

# «به نام خدا»

تکلیف هفتم – سوال دوم – مرضیه علیدادی – 9631983  
(کد های مربوط، در دو فرمت py و ipynb. ضمیمه شده اند.)

## 2.

(a) این دیتاست از نوع `sklearn.utils.bunch` است. یعنی دیتا را در قالب `Obj` های دیکشنری مانند و دسته هایی شامل یک سری کلید ذخیره می کند.

شامل 4 تا کلید است:

`data`: شامل اطلاعات خانه های مختلف است.

`target`: قیمت خانه

`feature_names`: اسم feature ها

`DESCR`: دیتابیس را توصیف می کند.

```
boston = load_boston()
type(boston)
```

```
sklearn.utils.Bunch
```

```
boston.keys()
```

```
dict_keys(['data', 'target', 'feature_names', 'DESCR', 'filename'])
```

برای این که بیشتر درباره ی `feature` ها اطلاعات به دست آوریم، از `DESCR` استفاده می کنیم:  
اطلاعات `attribute` ها نمایش داده شده است. در کل 13 تا `attribute` دارد. که `MEDV` همان `target` است.

**\*\*Data Set Characteristics:\*\***

:Number of Instances: 506

:Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

:Attribute Information (in order):

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per \$10,000
- PTRATIO pupil-teacher ratio by town
- B  $1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in \$1000's

:Missing Attribute Values: None

(b)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88

506 rows × 13 columns

(c)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Price
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88	11.9

506 rows × 14 columns

(i) هیچ مقدار null ای در دیتاست وجود ندارد.

```
df.isnull().sum()
```

```
CRIM      0
ZN        0
INDUS     0
CHAS      0
NOX       0
RM        0
AGE       0
DIS       0
RAD       0
TAX       0
PTRATIO   0
B         0
LSTAT     0
Price     0
dtype: int64
```

دو متغیر RAD و CHAS از نوع دسته ای هستند. بقیه عددی هستند.

پس از اعمال dummies روی این دو متغیر:

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	TAX	PTRATIO	B	...	CHAS_1.0	RAD_1.0	RAD_2.0	RAD_3.0	RAD_4.0	RAD_5.0	RAD_6.0	RAD_7.0	RAD_8.0	RAD_9.0
0	0.00632	18.0	2.31	0.538	6.575	65.2	4.0900	296.0	15.3	396.90	...	0	1	0	0	0	0	0	0	0	0
1	0.02731	0.0	7.07	0.469	6.421	78.9	4.9671	242.0	17.8	396.90	...	0	0	1	0	0	0	0	0	0	0
2	0.02729	0.0	7.07	0.469	7.185	61.1	4.9671	242.0	17.8	392.83	...	0	0	1	0	0	0	0	0	0	0
3	0.03237	0.0	2.18	0.458	6.998	45.8	6.0622	222.0	18.7	394.63	...	0	0	0	1	0	0	0	0	0	0
4	0.06905	0.0	2.18	0.458	7.147	54.2	6.0622	222.0	18.7	396.90	...	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.573	6.593	69.1	2.4786	273.0	21.0	391.99	...	0	1	0	0	0	0	0	0	0	0
502	0.04527	0.0	11.93	0.573	6.120	76.7	2.2875	273.0	21.0	396.90	...	0	1	0	0	0	0	0	0	0	0
503	0.06076	0.0	11.93	0.573	6.976	91.0	2.1675	273.0	21.0	396.90	...	0	1	0	0	0	0	0	0	0	0
504	0.10959	0.0	11.93	0.573	6.794	89.3	2.3889	273.0	21.0	393.45	...	0	1	0	0	0	0	0	0	0	0
505	0.04741	0.0	11.93	0.573	6.030	80.8	2.5050	273.0	21.0	396.90	...	0	1	0	0	0	0	0	0	0	0

506 rows × 23 columns

df.columns

```
Index(['CRIM', 'ZN', 'INDUS', 'NOX', 'RM', 'AGE', 'DIS', 'TAX', 'PTRATIO', 'B',
      'LSTAT', 'Price', 'CHAS_0.0', 'CHAS_1.0', 'RAD_1.0', 'RAD_2.0',
      'RAD_3.0', 'RAD_4.0', 'RAD_5.0', 'RAD_6.0', 'RAD_7.0', 'RAD_8.0',
      'RAD_9.0'],
      dtype='object')
```

پس از نرمال سازی متغیرهای عددی:

	CRIM	ZN	INDUS	NOX	RM	AGE	DIS	TAX	PTRATIO	B	...	CHAS_1.0	RAD_1.0	RAD_2.0	RAD_3.0	RAD_4.0	RAD_5.0	RAD_6.0	RAD_7.0	RAD_8.0	RAD_9.0
0	0.000013	0.035997	0.004620	0.001076	0.013149	0.130388	0.008179	0.591947	0.030597	0.793728	...	0	1	0	0	0	0	0	0	0	0
1	0.000058	0.000000	0.014977	0.000994	0.013602	0.167142	0.010522	0.512653	0.037708	0.840793	...	0	0	1	0	0	0	0	0	0	0
2	0.000059	0.000000	0.015175	0.001007	0.015421	0.131141	0.010661	0.519414	0.038205	0.843146	...	0	0	1	0	0	0	0	0	0	0
3	0.000071	0.000000	0.004785	0.001005	0.015360	0.100529	0.013306	0.487279	0.041046	0.866193	...	0	0	0	1	0	0	0	0	0	0
4	0.000151	0.000000	0.004755	0.000999	0.015588	0.118212	0.013222	0.484188	0.040785	0.865649	...	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.000130	0.000000	0.024679	0.001185	0.013638	0.142942	0.005127	0.564736	0.043441	0.810883	...	0	1	0	0	0	0	0	0	0	0
502	0.000093	0.000000	0.024421	0.001173	0.012528	0.157005	0.004683	0.558833	0.042987	0.812456	...	0	1	0	0	0	0	0	0	0	0
503	0.000124	0.000000	0.024301	0.001167	0.014210	0.185364	0.004415	0.556092	0.042776	0.808472	...	0	1	0	0	0	0	0	0	0	0
504	0.000225	0.000000	0.024455	0.001175	0.013927	0.183053	0.004897	0.559614	0.043047	0.806521	...	0	1	0	0	0	0	0	0	0	0
505	0.000097	0.000000	0.024389	0.001171	0.012327	0.165182	0.005121	0.558102	0.042931	0.811394	...	0	1	0	0	0	0	0	0	0	0

506 rows × 23 columns

(e)

- پارامترهای مربوط به مدل رگرسیون خطی تکلیف قبلی به این شکل بود:

$$MAE = 3.48 \quad \text{و} \quad MSE = 32.11 \quad \text{و} \quad RMSE = 5.66$$

و همچنین با در نظر گرفتن cross validation برای آن، میانگین دقت مدل برابر 0.75 شد.

- و پارامترهای مربوط به مدل پواسون در این تکلیف، به این شکل است:

$$MAE = 5.25 \quad \text{و} \quad MSE = 49.84 \quad \text{و} \quad RMSE = 7.06$$

و همچنین با در نظر گرفتن cross validation برای آن، میانگین دقت مدل برابر 0.26 شد.

← این بیانگر این است که مدل رگرسیون تکلیف قبلی، دارای خطای کمتر و دقت بیشتری بوده نسبت به مدل پواسن این تکلیف.