Благодаря высокой скорости получения результатов работы модели, терпимости к сложности данных нейронных сетей, относительной зависимости системы на этапе построения от разработчика, а также гибкости и компактности удалось разработать систему, основанную на изучаемых данных, которая способна на основе краткосрочных наблюдений за поведением клиента предсказать вероятность положительной кредитной истории в дальнейшем.

СПИСОК ЛИТЕРАТУРЫ

1. *MacLennan J., Crivat B.* Data Mining with Microsoft SQL Server 2008 / John Wiley and Sons. 2009. 672 p.
2. *Inmon W. H.* Building the Data Warehouse, Third Edition / Wiley Computer Publishing, 2002. 576 p.
3. *Hawkins H.* Data Warehousing Architecture and Implementation / New Jersey : Prentice Hall PTR, 1998. 362 p.
4. *Lichman M.* UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science [Electronic resource]. URL: http://archive.ics.uci.edu/ml (accessed: 03.05.2018).
5. *Berson A., Smith S.* Data Warehousing, Data Mining, and OLAP / N.Y. : McGraw Hill Education, 2017. 638 p.
6. *Zaki M.* Data mining and analysis: Fundamental concepts and algorithms / Cambridge : Cambridge India, 2016. 567 p.
7. *Han J., Kamber M.* Data Mining: Concepts and Techniques / Waltham : Elsevier, 2006. 772 p.
8. *Kantardzic M.* Data Mining: Concepts, Models, Methods, and Algorithms / N.Y. : John Wiley and Sons, 2003. 343 p.
9. *Kovalerchuk B., Vityaev E.* Data Mining in Finance: Advances in Relational and Hybrid Methods / N.Y. : Kluwer Academic Publishers, 2002. 322 p.
10. *Witten I. H., Frank E., Hall M. A.* Data Mining: Practical Machine Learning Tools and Techniques / Waltham : Morgan Kaufmann, 2016. 654 p.

МЕТОДЫ АППРОКСИМАЦИИ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

Д. А. Зайнышева

Саратовский государственный университет, Россия
E-mail: d.zaynysheva@gmail.com

В регрессионном анализе нередко приходится сталкиваться с ситуацией, когда среди объясняющих переменных наблюдается зависимость. Тогда говорят о наличии мультиколлинеарности. В таких ситуациях МНК-оценки формально существуют, но обладают «плохими» статистическими свойствами. Регуляризация — метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу. Эта информация часто имеет вид штрафа за сложность модели. Методы регрессии Ridge и Lasso осуществляют регуляризацию параметров и позволяют преодолеть некоторые недостатки метода наименьших квадратов.

Методы регуляризации зачастую позволяют добиться уменьшения дисперсии прогноза за счет незначительного увеличения его смещения. В результате точность прогноза растет.

Также, в результате применения методов регрессии Ridge и Lasso коэффициент (вес) при некоторых предикторах линейной модели приближается к нулю (или становится равным нулю), благодаря этому модель легче интерпретировать.

METHODS OF APPROXIMATION IN MACHINE LEARNING TASKS

D. A. Zaynysheva

The variables in a regression might be correlated with each other. This concept is known as the multicollinearity. In such situations, OLS estimates formally exist, but have poor statistical properties. Regularization is a method of adding some information to a model to solve some problems. This information often has the form of a penalty for the complexity of the model. The regression methods of Ridge and Lasso regularize the parameters and allow some flaws in the least squares method to be overcome.

Regularization methods often make it possible to reduce the dispersion of prognosis by slightly increase in its displacement. As a result, the accuracy of the forecast increases. Also, as a result of applying the Ridge and Lasso regression methods, the coefficient (weight) at some predictors of the linear model approaches zero (or becomes zero), which makes the model easier to interpret.

Существует много задач, требующих изучения отношения между двумя и более переменными. Для решения таких задач используется регрессионный анализ. В настоящее время регрессия получила широкое применение, включая задачи прогнозирования и управления. Целью регрессионного анализа является определение зависимости между исходной переменной и множеством внешних факторов (регрессоров).

Нередко приходится сталкиваться с ситуацией, когда среди объясняющих переменных наблюдается зависимость. Тогда говорят о наличии мультиколлинеарности. В таких ситуациях МНК-оценки формально существуют, но обладают «плохими» статистическими свойствами.

Регуляризация — метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу. Методы регрессии Ridge и Lasso осуществляют регуляризацию параметров и позволяют преодолеть некоторые недостатки метода наименьших квадратов.

Ridge-регрессия очень похожа на метод наименьших квадратов, за исключением добавления «гребня». Коэффициент λ умножается на l_2-норму вектора коэффициентов. Как и в случае с методом наименьших квадратов, Ridge-регрессия ищет оценки коэффициентов, которые хорошо подходят для данных, минимизируя RSS. Однако добавление штрафа за сокращение приводит к стремлению оценок коэффициентов к нулю.

У Ridge-регрессии есть один недостаток. Штраф λ сжимает все коэффициенты до нуля, но он не будет устанавливать ни одного из них точно в ноль (кроме случая, когда λ равно бесконечности).

Lasso, в отличие от Ridge-регрессии, не только осуществляет регуляризацию, но и приравнивает некоторые из коэффициентов к нулю при достаточно большом значении λ. То есть дополнительно осуществляет выбор подмножест-

ва переменных, что позволяет легче интерпретировать модель. В Lasso используется l_1-норма вектора коэффициентов.

Как и при Ridge-регрессии, оценки метода наименьших квадратов имеют чрезмерно высокую дисперсию, Lasso может дать уменьшение дисперсии за счет небольшого увеличения смещения и, следовательно, может генерировать более точные прогнозы.

Можно ожидать, что Lasso лучше работает в моделях, где относительно небольшое число предикторов имеет существенные коэффициенты, а остальные предикторы имеют очень малые или равные нулю коэффициенты. Ridge-регрессия будет лучше работать, когда ответ будет функцией многих предикторов, все с коэффициентами примерно равного размера. Однако число предикторов, связанных с ответом, никогда не известно заранее для реальных наборов данных.

Рассмотрим реализацию Lasso и Ridge-регрессии на примере тестовых данных в R. Воспользуемся пакетом glmnet.

В качестве тестовых данных будут использоваться данные Hitters. Этот набор данных содержит основные бейсбольные данные лиги за 1986 и 1987 года. Он состоит из 322 наблюдений основных игроков лиги по 20 переменным. В качестве зависимой переменной выбран годовой оклад.

```
> x=model.matrix (Salary~.,Hitters )[,-1]
> y=Hitters$Salary
```

Функция glmnet() содержит аргумент alpha, который определяет, какой вид регрессии используется. Если alpha = 0, то это Ridge-регрессия, если alpha = 1 – Lasso.

Разделим данные на обучающий набор и тестовый, чтобы оценить погрешность теста Ridge-регрессии и Lasso.

```
> train=sample (1: nrow(x), nrow(x)/2)
> test=(- train )
> y.test=y[test]
```

Рассмотрим, какие значения будут получены, при использовании метода наименьших квадратов.

```
> ridge.pred=predict (ridge.mod,s=0, newx=x[test ,])
> mean(( ridge.pred -y.test)^2)
[1] 114723.6
> lm(y~x, subset =train)
Call:
lm(formula = y ~ x, subset = train)
Coefficients:
(Intercept)       xAtBat        xHits       xHmRun
  299.42849     -2.54027      8.36682     11.64512
       xRuns        xRBI       xWalks       xYears
   -9.09923      2.44105      9.23440    -22.93673
      xCAtBat       xCHits      xCHmRun       xCRuns
   -0.18154     -0.11598     -1.33888      3.32838
       xCRBI      xCWalks     xLeagueN   xDivisionW
```

```
  0.07536    -1.07841    59.76065   -98.86233
xPutOuts    xAssists     xErrors  xNewLeagueN
 0.34087     0.34165    -0.64207    -0.67442
```

В общем случае удобно использовать кросс-валидацию для выбора параметра λ. Можно сделать это, используя встроенную функцию перекрестной проверки cv.glmnet().

```
> cv.out =cv.glmnet (x[train ,],y[train],alpha =0)
> bestlam =cv.out$lambda.min
> bestlam
[1] 211.7416
```

Видно, что значение λ, которое приводит к наименьшей ошибке, равно 212. Найдем MSE для этого значения λ.

```
> ridge.pred=predict (ridge.mod ,s=bestlam ,newx=x[test ,])
> mean(( ridge.pred -y.test)^2)
[1] 96015.51
```

Построим модель Ridge-регрессии на полном наборе данных, используя значение λ, полученное путем кросс-валидации.

```
> out=glmnet (x,y,alpha =0)
> predict (out ,type="coefficients",s=bestlam )[1:20 ,]
 (Intercept)        AtBat         Hits        HmRun
 9.88487157   0.03143991   1.00882875   0.13927624
        Runs          RBI        Walks        Years
 1.11320781   0.87318990   1.80410229   0.13074383
     CAtBat        CHits       CHmRun        CRuns
0.01113978   0.06489843   0.45158546   0.12900049
       CRBI       CWalks      LeagueN    DivisionW
0.13737712   0.02908572  27.18227527 -91.63411282
    PutOuts      Assists       Errors   NewLeagueN
0.19149252   0.04254536  -1.81244470   7.21208394
```

Ridge-регрессия с мудрым выбором λ может превосходить результаты, полученные методом наименьших квадратов. Теперь необходимо узнать, может ли Lasso дать либо более точную, либо более интерпретируемую модель, чем Ridge-регрессия. Снова используем функцию glmnet(), однако на этот раз используем аргумент alpha = 1. Выполним кросс-проверку и вычислим тестовую ошибку.

```
> cv.out =cv.glmnet (x[train ,],y[train],alpha =1)
> bestlam =cv.out$lambda.min
> lasso.pred=predict (lasso.mod ,s=bestlam ,newx=x[test ,])
> mean(( lasso.pred -y.test)^2)
[1] 100743.4
```

Это значение существенно ниже, чем тестовое значение MSE метода наименьших квадратов и очень похоже на тестовое значение MSE Ridge-регрессии с λ, выбранным путем перекрестной проверки. Однако Lasso имеет существенное преимущество перед Ridge-регрессией. Видно, что 12 из 19 оценок коэффициентов равны нулю. Таким образом, модель лассо с λ, выбранная

перекрестной валидацией, содержит только семь переменных.

(Intercept)	AtBat	Hits	HmRun
18.5394844	0.0000000	1.8735390	0.0000000
Runs	RBI	Walks	Years
0.0000000	0.0000000	2.2178444	0.0000000
CAtBat	CHits	CHmRun	CRuns
0.0000000	0.0000000	0.0000000	0.2071252
CRBI	CWalks	LeagueN	DivisionW
0.4130132	0.0000000	3.2666677	-103.4845458
PutOuts	Assists	Errors	NewLeagueN
0.2204284	0.0000000	0.0000000	0.0000000

Ridge-регрессия – усовершенствование линейной регрессии с повышенной устойчивостью к ошибкам, налагающая ограничения на коэффициенты регрессии для получения куда более приближенного к реальности результата.

Lasso-регрессия сходна с Ridge, за исключением того, что коэффициенты регрессии могут равняться нулю (часть признаков при этом исключается из модели).

Оба метода успешно решают проблемы мультиколлинеарности, переобучения, и уменьшают разброс коэффициентов. Ridge-регрессия использует все предикторы, стараясь «выжать максимум» из всей имеющейся информации. Lasso производит отбор предикторов, что предпочтительнее, когда среди признаков имеются шумовые, или измерения признаков связаны с ощутимыми затратами.

СПИСОК ЛИТЕРАТУРЫ

1. *Casella G. S.* An Introduction to Statistical Learning with Application in R // Casella G.S., Fienberg I. O. New York: Springer. 2015.

2. *Buhlmann P.* Statistics for High-Dimensional Data // Buhlmann P., Geer S. Berlin: Springer. 2011.

3. *Иконникова И. А.* Эконометрика: учебно-методическое пособие / И.А. Иконникова, Н.А. Вихорь. Томск: Изд-во Том. гос. архит.-строит. ун-та. 2012. 88 с.

4. *Носко В. П.* Эконометрика: учебник для студентов высш. учеб. заведений, обучающихся по экон. специальностям. В 2 ч. М. : Дело, 2011.

5. *Аникина Е. А.* Экономическая теория: учебник / Е.А. Аникина, Л.И. Гавриленко. Томск: Изд-во Томского политехнического университета, 2014. 413 с.

6. *Мардас А. Н.* Эконометрика. СПб: Питер, 2001. 144 с.

7. *Носко В. П.* Эконометрика: учебник для студентов высш. учеб. заведений, обучающихся по экон. специальностям. В 2 ч. М.: Дело, 2011.

8. *Tibshirani R.* Sparsity and the Lasso / R. Tibshirani., 2015.

9. *Савельев А. А.* Основные понятия языка R / А. А. Савельев, С. С. Мухамарова. Казань: Казань, 2007.

10. *Paradis E.* R for beginners / E. Paradis, 2000.