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

```
import pandas as pd  #cho du lieu tu file
import numpy as np   #xu li mang
import matplotlib.pyplot as plt #truc quan hóa dữ liệu
from sklearn.model_selection import train_test_split #phân
chia dữ liệu
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
#Tiền xử lí dữ liệu
dataset =pd.read_csv("Salary_Data.csv")
X=np.array(dataset.iloc[:,:-1].values)
Y=np.array(dataset.iloc[:,1].values)
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,train_size=0
.8, random_state=0)
reg=LinearRegression()
reg.fit(X_train, Y_train)
Y_train_pred=reg.predict(X_train)
Y_test_pred=reg.predict(X_test)
#hoàn thiện
plt.scatter(X_test, Y_test, color = 'red')
plt.scatter(X_test, Y_test_pred, color="BLACK")
plt.plot(X_train,Y_train_pred, color="BLUE")
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
def compare(i_example):
     x=X_test[i_example : i_example+1]
     y=Y_test[i_example]
     y_pred=reg.predict(x)
     print(x,y,y_pred)
for i in range(len(X_test)):
     compare(i)
#Đánh giá mô hình
print('(R)2 train = ',reg.score(X_train, Y_train))
print('(R)2 test = ',reg.score(X_test, Y_test))
r2=r2_score(Y_test,Y_test_pred)
```

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

3. Kết quả

```
In [18]: runfile('F:/MAY HỌC/THBuổi2/BaiThucHanh.py', wdir='F:/MAY HỌC/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]:
```

IPython console History

Nhận xét:

- Khi dùng với việc lấy kết quả ngâu 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 hoa dưới đây là R

Console 1/A

```
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
```

IPython console History