Regression: Animal Species Sleeping Hours

Guidance teacher : Liu Haixia

杨 准 , 黄 子 玲 and 李 沛

华中科技大学——数据挖掘

华中科技大学

Introduction

To solve what might affect the sleep that an animal needs. We use given data, there are 62 species with several features including the average sleeping hours per day (sleep) but some values are missing(NA). Finally, the research results are displayed with graphs.

Methods

1. The lost NA is compensated by the mean value.
2. Divide data into training set and data set.
3. Use multiple linear regression to analyze the value.
4. Analyze data using logistic regression.
5. Compare linear regression data with logistic regression data.
6. Analyze data using MSE.
7. Analyze data using cross-validation.
8. Analyze data using ridge regression.
9. Compare the results obtained by various methods.

Find The NA

1. Change the size of data from 0.2 to 0.35 for a better prediction result.
2. Fill the NA with mean value for the complete data table.

	species	slowWaveSleep	dreamSleep	sleep	body	brain	life	gestation	predation	sleepExposure	danger
0	African_elephant	8.672917	1.972	10.644917	6654.000	5712.0	38.600000	645.0	3	5	3
1	African_giant_pouched_rat	6.300000	2.000	8.300000	1.000	6.6	4.500000	42.0	3	1	3
2	Arctic_Fox	8.672917	1.972	10.644917	3.385	44.5	14.000000	60.0	1	1	1
3	Arctic_ground_squirrel	8.672917	1.972	10.644917	0.920	5.7	19.877586	25.0	5	2	3
4	Asian_elephant	2.100000	1.800	3.900000	2547.000	4603.0	69.000000	624.0	3	5	4
...	...	...	...	...	...	...	...	...	...	...	...
57	Tree_hyrax	4.900000	0.500	5.400000	2.000	12.3	7.500000	200.0	3	1	3
58	Tree_shrew	13.200000	2.600	15.800000	0.104	2.5	2.300000	46.0	3	2	2
59	Vervet	9.700000	0.600	10.300000	4.190	58.0	24.000000	210.0	4	3	4
60	Water_opossum	12.800000	6.600	19.400000	3.500	3.9	3.000000	14.0	2	1	1
61	Yellow-bellied_marmot	8.672917	1.972	10.644917	4.050	17.0	13.000000	38.0	3	1	1

image1: complete data table

Linear Regression

The linear relationship between shoe predict and test was found by using the multiple linear regression medel.

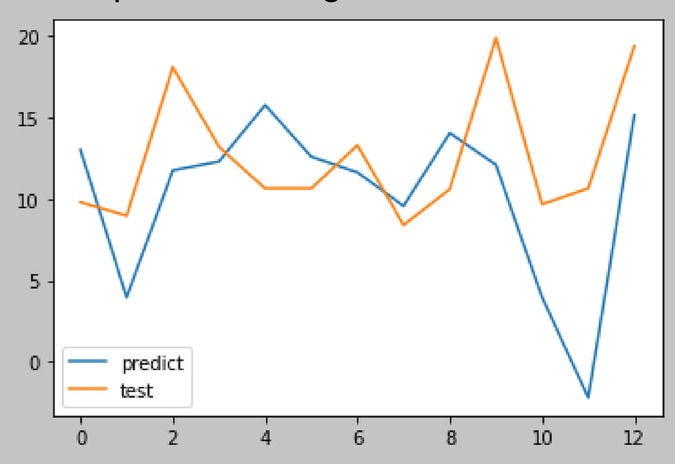

image2: shoe predict and test linear relationship

Logistic Regression

From linear regression, we get r2 score is -1.063924951753823.

From logistic regression, we get r2 score is -1.4193915578390914.

The data from logistic regression is smaller than linear regression so we choose to use linear regression to support our result.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=0).fit(x_train, y_train.astype('int'))
y_pred_lr = model.predict(x_test)
r2_score(y_test, y_pred_lr)
```

image3: process from logistic regression

MSE

We use the MSE method to assist the relationship obtained by linear regression for reduce error.

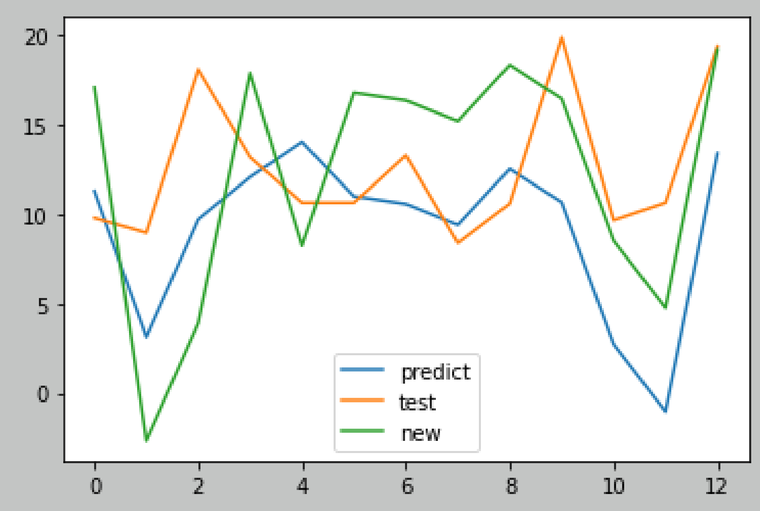

image4: the relationship with MSE prediction

Cross-validation

Use cross-validation to remove outliers and select the optimal r2 score.

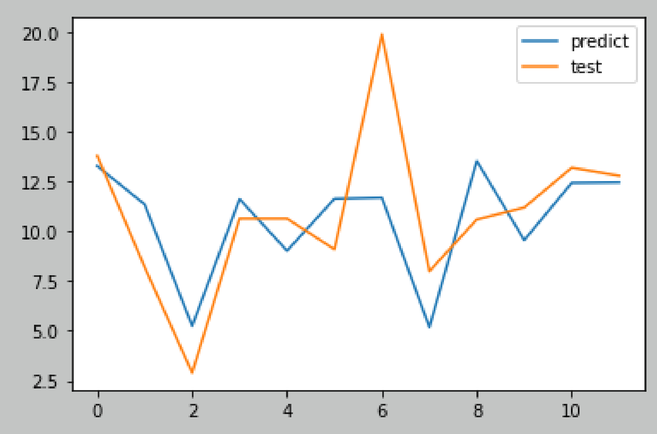

image5: shoe predict and test cross-validation relationship

Ridege Regression

Ridge regression was used to observe significant changes in the data.

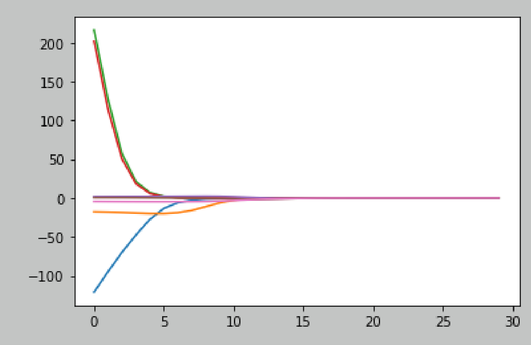

image6: ridege weights

Comparison And Analysis

We tried to analyze data using both linear regression and logistic regression, and found that this data is more suitable for the linear regression method. the methods MSE and cross-validation are used to reduce the error and make the final value closer to the real value.

According to the MSE graph obtained at last, the reasons that affect the sleep of animals are complex. 'x' is the reason and 'y' is the result, when 'x' is 1, the value of 'y' is the smallest, and when'x' is 12, the value of 'y' is the largest. Most of reasons have a positive impact on the result.

Division of contribution

Yang Zhun:	33.33%
Huang Ziling:	33.33%
Li Pei:	33.33%