимущества L1 и L2 регуляризации, а также более сложных подходов, например, регуляризация в байесовских методах или методов, основанных на априорных знаниях.

Список литературы

1. Зубарев К.М., Безрученко Т.С. Анализ эффективности маркетинговой кампании методами машинного обучения // Дневник науки. 2024. № 6 (90). URL: https://dnevniknauki.ru/images/publications/2024/6/physics/Zubarev_Bezruchenko.pdf (дата обращения: 18.08.2024).

2. Шершакова А.О., Пархоменко В.П. Методы интеллектуального анализа данных в модели наукастинга опасных явлений // Математическое моделирование и численные методы. 2021. № 3 (31). С. 88–104.

3. Николенко С.И., Кадурин А.А., Архангельская Е.О. Глубокое обучение. Погружение в мир нейронных сетей. М.: ИД «Питер», 2018. 481 с.

4. Strijov V., Krymova E., Weber G.W. Evidence optimization for consequently generated models // Mathematical and Computer Modelling. 2013. Vol. 57, Is. 1–2. P. 50–56.

5. Борхани Р., Катсаггелос А., Уатт Д. Машинное обучение: основы, алгоритмы и практика применения / Пер. с англ. СПб.: БХВ-Петербург, 2022. 640 с.

6. Облакова Т.В., Зубарев К.М., Яковлев Д.Ю. Анализ распределения высоты морских волн. Сравнение оценок и применение критерия согласия Пирсона // Дневник науки. 2023. № 12 (84). URL: https://dnevniknauki.ru/images/publications/2023/12/physics/Oblakova_Zubarev_Yakovlev.pdf. (дата обращения: 18.08.2024). DOI: 10.51691/2541-8327_2023_12_32.

7. Сухарев А.Н. Алмазы и бриллианты как инвестиционные инструменты, оценка их стоимости // Финансы и кредит. 2013. № 37 (565). С. 18–23.

8. Su M., Zhong Q., Peng H. Regularized multivariate polynomial regression analysis of the compressive strength of slag-metakaolin geopolymer pastes based on experimental data // Construction and Building Materials. 2021. Vol. 303. DOI: 10.1016/j.conbuildmat.2021.124529.

9. Saqib M. Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model // Applied Intelligence. 2021. Vol. 51. P. 2703–2713.

10. Tyagi K. et al. Regression analysis // Artificial intelligence and machine learning for EDGE computing. Academic Press. 2022. P. 53–63.