Knowledge Discovery and Data Mining

Lab 5 Linear Regression

Xuan Song
Songx@sustech.edu.cn

Topics

Implement linear regression based on scikit-learn

Scikit-learn

scikit-learn

Machine Learning in Python

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

https://scikit-learn.org/stable/

Implement Linear Regression based on scikit-learn

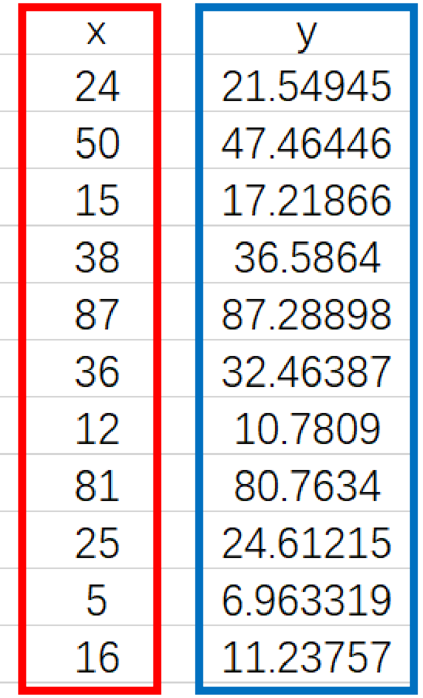
- Dataset:

explanatory variables

dependent variables

test.csv

train.csv



x	y
24	21.54945
50	47.46446
15	17.21866
38	36.5864
87	87.28898
36	32.46387
12	10.7809
81	80.7634
25	24.61215
5	6.963319
16	11.23757

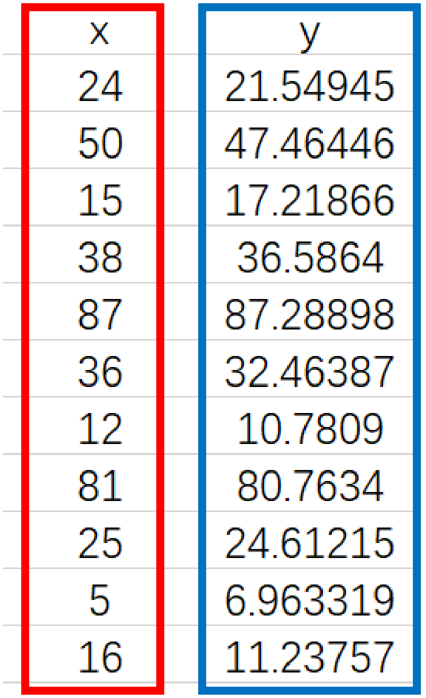
Implement Linear Regression based on scikit-learn

- Dataset:

explanatory variables   dependent variables

test.csv

train.csv



x	y
24	21.54945
50	47.46446
15	17.21866
38	36.5864
87	87.28898
36	32.46387
12	10.7809
81	80.7634
25	24.61215
5	6.963319
16	11.23757

$$y = wx + b$$

where

x: explanatory variable (feature)

y: dependent variable

w: slope coefficient for explanatory variable

b: y−intercept

Implement Linear Regression based on scikit-learn

- 1. Load training data and test data from csv files

- 2. Data cleaning

  pandas.Dataframe.dropna()

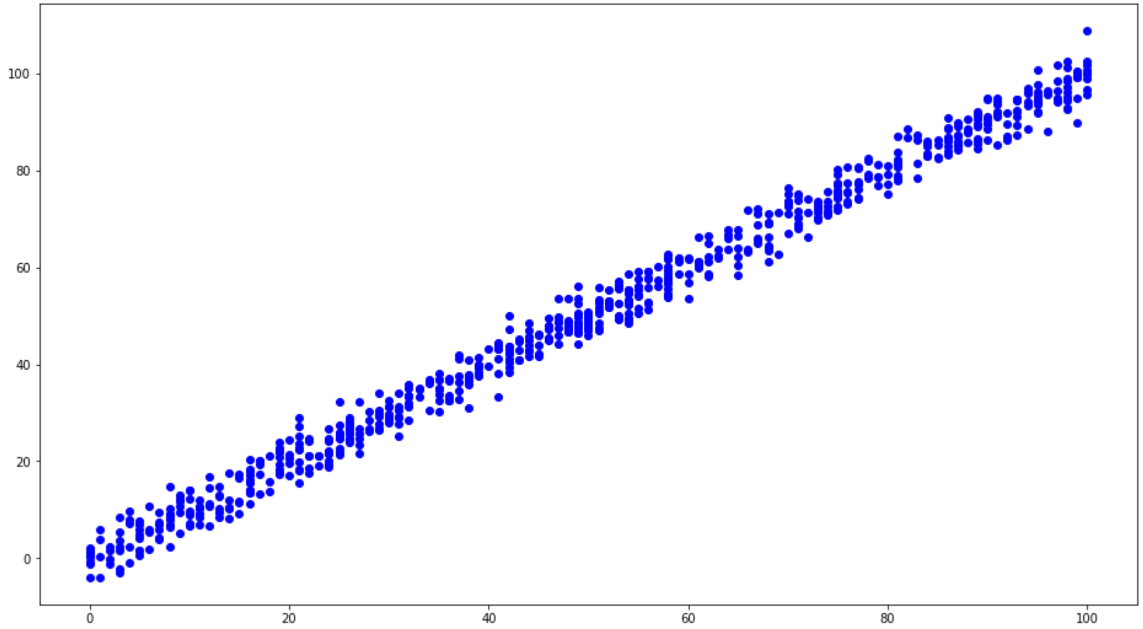
- 3. Get explanatory variables and dependent variables from training data and test data.

  For example, you can use **X_train = df.iloc[:, :-1].values.reshape(-1,1)** to obtain the independent variables of training data.

Implement Linear Regression based on scikit-learn

- 4. Visualize training data to further understand data

```
import matplotlib.pyplot as plt
plt.figure(figsize=(16,9))
plt.scatter(X_train, Y_train, color='blue')
plt.show()
```



If you have any problems about the functions of matplotlib,
you could refer to the following link: https://matplotlib.org/3.3.2/api/pyplot_summary.html

Implement Linear Regression based on scikit-learn

- 5. Build a linear regression model based on scikit-learn library.

```
from sklearn.linear_model import LinearRegression
LR_Model = LinearRegression()
```

- 6. Data fitting

```
LR_Model.fit(X_train, Y_train)

print(LR_Model.intercept_)
print(LR_Model.coef_)
```

w = 1.00065638
b =-0.10726546

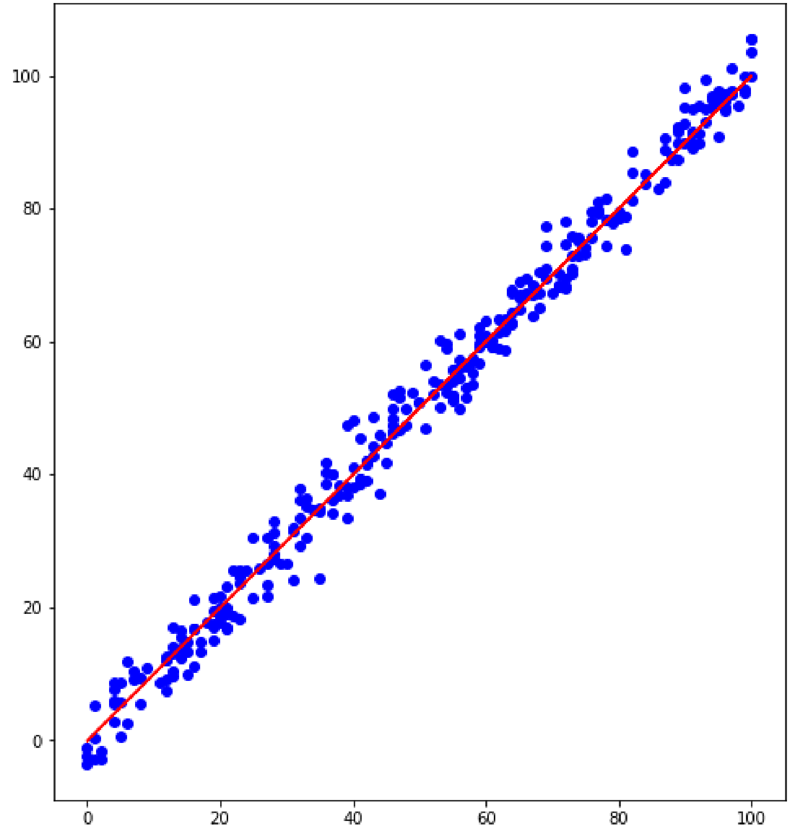
Y= 1.00065638 X -0.10726546

Implement Linear Regression based on scikit-learn

- 7. Visualization

You can use **plt.plot(x, y)** to plot the function $Y = 1.00065638\,X - 0.10726546$.

.



https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.plot.html

Implement Linear Regression based on scikit-learn

- 8. Evaluation

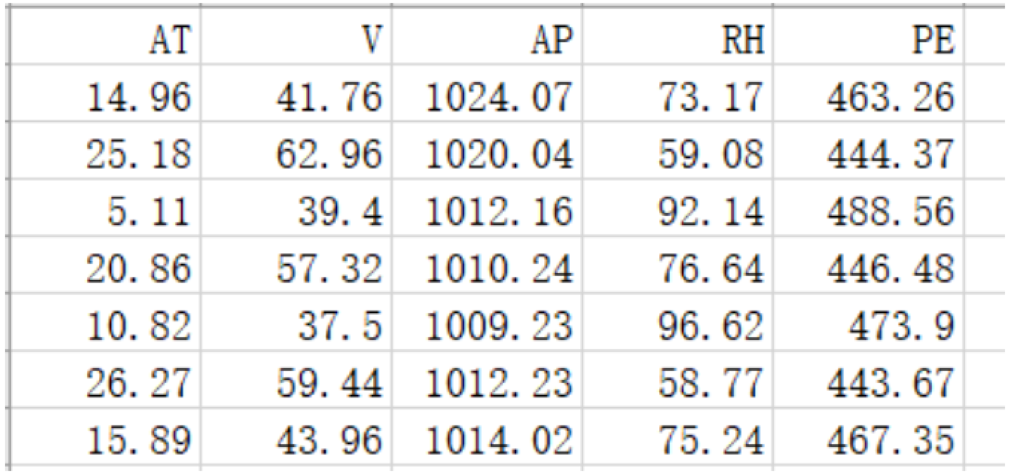
```
from sklearn import metrics

Y_pred = LR_Model.predict(X_test)

# 用scikit-learn计算MSE
print("MSE  ",metrics.mean_squared_error(Y_test, Y_pred))

# 用scikit-learn计算RMSE
print("RMSE ",np.sqrt(metrics.mean_squared_error(Y_test, Y_pred)))
```

```
MSE   9.43292219203932
RMSE 3.0713062680298298
```

Class Work: Implement linear regression based on a given data set

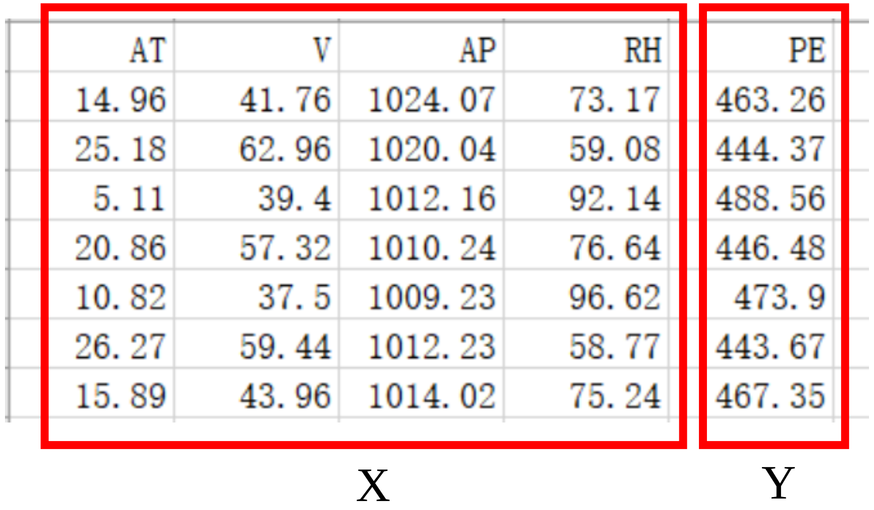
CCPP.csv

- Data name: Combined Cycle Power Plant Data Set
- Data description:
  - The dataset contains 9568 data points collected from a Combined Cycle Power Plant, when the power plant was set to work with full load.
  - Features consist of hourly average ambient variables Temperature (**T**), Ambient Pressure (**AP**), Relative Humidity (**RH**) and Exhaust Vacuum (**V**) to predict the net hourly electrical energy output (**PE**) of the plant



AT	V	AP	RH	PE
14.96	41.76	1024.07	73.17	463.26
25.18	62.96	1020.04	59.08	444.37
5.11	39.4	1012.16	92.14	488.56
20.86	57.32	1010.24	76.64	446.48
10.82	37.5	1009.23	96.62	473.9
26.27	59.44	1012.23	58.77	443.67
15.89	43.96	1014.02	75.24	467.35

Task: Implement linear regression based on a given data set



AT	V	AP	RH	PE
14.96	41.76	1024.07	73.17	463.26
25.18	62.96	1020.04	59.08	444.37
5.11	39.4	1012.16	92.14	488.56
20.86	57.32	1010.24	76.64	446.48
10.82	37.5	1009.23	96.62	473.9
26.27	59.44	1012.23	58.77	443.67
15.89	43.96	1014.02	75.24	467.35

$$\mathrm{PE} = w_1 * \mathrm{AT} + w_2 * \mathrm{V} + w_3 * \mathrm{AP} + w_4 * \mathrm{RH} + b$$

Where

w_1, w_2, w_3, w_4: slope coefficients for each explanatory variable

b : y-intercept

Task

- Implement linear regression based on a given data set

Functions you may need:

(1) sklearn.model_selection.train_test_split(X,Y, test_size=0.2, shuffle=False)

Other Resources

- Data Visualization:
  - https://matplotlib.org/3.3.2/api/pyplot_summary.html
  - https://matplotlib.org/3.3.2/api/_as_gen/matplotlib.pyplot.xticks.html#matplotlib.pyplot.xticks
- sklearn.linear_model.LinearRegression
  - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

南方科技大学

End of Lab5