

BÀI TUẦN 02: HỒI QUY TUYẾN TÍNH ĐƠN BIẾN

(SIMPLE LINEAR REGRESSION)

1.Thông tin sinh viên

Họ tên: Dương Minh Lượng

MSSV: 18521071

LỚP: Học máy thống kê-DS102.L12.CNCL

2.Source

```
1. import pandas as pd #cho du lieu tu file
2. import numpy as np #xu li mang
3. import matplotlib.pyplot as plt #truc quan hoa du lieu
4. from sklearn.model_selection import train_test_split #phan
   chia du lieu
5. from sklearn.linear_model import LinearRegression
6. from sklearn.metrics import r2_score
7. #Tien xu li du lieu
8. dataset =pd.read_csv("Salary_Data.csv")
9. X=np.array(dataset.iloc[:, :-1].values)
10. Y=np.array(dataset.iloc[:, 1].values)
11. X_train,X_test,Y_train,Y_test=train_test_split(X,Y,s
   huffle=True ,train_size=0.8, random_state=0)
12. reg=LinearRegression()
13. reg.fit(X_train, Y_train)
14. Y_train_pred=reg.predict(X_train)
15. Y_test_pred=reg.predict(X_test)
16. #hoan thien
17. plt.scatter(X_test, Y_test, color = 'red')
18. plt.scatter(X_test, Y_test_pred, color="BLACK")
19. plt.plot(X_train,Y_train_pred, color="BLUE")
20. plt.title('Salary vs Experience (Test set)')
21. plt.xlabel('Years of Experience')
22. plt.ylabel('Salary')
23. plt.show()
24. def compare(i_example):
25. x=X_test[i_example : i_example+1]
26. y=Y_test[i_example]
27. y_pred=reg.predict(x)
28. print(x,y,y_pred)
29. for i in range(len(X_test)):
30. compare(i)
31. #Đánh giá mô hình
32. print('(R)2 train = ',reg.score(X_train, Y_train))
33. print('(R)2 test = ',reg.score(X_test, Y_test))
34. r2=r2_score(Y_test,Y_test_pred)
```

```

35.     print('(R)2 test = ', r2)
36.     if ((reg.score(X_train,
    Y_train)>=0.8)&(reg.score(X_test, Y_test)>=0.8)):
37.         print('Mô hình tốt')
38.     elif ((reg.score(X_train,
    Y_train)==1)&(reg.score(X_test, Y_test)==1)):
39.         print('Mô hình cơ sở')
40.     else:
41.         print('Cần xem lại')

```

3. Kết quả



```

In [18]: runfile('F:/MAY HOC/THBuổi2/BaiThucHanh.py', wdir='F:/MAY HOC/THBuổi2')
[[1.5]] 37731.0 [40748.96184072]
[[10.3]] 122391.0 [122699.62295594]
[[4.1]] 57081.0 [64961.65717022]
[[3.9]] 63218.0 [63099.14214487]
[[9.5]] 116969.0 [115249.56285456]
[[8.7]] 109431.0 [107799.50275317]
(R)2 train = 0.9411949620562126
(R)2 test = 0.988169515729126
(R)2 test = 0.988169515729126
Mô hình tốt

```

```

In [19]:

```

Nhận xét:

- Khi dùng với việc lấy kết quả ngẫu nhiên thì $R > 0.8$ -> Mô hình tốt
- Các Y_{test} thực tế và Y_{test_pred} dự đoán gần nhau các điểm gần nhau .
- Nếu mà ta dùng với việc `shuffle=False` Thì mô hình không được tốt vì do dữ liệu xếp năm kinh nghiệm theo lương tăng dần hình minh họa dưới đây là R

```
Console 1/A x
In [13]: runfile('F:/MAY HỌC/THBuổi2/BaiThucHanh.py', wdir='F:/MAY HỌC/THBuổi2')
[[8.7]] 109431.0 [111119.08832991]
[[9.]] 105582.0 [114134.92418014]
[[9.5]] 116969.0 [119161.31726387]
[[9.6]] 112635.0 [120166.59588062]
[[10.3]] 122391.0 [127203.54619784]
[[10.5]] 121872.0 [129214.10343133]
(R)2 train = 0.9179154343152582
(R)2 test = 0.07028895951395653
(R)2 test = 0.07028895951395653
Cần xem lại
```