Homework

September 2, 2021

1 Đề bài về nhà

1.1 Yêu cầu

- Tự viết code cho mô hình Linear Regression theo công thức đã được dạy trong buổi lý thuyết trên lớp.
- Tự viết hàm dự đoán.
- Huấn luyện cả mô hình của thư viện và mô hình mình tự viết.
- In ra các trọng số: w0, w1, w2, ..., wn của cả 2 mô hình đã huấn luyện để quan sát và so sánh.
- Dự đoán dữ liệu tập test bằng cả 2 mô hình (mô hình thư viện thì dùng hàm predict() của thư viện, mô hình tự viết dùng hàm dự đoán tự viết), in ra kết quả bằng Dataframe như trong bài thực hành trên lớp.
- Tính RMSE trên tập test cho cả 2 mô hình và so sánh.

1.2 Dữ liệu

Tập dữ liệu giá nhà ở Boston đã có sẵn trên sklearn, dữ liệu đã được chuẩn hóa và chia thành tập train, tập test

[17]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import math

from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score
```

2 Đọc dữ liệu

Dữ liệu về giá nhà ở Boston được hỗ trợ bởi sklearn, đọc dữ liệu thông qua hàm `datasets.load_boston()`

Xem thêm các bộ dữ liệu khác tại https://scikit-learn.org/stable/datasets/index.html#toy-datasets.

Dữ liệu được chia thành các thành phần data và target như tập diabetes. Dữ liệu cũng đã được chuẩn hóa, chỉ cần gọi ra và huấn luyện

[32]:
```
# lay du lieu dataset - du lieu ve giá nhà
dataset = datasets.load_boston()
print("Số chiều dữ liệu input: ", dataset.data.shape)
print("Số chiều dữ liệu target: ", dataset.target.shape)
print()

print("5 mẫu dữ liệu đầu tiên:")
print("input: ", dataset.data[:5])
print("target: ",dataset.target[:5])
```

```
Số chiều dữ liệu input:  (506, 13)
Số chiều dữ liệu target:  (506,)

5 mẫu dữ liệu đầu tiên:
input:  [[6.3200e-03 1.8000e+01 2.3100e+00 0.0000e+00 5.3800e-01 6.5750e+00
  6.5200e+01 4.0900e+00 1.0000e+00 2.9600e+02 1.5300e+01 3.9690e+02
  4.9800e+00]
 [2.7310e-02 0.0000e+00 7.0700e+00 0.0000e+00 4.6900e-01 6.4210e+00
  7.8900e+01 4.9671e+00 2.0000e+00 2.4200e+02 1.7800e+01 3.9690e+02
  9.1400e+00]
 [2.7290e-02 0.0000e+00 7.0700e+00 0.0000e+00 4.6900e-01 7.1850e+00
  6.1100e+01 4.9671e+00 2.0000e+00 2.4200e+02 1.7800e+01 3.9283e+02
  4.0300e+00]
 [3.2370e-02 0.0000e+00 2.1800e+00 0.0000e+00 4.5800e-01 6.9980e+00
  4.5800e+01 6.0622e+00 3.0000e+00 2.2200e+02 1.8700e+01 3.9463e+02
  2.9400e+00]
 [6.9050e-02 0.0000e+00 2.1800e+00 0.0000e+00 4.5800e-01 7.1470e+00
  5.4200e+01 6.0622e+00 3.0000e+00 2.2200e+02 1.8700e+01 3.9690e+02
  5.3300e+00]]
target:  [24.  21.6 34.7 33.4 36.2]
```

Chia dữ liệu làm 2 phần training 362 mẫu và testing 80 mẫu

[19]:
```
# cat nho du lieu, lay 1 phan cho qua trinh thu nghiem,
# chia train test cac mau du lieu
# dataset_X = dataset.data[:, np.newaxis, 2]
dataset_X = dataset.data

dataset_X_train = dataset_X[:404]
dataset_y_train = dataset.target[:404]

dataset_X_test = dataset_X[405:]
dataset_y_test = dataset.target[405:]
```

[ ]: