TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CNTT

----- ๛ 🕮 ๛ -----

ĐỒ ÁN MÔN HỌC III

Đề tài: Linear Regression

Nhóm và sinh viên thực hiện:

Họ và tên: Phan Nguyễn Phước Nguyên

MSSV: 20127577

MỤC LỤC

MỞ ĐẦU

Dữ liệu tuổi thọ trung bình được thu thập từ tổ chức WHO và trang web United Nations từ năm 2000 đến 2015 trên tất cả quốc gia.

Gồm:

1180 dòng dữ liệu

11 cột dữ liệu gồm:

1 giá trị mục tiêu (y): Life expectancy

10 đặc trưng giải thích (X) (đặc trưng giúp tìm giá trị mục tiêu) gồm: Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19, Thinness age 5-9, Income composition of resources, Schooling, Life expectancy

CHƯƠNG I. CÁC THƯ VIỆN SỬ DỤNG

a)

import pandas as pd : Hàm chuyên biệt cho xử lý dataframe, đọc file csv và biến đổi dữ liệu

import numpy as np : Hàm cơ bản xử lý ma trận và các vấn đề về toán

b)

from sklearn.model_selection import Kfold : Hàm giúp chia dữ liệu thành k phần ở đây dùng k = 5 với 4 phần train và 1 phần test sau đó di chuyển lần lượt train/set sang các phần khác

c)

from sklearn.preprocessing import StandardScaler : Hàm chuẩn hóa dữ liệu để phục vụ cho việc biến đổi (cộng, trừ, nhân,… 2 features)

from sklearn.linear_model import LinearRegression : Kiểm tra đối chiếu với model train được và model có sẵn của thư viện

CHƯƠNG II. Ý TƯỞNG, TÊN HÀM

Đặt vấn đề

Dự đoán tuổi thọ trung bình là một đề tài phổ biến. Ở đồ án này ta được cung cấp 2 file dữ liệu là train.csv và test.csv

Trong đó mỗi file có thành phần gồm 10 features liên quan đến sự ảnh hưởng tuổi thọ và 1 label là tuổi thọ trung bình

Có 3 yêu cầu:

1. Dự đoán tuổi thọ trung bình từ 10 feature cho sẵn
2. Dự đoán tuổi thọ trung bình nhưng chỉ sử dụng từng feature lần lượt
3. Kết hợp hai hay nhiều feature và có thể biến đổi cho phù hợp để đạt kết quả với RMSE thấp nhất (sai số trung bình lấy căn bậc hai). Sau đó dùng model vừa tạo được dự đoán tuổi thọ trung bình

Các hàm sử dụng

a)

class OLSLinearRegression:

def fit : train model tu X_train

def predict : predict label sau khi fit

def rmse(predictions, targets): tính sai số trung bình lấy căn bậc hai từ y_predict và y_test

b)

def rmse_mean_cal(model, X, X_test, y, kfold):

train/test tren 5fold va tra ve gia trị trung bình của 5 lần

Đầu tiền phần train và test sẽ được concat lại sau đó sử dụng hàm Kfold để chia dữ liệu thành 5 phần rồi lần lượt thực hiện train ở 4 phần và test ở phần còn lại. Lần lượt hết 5 phần thì dừng và trả về trung bình rmse của 5 lần train

c)

scale= StandardScaler()

X2_train = scale.fit_transform(X2_train)

X2_test = scale.fit_transform(X2_test)

Chuẩn hóa dữ liệu để kết hợp

convert back to dataframe

X2_train = pd.DataFrame(X2_train, columns = X_train.columns)

X2_test = pd.DataFrame(X2_test, columns = X_train.columns)

NHẬN XÉT

Life expectancy sẽ được tính bằng hồi quy tuyến tính với công thức

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{10} x_{10} + \varepsilon$. Trong đó $\beta_0, \beta_1, \ldots, \beta_r$ là hệ số tương quan và ε là sai số ngẫu nhiên

Bảng hệ số tương quan

	Life expectancy
Adult Mortality	-0.685290
BMI	0.610687
Polio	0.365705
Diphtheria	0.406798
HIV/AIDS	-0.586578
GDP	0.486761
Thinness age 10-19	-0.503652
Thinness age 5-9	-0.500231
Income composition of resources	0.779404
Schooling	0.754864
Life expectancy	1.000000

a)

Ở câu a ta sẽ dùng 10 feature đề cho để train và sau đó test trên file test.csv

Kết quả thu được

RMSE = 7.06404643058404

Vì 10 feature chưa được sàng lọc nên kết quả ở mức tương đối và có thể tối ưu

b)

STT	Mô hình với 1 đặc trưng	RMSE
1	Adult Mortality	45.62493038094085
2	BMI	26.21729921152169
3	Polio	17.98546983493596
4	Diphtheria	15.553270331802272
5	HIV/AIDS	69.02769971532624
6	GDP	59.711600292023334
7	Thinness age 10-19	53.44080927591617
8	Thinness age 5-9	52.90478883918289
9	Income composition of resources	9.463088664795228
10	Schooling	10.26194576513432

Đặc trưng tốt nhất: Income composition of resources : 9.463088664795228

Ở câu b ta xét riêng từng feature và nhìn vào bảng kết quả cho thấy một số feature có sức ảnh hưởng rất yếu trong khi số còn lại cao hơn rõ rệt, Vì vậy việc sàng lọc các feature là điều thiết yếu để tối ưu sai số.

Với kết quả trên HIV/AIDS có ảnh hưởng thấp đến tuổi trọ trung bình (RMSE = 69.03)

Và Income composition of resources (Thu nhập cá nhân) có sai số thấp chứng tỏ có liên hệ mật thiết đến tuổi thọ trung bình

RMSE = 9.462641625274049

c)

model1 lay ngau nhien 3 feature 0,5,6

model1 = [0, 5, 6]

Đây là model kết hợp ngẫu nhiên 3 feature nên kết quả mang tính tương đối tuy nhiên ở trường hợp này cho ra một kết quả khá tốt (RMSE = 4.69)

model2 tao 1 feature moi (NewCol10) bang tong 2 feature Income composition of resources va Schooling (sau khi chuan hoa)

model2 = [10]

Đây là model sử dụng thuộc tính thứ 11 được tạo ra bằng tổng của Income composition of resources va Schooling

Model này dựa trên các feature có biểu hiện tốt trên câu b để kết hợp lại với nhau tạo ra 1 feature mới có kết quả tốt hơn (RMSE = 4.532574940459634

)

model3 tao 1 feature moi (NewCol11) bang tong 2 feature 0.5GDP va Schooling

#sau do ket hop voi cac feature tu model 1

model3 = [0, 5, 6, 8, 11]

Model này có tinh chỉnh các thuộc tính bằng cách nhân với hệ số tự do để thay đổi mức ảnh hưởng lên kết quả đầu ra, mặt khác còn sử dụng kết hợp nhiều model với nhau nên đầu ra kết quả tương đối tốt. Đây là model tốt nhất được tạo ra và đặt tên là my_best_model

(RMSE = 4.42434132052356)

Sau đó sử dụng my_best_model để predict trên file test

Kết quả thu được (RMSE = 4.972446615817532)

TÀI LIỆU THAM KHẢO

[1] https://www.askpython.com/python/examples/standardize-data-in-python

[2] https://datatofish.com/numpy-array-to-pandas-dataframe/

[3] https://www.marsja.se/adding-new-columns-dataframe-pandas-examples/