

BÀI TUẦN 02: SIMPLE LINEAR REGRESSION

1. Thông tin sinh viên

DƯƠNG MINH LƯỢNG-18521071

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. dataset1 =pd.read_csv("Salary_Data_Test.csv")
10. X=np.array(dataset.iloc[:, :-1].values)
11. Y=np.array(dataset.iloc[:, 1].values)
12. X1_test=np.array(dataset1.iloc[:, :-1].values)
13. Y1_test=np.array(dataset1.iloc[:, 1].values)
14. X_train,X_test,Y_train,Y_test=train_test_split(X,Y,t
   rain_size=0.5, random_state=0)
15. reg=LinearRegression()
16. reg.fit(X_train, Y_train)
17. Y_train_pred=reg.predict(X_train)
18. Y_test_pred=reg.predict(X1_test)
19. #hoan thien
20. plt.scatter(X1_test, Y1_test, color = 'red')
21. plt.scatter(X1_test, Y_test_pred, color="BLACK")
22. plt.plot(X_train,Y_train_pred, color="BLUE")
23. plt.title('Salary vs Experience (Test set)')
24. plt.xlabel('Years of Experience')
25. plt.ylabel('Salary')
26. plt.show()
27. def compare(i_example):
28.     x=X1_test[i_example : i_example+1]
29.     y=Y1_test[i_example]
30.     y_pred=reg.predict(x)
31.     print(x,y,y_pred)
32. for i in range(len(X1_test)):
33.     compare(i)
34. #Đánh giá mô hình
35. from sklearn.metrics import mean_squared_error
36. print(mean_squared_error(Y1_test, Y_test_pred))
```

```
37. print('(R)2 test = ',reg.score(X1_test, Y1_test))
38. r2=r2_score(Y1_test,Y_test_pred)
39. print('(R)2 test = ',r2)
40.
41.
42.
43.
```

3. Kết quả

0.5



0.6



0.7



0.8



0.9



	R²	MSE (Mean Square Error)
0.5	0.9192840479224546	43785512.66922273
0.6	0.9261041740558313	40085838.54617234
0.7	0.9237126514543206	41383153.89056159
0.8	0.9215791609988908	42540496.03869277
0.9	0.9228800977916074	41834784.429609336

Nhận xét:

- Ta thấy với việc chia 60% của 30 điểm dữ liệu cho Train thì dữ liệu test cho kết quả tốt nhất.
- Với việc chia 50% của 30 điểm dữ liệu cho Train thì lỗi sẽ cao nhất với dữ liệu test.
- Ta cần tìm MSE (min) or tối thiểu để tăng độ chính xác.
- MSE ít cost function thì parameter phải tốt nhất.
- Với bộ dữ liệu test cố định này thì kết quả giữa các cách chia train có sự biến đổi nhưng việc biến đổi như thế có thể làm độ chính xác giảm xuống. Nếu ta lấy bộ test khác có thể kết quả cũng sẽ khác.