علی قربان‌پور ـ ۹۶۱۰۵۹۹۴
تمرین سری سوم درس یادگیری ماشین ـ دکتر حسینی

مقدمه

در این تمرین می‌خواهیم با دیتاست load_boston از کتابخانه‌ی sklearn کار کنیم. این دیتاست اطلاعاتی از قیمت خانه‌های بوستون به همراه تعدادی ویژگی از آن مکان‌ها را در اختیار ما قرار می‌دهد. هدف از این تمرین پیاده‌سازی فرم بسته‌ی تابع تخمین رابطه‌ی قیمت خانه بر اساس سایر ویژگی‌های ارائه‌شده در دیتاست است. در ابتدا دیتاست را بررسی کرده و به بررسی ویژگی‌های ستون‌ها و رابطه‌ی کورلیشن ستون قیمت با سایر ستون‌ها می‌پردازیم. سپس به دنبال یافتن تخمینی از رابطه‌ی این قیمت با سایر ویژگی‌ها می‌رویم. سپس توابع پایه‌ی دیگری را نیز امتحات می‌کنیم و با کمک آن‌ها به تخمین تابع می‌پردازیم. در نهایت نیز خطاهای توابع مختلف بدست آمده را بر روی داده‌های آموزش و آزمایش می‌سنجیم و به عنوان خروجی نهایی گزارش می‌کنیم.

۱

در نخستین گام دیتاست را لود کرده و آن را به فرمت یک دیتافریم در می‌آوریم. سپس ستون هدف یا همان قیمت خانه را به دیتافریم بدست آمده اضافه می‌کنیم و بررسی می‌کنیم که آیا داده‌های Nan در ستون‌ها وجود دارد یا خیر که در این دیتاست داده‌ی نامشخص نداشتیم.

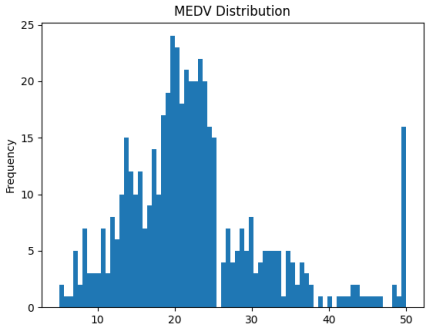
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.datasets import load_boston

dataset = load_boston()
df = pd.DataFrame(dataset.data, columns=dataset.feature_names)
df['MEDV'] = dataset.target
print(df.isnull().sum())
```

در مرحله‌ی بعد برای آنکه دید بهتری نسبت به داده‌ها داشته باشیم نمودار توزیع داده‌های قیمت را رسم می‌کنیم. سپس داده‌های پرت ستون قیمت را از مجموعه‌ی داده‌های موجود حذف می‌کنیم تا بتوانیم پیاده‌سازی بهتری در ادامه داشته باشیم. همانطور که در تصویر مشاهده می‌کنید تعدادی داده‌ی پرت در انتهای داده‌های قیمت وجود دارد که مقدارشان برابر ۵۰ و تعدادشان ۱۶ تا بود. با حذف این داده‌ها منحنی توزیع داده‌ها بتقریب نرمال خواهد بود که شروع خوبی برای کار با داده‌هاست :)

```
plt.hist(df['MEDV'], bins=75)
plt.gca().set(title='MEDV Distribution', ylabel='Frequency')
plt.show()

print(len(df[df['MEDV'] == df['MEDV'].max()]))
df = df[df['MEDV'] != df['MEDV'].max()]
```



در گام آخر با استفاده از توابع کتابخانه‌ی pandas داده‌های بدست آمده را به دو دسته‌ی داده‌های آموزش و آزمایش تقسیم می‌کنیم. این دو بخش را در فایل‌های جداگانه با پسوند csv. می‌نویسیم تا در بخش‌های بعد از آن‌ها استفاده کنیم. توجه کنید در تمام مراحل یادگیری به هیچ عنوان به سراغ داده‌های آزمایش نمی‌رویم. زیرا این کار می‌تواند باعث شود آموزش بایاس شود بر روی داده‌های آزمایش و بر روی آن‌ها خوب عمل کند. در حالی که بر روی داده‌هایی که تا به حال ندیده است خطای زیادی داشته باشد.

```
train = df.sample(frac=0.8, random_state=int(time.time()))
test = df.drop(train.index)
print(train.shape, test.shape)

train.to_csv(r'train.csv', index=False, header=True)
test.to_csv(r'test.csv', index=False, header=True)
```

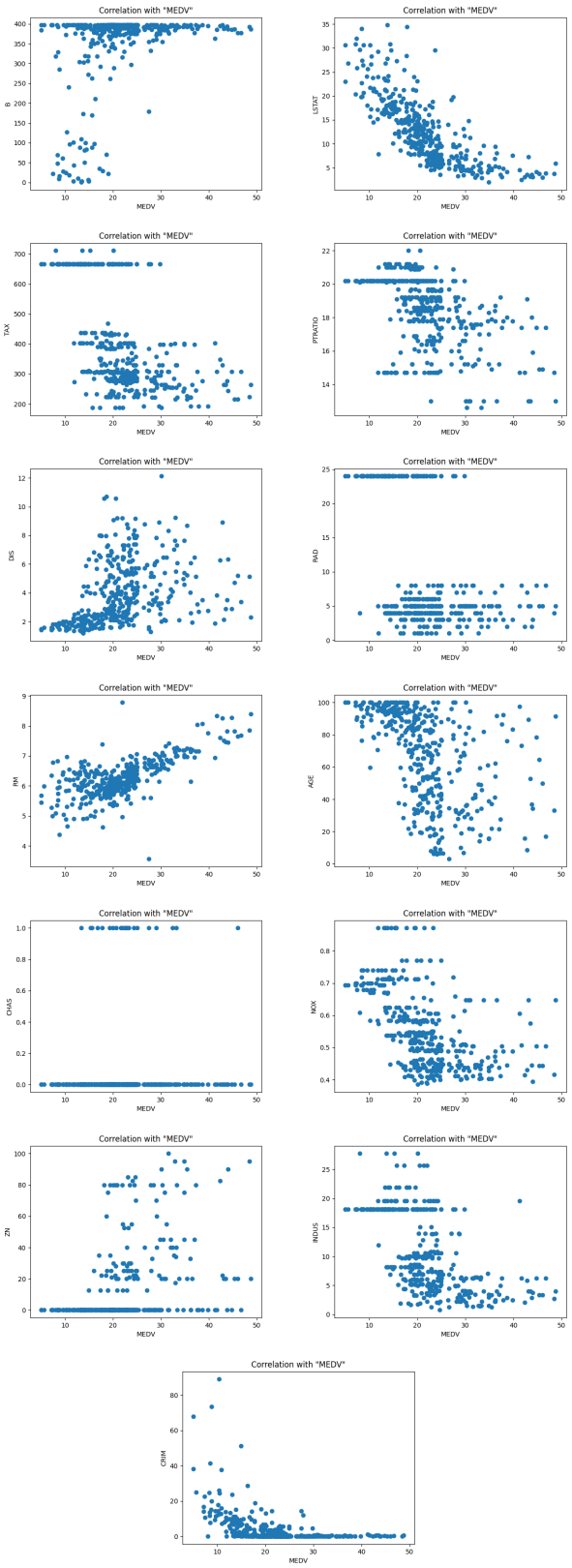
۲

در این بخش مقادیر ستون هدف یا همان MEDV را بر اساس سایر ستون‌ها رسم می‌کنیم تا روند تغییرات این ستون را نسبت به همه ستون‌ها مشاهده کنیم. همچنین کورلیشن دو به دوی ستون‌ها را نیز در خروجی زیر مشاهده می‌کنید. همانطور که از مقادیر کورلیشن در ستون MEDV مشاهده می‌کنید ستون‌های LSTAT, RM, INDUS به ترتیب بیشترین همبستگی را با توزیع ستون هدف دارند. این موارد به روشنی در نمودارهای زیر مشهود است.

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM	1.00	-0.20	0.41	-0.05	0.48	-0.17	0.35	-0.38	0.61	0.57	0.28	-0.36	0.44	-0.44
ZN	-0.20	1.00	-0.53	-0.06	-0.51	0.30	-0.55	0.66	-0.32	-0.31	-0.36	0.18	-0.43	0.40
INDUS	0.41	-0.53	1.00	0.04	0.77	-0.38	0.65	-0.72	0.63	0.73	0.38	-0.38	0.65	-0.59
CHAS	-0.05	-0.06	0.04	1.00	0.08	0.01	0.07	-0.08	-0.02	-0.06	-0.09	0.04	0.02	0.05
NOX	0.40	-0.51	0.77	0.08	1.00	-0.30	0.73	-0.77	0.61	0.67	0.16	-0.40	0.62	-0.51
RM	-0.17	0.30	-0.38	0.01	-0.30	1.00	-0.24	0.22	-0.14	-0.23	-0.28	0.08	-0.59	0.68
AGE	0.35	-0.55	0.65	0.07	0.73	-0.24	1.00	-0.74	0.47	0.51	0.25	-0.29	0.64	-0.48
DIS	-0.38	0.66	-0.72	-0.08	-0.77	0.22	-0.74	1.00	-0.50	-0.54	-0.22	0.32	-0.55	0.36
RAD	0.61	-0.32	0.63	-0.02	0.61	-0.14	0.47	-0.50	1.00	0.91	0.47	-0.47	0.51	-0.47
TAX	0.57	-0.31	0.73	-0.06	0.67	-0.23	0.51	-0.54	0.91	1.00	0.46	-0.46	0.56	-0.56
PTRATIO	0.28	-0.36	0.38	-0.09	0.16	-0.28	0.25	-0.22	0.47	0.46	1.00	-0.18	0.34	-0.51
B	-0.36	0.18	-0.38	0.04	-0.48	0.08	-0.29	0.32	-0.47	-0.46	-0.18	1.00	-0.37	0.37
LSTAT	0.44	-0.43	0.65	0.02	0.62	-0.59	0.64	-0.55	0.51	0.56	0.34	-0.37	1.00	-0.75
MEDV	-0.44	0.40	-0.59	0.05	-0.51	0.68	-0.48	0.36	-0.47	-0.56	-0.51	0.37	-0.75	1.00

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00	392.00
mean	3.69	11.84	10.89	0.05	0.55	6.26	66.85	3.94	9.52	406.54	18.51	354.47	12.66	21.75
std	9.14	23.71	6.77	0.23	0.12	0.64	28.28	2.17	8.71	168.89	2.08	96.94	6.84	7.91
min	0.01	0.00	0.74	0.00	0.38	3.56	2.90	1.17	1.00	187.00	12.60	0.32	1.98	5.00
25%	0.08	0.00	5.13	0.00	0.45	5.88	42.28	2.12	4.00	279.00	17.40	376.46	7.22	17.03
50%	0.23	0.00	8.56	0.00	0.52	6.19	74.85	3.38	5.00	330.00	19.10	392.17	11.64	20.95
75%	3.59	20.00	18.10	0.00	0.62	6.57	92.12	5.40	24.00	666.00	20.20	396.90	16.98	24.70
max	88.98	100.00	27.74	1.00	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	34.77	48.80

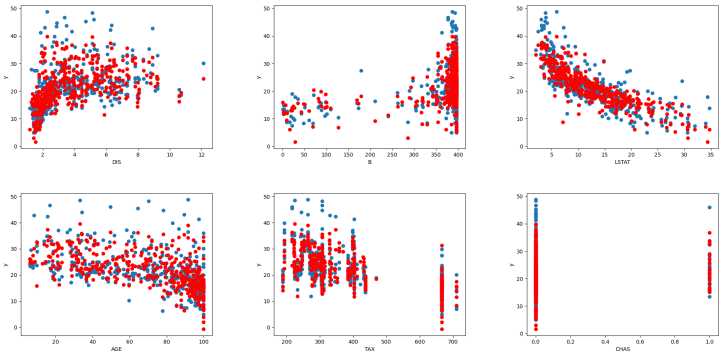
شکل ۱: مقادیر کورلیشن زوج‌های مرتب ستون‌ها.



شکل ۲: ستون هدف بر حسب سایر ستون‌ها.

۳

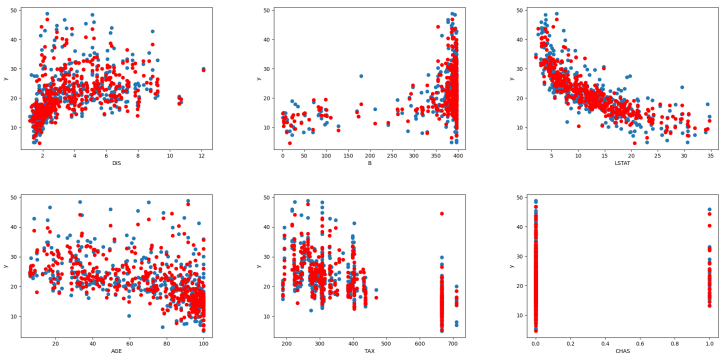
در این بخش می‌خواهیم با استفاده از ستون‌های در دست به تخمین مقادیر ستون هدف بپردازیم. ابتدا باید توجه کرد برای استفاده از ستون‌های مختلف به منظور تخمین، نباید بین ستون‌ها کورلیشن قابل توجهی وجود داشته باشد. بدین ترتیب با بررسی این مورد متوجه می‌شویم بین دو ستون RAD, TAX کورلیشن بالای ۹.۰ وجود دارد. به همین علت یکی از ستون‌ها را به انتخاب حذف می‌کنیم تا تخمین بهتری داشته باشیم. در نهایت به تخمینی بر حسب ستون‌های در دست می‌رسیم و این تخمین را در مقایسه با مقدار واقعی در یک نمودار مشخص نمایش می‌دهیم. این نمودارها بر حسب ستون‌های مختلف رسم شده‌اند.



شکل ۳: ستون هدف بر حسب سایر ستون‌ها با استفاده از فرم بسته.

۴

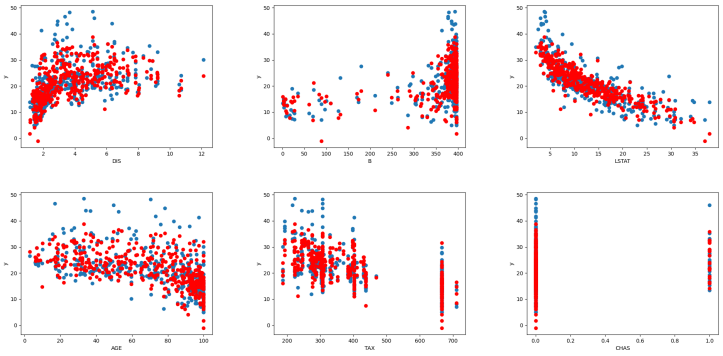
در این بخش به مانند بخش پیشین عمل می‌کنیم. تنها باید به ازای ستون‌های ماتریس ورودی یک ستون توان دوی آن را اضافه کنیم تا یادگیری با استفاده از مقادیر مطلق و درجه‌ی دوی هر ستون صورت گیرد. سپس با استفاده از همان فرم بسته‌ی معادله و افزودن ورودی های توان دو به ماتریس X پارامترهای بهینه را بدست آورده و خروجی را تخمین می‌زنیم. در ادامه خروجی تخمین توابع در تناظر با برخی ستون ها به همراه مقدار واقعی مشاهده می‌کنید.



شکل ۴: ستون هدف بر حسب سایر ستون‌ها با افزودن تابع درجه دو.

۵

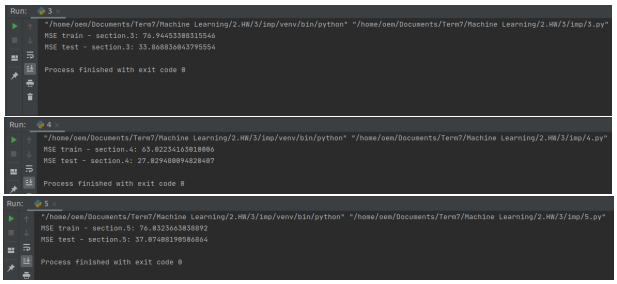
در این بخش مشابه بخش پیشین به افزودن ستون‌هایی به استفاده از فرمول گاوسی می‌پردازیم. بدین ترتیب ۱۰ ستون به ماتریس ورودی آموزش و آزمایش افزوده می‌شود و با ماتریس جدید افزوده شده به تخمین می‌پردازیم. مشابه بخش‌های پیشین به رسم نمودار خروجی تخمین زده شده در کنار خروجی مورد انتظار، بر حسب برخی از ستون‌ها می‌پردازیم.



شکل ۵: ستون هدف بر حسب سایر ستون‌ها با افزودن تابع درجه دو.

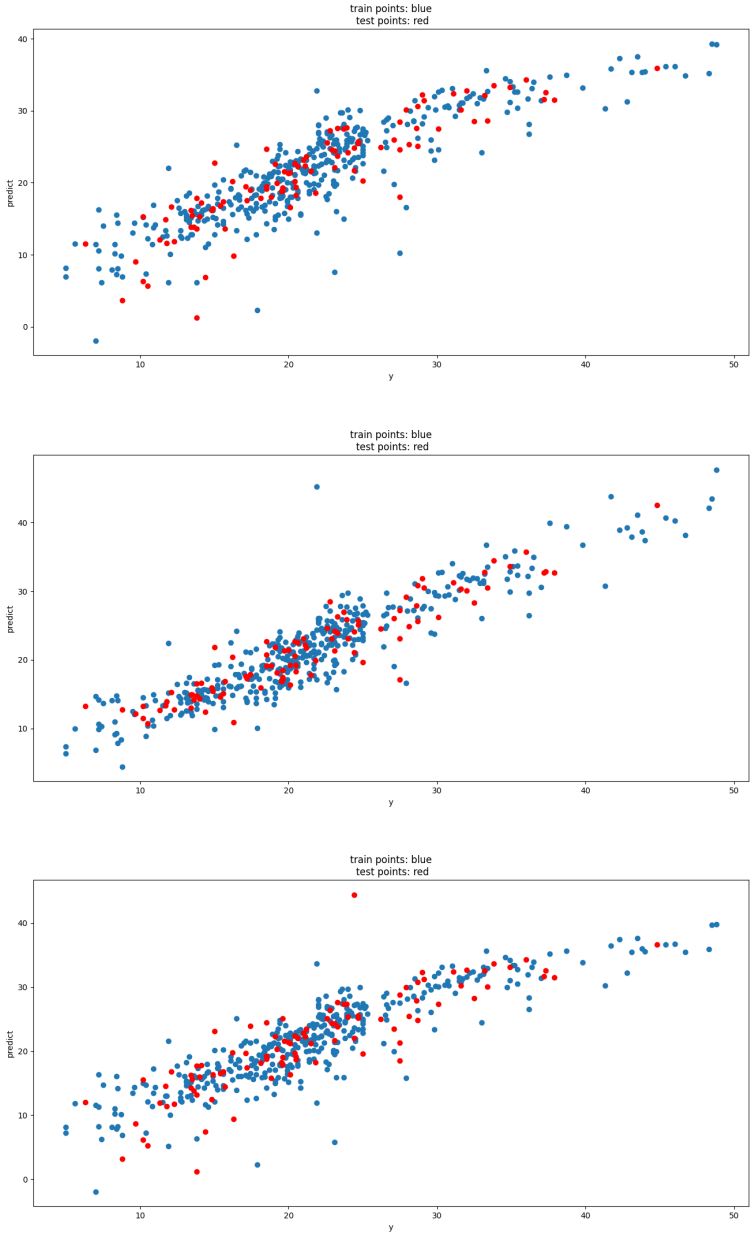
۶

در بخش نهایی به محاسبه‌ی خطای اندازه‌گیری در هر یک از ۳ بخش پیشین بر حسب داده‌های آزمایش می‌پردازیم. توجه کنید این داده‌ها از ابتدای تمرین از فایل اولیه جدا شده‌اند و هیچ نقشی در اندازه‌گیری پارامترها تا کنون نداشته‌اند. از این داده‌ها صرفا برای بررسی دقت اندازه‌گیری استفاده می‌شود.



شکل ۶: میزان خطا بر روی داده‌های آموزش و آزمایش در سه رویکرد بخش‌های پیشین.

همچنین نمودار مقادیر مورد انتظار بر حسب مقادیر تخمین زده شده بر روی هر دو داده‌ی آموزش و آزمایش را برای سه رویکرد پیشین به ترتیب مشاهده می‌کنید.



شکل ۷: تخمین بر حسب مقدار مورد انتظار به ترتیب بخش‌های ۳، ۴ و ۵

منابع

https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-

f409b7e4a155

https://www.youtube.com/watch?v=VEluK6Mp340

https://www.youtube.com/watch?v=MleRltu3BUk&t=481s