PORÓWNANIE METOD REGRESJI – PAKIET R

Autorzy:

Jakub Kopka

Mateusz Grzela

Toyota Corolla

Opis danych

	Price	Age	KM	FuelType	HP	MetColor	Automatic	CC	Doors	Weight
1	13500	23	46986	Diesel	90	1	0	2000	3	1165
2	13750	23	72937	Diesel	90	1	0	2000	3	1165
3	13950	24	41711	Diesel	90	1	0	2000	3	1165
4	14950	26	48000	Diesel	90	0	0	2000	3	1165
5	13750	30	38500	Diesel	90	0	0	2000	3	1170
6	12950	32	61000	Diesel	90	0	0	2000	3	1170
7	16900	27	94612	Diesel	90	1	0	2000	3	1245
8	18600	30	75889	Diesel	90	1	0	2000	3	1245
9	21500	27	19700	Petrol	192	0	0	1800	3	1185
10	12950	23	71138	Diesel	69	0	0	1900	3	1105
11	20950	25	31461	Petrol	192	0	0	1800	3	1185

Plik z danymi „ToyotaCorolla.csv” to plik zwierający dane dotyczące samochodów Toyota Corolla. Składa się on na 10 kolumn i 1436 wierszy z danymi. Zadanie jakie postawiliśmy sobie dla tych danych to możliwe jak najlepsze przewidywanie wartości samochodu na podstawie innych cech.

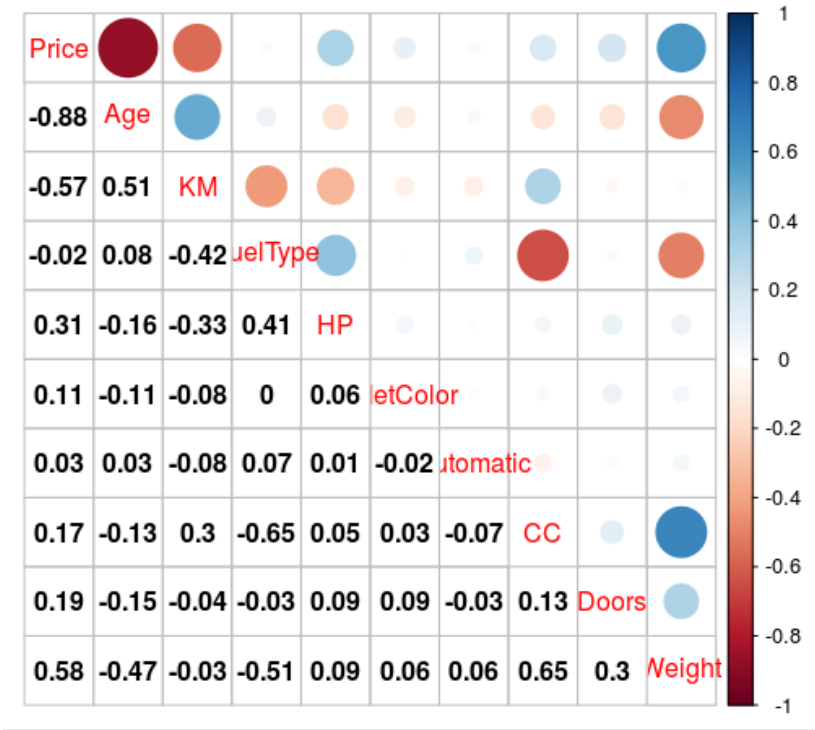
Kolumnami w pliku są:

- „Price” – cena
- „Age” – wiek
- „Km” – liczba przejechanych km
- „FuelType” – typ paliwa (Diesel/Petrol)
- „HP” – moc (konie mechaniczne)
- „MetColor” – kolor nadwozia (true/false)
- „Automatic” – informacja czy skrzynia biegów jest automatyczna
- „CC” – pojemność silnika
- „Doors” – liczba drzwi
- „Weight” – waga pojazdu.

Podsumowanie danych:

```
> summary(corolla)
     Price            Age              KM            FuelType
 Min.   : 4350   Min.   : 1.00   Min.   :     1   CNG    :  17
 1st Qu.: 8450   1st Qu.:44.00   1st Qu.: 43000   Diesel: 155
 Median : 9900   Median :61.00   Median : 63390   Petrol:1264
 Mean   :10731   Mean   :55.95   Mean   : 68533
 3rd Qu.:11950   3rd Qu.:70.00   3rd Qu.: 87021
 Max.   :32500   Max.   :80.00   Max.   :243000
       HP            MetColor        Automatic            CC
 Min.   : 69.0   Min.   :0.0000   Min.   :0.00000   Min.   :1300
 1st Qu.: 90.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1400
 Median :110.0   Median :1.0000   Median :0.00000   Median :1600
 Mean   :101.5   Mean   :0.6748   Mean   :0.05571   Mean   :1567
 3rd Qu.:110.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1600
 Max.   :192.0   Max.   :1.0000   Max.   :1.00000   Max.   :2000
     Doors           Weight
 Min.   :2.000   Min.   :1000
 1st Qu.:3.000   1st Qu.:1040
 Median :4.000   Median :1070
 Mean   :4.033   Mean   :1072
 3rd Qu.:5.000   3rd Qu.:1085
 Max.   :5.000   Max.   :1615
```

Korelacja zmiennych wygląda następująco:



Obserwacje: Istnieje duża negatywna korelacja pomiędzy ceną a wiekiem.

Obróbka danych oraz zbiór testowy i treningowy

```
Multiple R-squared (forward):  0.8615953
Multiple R-squared (backward):  0.8615953
Multiple R-squared (both):  0.8615953
```

Do jak najlepszego dopasowania danych do modelu użyto funkcji step, która automatycznie dobiera kolumny, tak aby regresja była jak najlepsza.

Dla danych ToyotaCorolla funkcja step dobrała model jednakowy dla każdej opcji przeszukiwania (forward, backward i both)

Cechy modelu dobrane automatycznie wyglądają następująco:

Price ~. Age + HP + Weight + KM + FuelType + CC + Doors

Modele regresji

W pakiecie zostały zaimplementowane trzy modele regresji dla danych ToyotaCorolla:

1) LM
2) SVM dla kerneli:
   a) Linear
   b) Polynomial
   c) Radial
   d) Sigmoid

3) GLM
    a) Gaussian
    b) Poisson
    c) Quasi
    d) Quasipoisson

Porównanie modeli i błędów RMSE dla danych Corolla

Model	Family /Kernel	**RMSE** Dane bez obróbki	**RMSE** dane po obróbce
LM	----	**1 207.097**	1 209.073
GLM	**gaussian**	1 207.097	1 209.073
	Gamma	8 073.813	8 073.813
	poisson	8 064.871	8 064.871
	quasi	1 207.097	1 209.073
	quasipoisson	8 064.871	8 064.871
SVM	**linear**	1 113.275	**1 110.204**
	polynomial	1 579.067	**1 372.606**
	radial	1 258.175	**1 145.888**
	sigmoid	**20 713.58**	28 884.79

Podsumowanie:

Najlepszy model z najniższym błędem RMSE dla danych ToyotaCorolla jest model SVM z funkcją kernelową liniową nauczony na danych obrobionych. Błąd ten wynosi 1110.204 w przewidywanej cenie samochodu.

Najlepszym modelem GLM jest model z rodziny gaussian i quasi. Modele te zostały nauczone i przetestowane na nieobrobionych danych. Błąd RMSE dla nich wynosi 1 207.097. W porównaniu do najlepszego modelu ze wszystkich jest to niewielka różnica. Trzema najgorszymi modelami dla tych danych są modele z rodziny gamma (RMSE = 8 073.813 i 8 073.813), quasipoisson (RMSE = 8 064.871 i 8 064.871) oraz Poisson (RMSE = 8 064.871, 8 064.871)

Najlepszym modelem SVM jest model z funkcją kernelową liniową (RMSE = 1 113.275, 1 110.204). Najgorszym modelem SVM jest model z funkcji kernelowej sigmoid (RMSE = 20 713.58, 28 884.79).

MtCars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1

Plik z danymi mtcars to plik zwierający dane dotyczące różnych samochodów. Zbior danych składa się na 11 kolumn i 32 wiersze z danymi. Zadanie jakie postawiliśmy sobie dla tych danych to możliwe jak najlepsze przewidywanie gsec (czas na ¼ mili) na podstawie innych cech.

Cechami danych są:

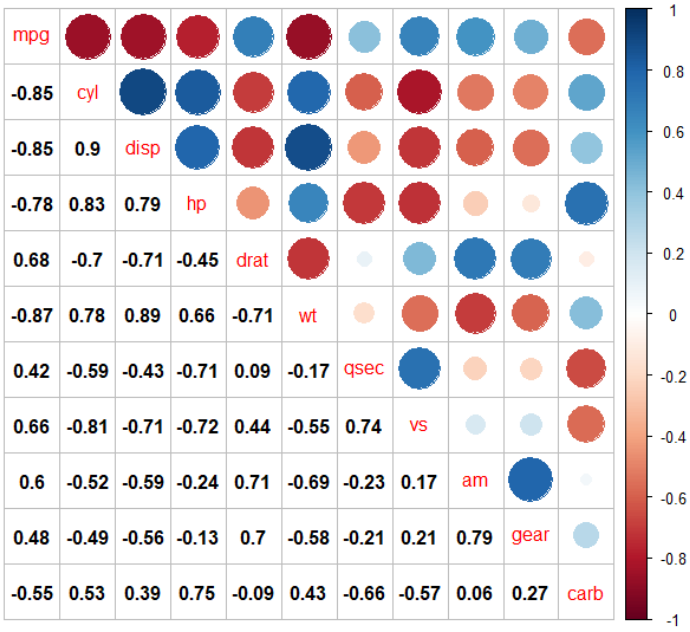
- „mpg” – spalanie (mil/galon)
- „cyl” – Liczba cylindrów
- „disp” – pojemność silnika w cu in
- „hp” – moc silnika w KM
- „drat” – przełożenie tylnej osi
- „wt” – waga (1000 lbs)
- „qsec” – czas na ¼ mili
- „vs” – typ silnika (0 – w kształcie V, 1 - prosty)
- „am” – skrzynia biegów (0 – automatyczna, 1 -manualna)
- „gear” – liczba biegów do jazdy do przodu
- „carb” – liczba gaźników

Podsumowanie danych:

mpg	cyl	disp	hp	drat	wt
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581
Median :19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695	Median :3.325
Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597	Mean :3.217
3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920	3rd Qu.:3.610
Max. :33.90	Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930	Max. :5.424

qsec	vs	am	gear	carb
Min. :14.50	Min. :0.0000	Min. :0.0000	Min. :3.000	Min. :1.000
1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2.000
Median :17.71	Median :0.0000	Median :0.0000	Median :4.000	Median :2.000
Mean :17.85	Mean :0.4375	Mean :0.4062	Mean :3.688	Mean :2.812
3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:4.000
Max. :22.90	Max. :1.0000	Max. :1.0000	Max. :5.000	Max. :8.000

Korelacja zmiennych wygląda następująco:



Obserwacje:

Istnieją duże negatywne korelacje pomiędzy:

a) spalaniem i ilością cylindrów,
b) spalaniem i pojemnością silnika,
c) spalaniem i wagą.

Istnieją duże pozytywne korelacje pomiędzy:

a) ilością cylindrów i pojemnością silnika
b) ilością cylindrów i mocą silnika
c) pojemnością silnika i wagą

Obróbka danych oraz zbiór testowy i treningowy

```
Multiple R-squared (forward):  0.8510283
Multiple R-squared (backward):  0.8475095
Multiple R-squared (both):  0.854394
```

Do jak najlepszego dopasowania danych do modelu użyto funkcji step, która automatycznie dobiera kolumny, tak aby regresja była jak najlepsza.

Dla danych mtcars funkcja step dobrała różne kolumny. Najlepiej wypadła dunkcja step z argumentem direction = „both”.

Cechy modelu dobrane automatycznie wyglądają następująco:

qsec ~ cyl + disp + wt + am + carb

Modele regresji

W pakiecie zostały zaimplementowane trzy modele regresji:

1) LM
2) SVM dla kerneli:
    e) Linear
    f) Polynomial
    g) Radial
    h) Sigmoid
3) GLM
    e) Gaussian
    f) Poisson
    g) Quasi
    h) Quasipoisson

Porównanie modeli i błędów RMSE dla danych MtCars

Model	**Family /Kernel**	**RMSE** Dane bez obróbki	**RMSE** dane po obróbce
LM	----	1.157114	1.050713
GLM	**gaussian**	1.157114	1.050713
	Gamma	16.60704	16.60762
	poisson	13.84408	13.84094
	quasi	1.157114	1.050713
	quasipoisson	13.84408	13.84094
SVM	**linear**	0.6156598	0.9736886
	polynomial	4.167068	4.424246
	radial	1.980707	1.91813
	sigmoid	1.471263	1.582125

Podsumowanie:

Najlepszy model z najniższym błędem RMSE dla danych MtCars to model SVM z funkcją kernelową liniową nauczony na danych bez obróbki. Błąd ten wynosi 0.6156598 w przewidywanym czasie na ¼ mili.

Najlepszym modelem GLM jest model z rodziny gaussian i quasi. Modele te zostały nauczone i przetestowane na obrobionych danych. Błąd RMSE dla nich wynosi 1.050713. W porównaniu do najlepszego modelu ze wszystkich jest to duża różnica. Trzema najgorszymi modelami dla tych danych są modele z rodziny gamma (RMSE = 16.60704 i 16.60704), quasipoisson (RMSE = 13.84408 i 13.84408) oraz Poisson (RMSE = 13.84408 i 13.84094)

Najlepszym modelem SVM jest model z funkcją kernelową liniową (RMSE = 0.6156598, 0.9736886). Najgorszym modelem SVM jest model z funkcji kernelowej polynomial (RMSE = 4.167068, 4.424246).

Irys

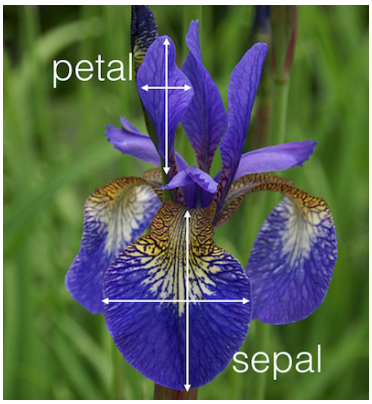
Opis danych

Zestaw pomiarów kwiatów irysa. Zazwyczaj zbioru tego używa się do wytrenowania systemu, który na podstawie 4 podanych parametrów, poda właściwą klasę kwiatu. Natomiast w naszym projekcie nie klasyfikujemy kwiatów, a staramy się przewidzieć za pomocą modeli regresji długość działki kielicha na podstawie: szerokości działki kielicha, długości płatka, szerokości płatka i klasy, w której znajduję się kwiat. Zbiór ten posiada 5 kolumn oraz 150 wierszy.

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Kolumnami w pliku są:

- „Sepal.Length” – długość działki kielicha (tą zmienną chcemy objaśniać)
- „Sepal.Width” – szerokość działki kielicha
- „Petal.Length” – długość płatka
- „Petal.Width” – szerokość płatka
- „Species” – klasa kwiatu



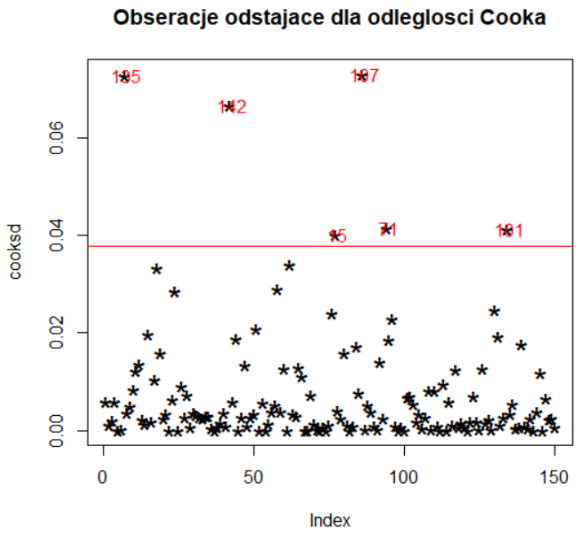
Podsumowanie danych:

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Korelacja zmiennych wygląda następująco:

Obserwacje: istnieje kilka silnych korelacji, jak widać długość działki kielicha jest silnie skorelowana z długością i szerokością płatka oraz klasą kwiatu.

Pozbycie się obserwacji odstających za pomocą odległości Cook'a



Modele regresji

W pakiecie zostały zaimplementowane trzy modele regresji:

4) LM
5) SVM dla kerneli:
   i) Linear
   j) Polynomial
   k) Radial
   l) Sigmoid
6) GLM
   i) Gaussian
   j) Poisson
   k) Quasi
   l) Quasipoisson

Porównanie modeli i błędów RMSE dla danych Iris

Model	**Family /Kernel**	**RMSE** dane bez obróbki	**RMSE** dane po obróbce
LM	----	0.2784828	0.2531975
GLM	**gaussian**	0.284817	0.1896056
	Gamma	5.853932	5.787837
	poisson	4.239627	4.201317
	quasi	0.284817	0.1896056
	quasipoisson	4.239627	4.201317
SVM	**linear**	0.2993042	0.2011303
	polynomial	0.3782467	0.2831165
	radial	0.3491633	0.3272615
	sigmoid	1.422121	1.019071

Podsumowanie:

Najlepszymi modelami z najniższym błędem RMSE dla danych Iris to modele GLM; family = gaussian oraz family = quasi. W obu przypadkach błąd RMSE wyniósł 0.1896056 w przewidywanej długości

działki kielicha dla irysa. Najgorszym modelem jest również model GLM w tym przypadku dla family = Gamma, którego błąd RMSE wyniósł 5.853932.

Dla modelu LM trochę mniejszy błąd RMSE wyszedł dla danych wcześniej przefiltrowanych i odrzuconych outlier'ów poprzez zastosowanie odległości Cook'a.

SVM uplasował się pomiędzy najlepszym wynikiem z GLM i LM, a jego błąd wyniósł odpowiednio 0.2011303 przypadek ten zaszedł dla funkcji kelnerowej liniowej na obrobionych danych.

Źródła:

- Wykłady dr inż. Robert Albert Kłopotek - https://rklopotek.blog.uksw.edu.pl/prowadzone-przedmioty/uczenie-maszynowe/
- „Przewodnik po pakiecie" – Przemysław Biecek
- „Analiza danych z programem R" – Przemysław Biecek
- http://r-statistics.co/Linear-Regression.html
- Grafika irysa - https://www.kaggle.com/biphili/seaborn-matplotlib-plot-to-visualize-iris-data