Thực hành Nguyên Lý Máy Học
Buổi 4: hồi quy tuyến tính

Mục tiêu:

- Củng cố lý thuyết bài toán hồi quy tuyến tính
- Cài đặt giải thuật hồi quy tuyến tính bằng phương pháp giảm gradient.
- Kiểm thử và đánh giá

A. HƯỚNG DẪN THỰC HÀNH

1. **Ví dụ dự đoán giá nhà (bài tập ví dụ trên lớp)**
   Cho tập dữ liệu gồm 3 phần tử như bảng bên dưới,

X	1	2	4
Y	2	3	6

   Anh/chị hãy thực hiện các yêu cầu sau:
   - Biểu diễn tập dữ liệu lên mặt phẳng toạ độ Oxy
   - Tìm hàm hồi quy h(x) với giá trị khởi tạo theta0=0, theta1=1, tốc độ học: 0.2, số bước lặp: 2
   - Vẽ đường hồi quy lên mặt phẳng toạ độ
   - Dự đoán giá trị y cho các phần tử có x có giá trị lần lượt là 0, 3, 5

Hướng dẫn

a. Biểu diễn dữ liệu lên mặt phẳng toạ độ

```
import numpy as np
import matplotlib.pyplot as plt

X = np.array([1,2,4])
Y = np.array([2,3,6])

plt.axis([0,5,0,8])
plt.plot(X,Y,"ro",color="blue")
plt.xlabel("Gia tri thuoc tinh X")
plt.ylabel("Gia tri du doan Y")
plt.show()
```

b. Tìm hàm hồi quy với theta0 = 0, theta1 =1, tốc độ học = 0.2, số lần lặp là 1

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

}

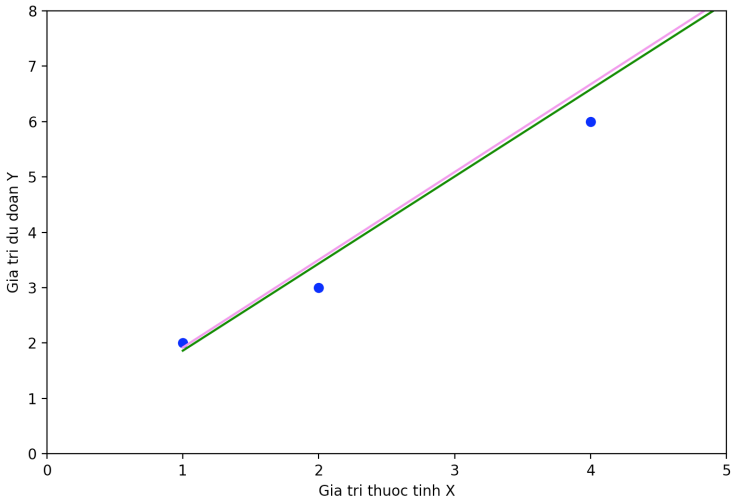
```
def LR1(X,Y,eta,lanlap, theta0,theta1):
    m = len(X) # so luong phan tu
    theta00 = theta0
    theta11 = theta1
    for i in range(0,lanlap):
        print("Lan lap: ", i)
        for j in range(0,m):
            #theta0
            h= theta0 + theta11*X[j]
            theta0 = theta0 + eta*(Y[j]-h)*1
            print ("Phan tu ", j, "y=", Y[j], "h=",h,"gia tri theta0 = ",theta0)
            #theta1
            h= theta00 + theta1*X[j]
            theta1 = theta1 + eta*(Y[j]-h)*X[j]
            print ("Phan tu ", j, "gia tri theta1 = ",theta1)
            theta00= theta0
            theta11= theta1
        print ("theta00 = ", theta00)
        print ("theta11 = ", theta11)
    return [theta0,theta1]

theta = LR1(X,Y,0.2,1,0,1)
theta
```

Kết quả (kiểm tra trong slide):

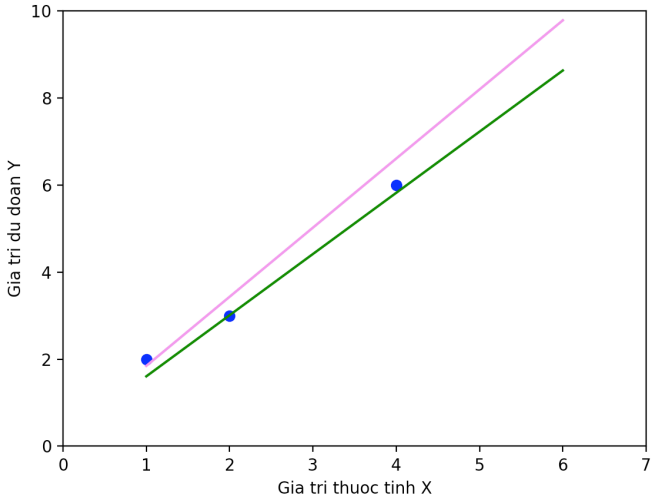
```
theta
        [0.3360000000000001, 1.584]
```

c. Vẽ đường hồi quy

```
theta = LR1(X,Y,0.2,1,0,1) # theta 1 bước
X1= np.array([1,6])
Y1= theta[0] + theta[1]*X1

theta2 = LR1(X,Y,0.2,2,0,1) # theta 2 bước lap
X2= np.array([1,6])
Y2= theta2[0] + theta2[1]*X2

plt.axis([0,7,0,10])
plt.plot(X,Y,"ro",color="blue")

plt.plot(X1,Y1,color="violet") # duong hoi quy lan lap 1
plt.plot(X2,Y2,color="green")  # duong hoi quy lan lap 2

plt.xlabel("Gia tri thuoc tinh X")
plt.ylabel("Gia tri du doan Y")
plt.show()
```



d. Thay đổi tốc độ học =0.1, số lần lặp bằng 2, anh/chị vẽ lại đường hồi quy cho bước lặp thứ 1 (màu hồng) và bước lặp thứ 2 (màu xanh)



e. Dự báo cho phần tử mới tới
Dự báo giá trị y cho 3 phần tử sau: x=0, x=3, x=5

```
# Du bao
y1 = theta[0] + theta[1]*0
y2 = theta[0] + theta[1]*3
y3 = theta[0] + theta[1]*5
```

Hoặc sử dụng vòng lặp for

```
# Du bao
XX = [0,3,5]
for i in range(0,3):
    YY = theta[0] + theta[1]*XX[i]
    print round(YY,3)
```

Kết quả kiểm tra:

```
0.336
5.088
8.256
```

2. Sử dụng thư viện scikit-learn của Python để tìm nghiệm

```
# đọc dữ liệu từ file Housing.csv
import pandas as pd
dt = pd.read_csv("Housing_2019.csv", index_col=0)
dt.ix[2:4,]
X= dt.ix[:,(1,2,3,4,10)]
X.ix[1:5,]
Y = dt.price

# huấn luyện mô hình
import sklearn
from sklearn import linear_model
lm = linear_model.LinearRegression()
lm.fit(X[1:520],Y[1:520])

print lm.intercept_
print lm.coef_

# dự báo giá nhà cho 20 phần tử cuối cùng trong tập dữ liệu
Y = dt.price
Y_test = Y[-20:]
X_test = X[-20:]
Y_pred = lm.predict(X_test)
```

Các câu hỏi cần trả lời:

1. Thuộc tính nào trong tập dữ liệu được sử dụng để dự báo giá nhà?
2. Cho biết có bao nhiêu “theta” và giá trị “theta” tương ứng?
3. Dữ liệu sử dụng để huấn luyện mô hình
4. Giá nhà dự báo cho 20 phần tử cuối cùng trong tập dữ liệu là bao nhiêu?
5. Giá nhà thực tế của 20 phần tử cuối cùng trong tập dữ liệu là bao nhiêu?