The Chinese University of Hong Kong, Shenzhen

DDA3020 Machine Learning

Assignment 1 Report

---

*Author:*Kangqi Yu *Student ID:*121090735

MARCH 2023

Contents

1 Written Problems

1.1 Question 1

(1) Since $\mathbf{X} \in \mathbb{R}^{h\times d}, \mathbf{y} \in \mathbb{R}^{h\times 1}$, we can denote $\mathbf{X}$ and $\mathbf{y}$ as:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{h1} & \cdots & x_{hd} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_h \end{bmatrix}$$

We can also denote $\mathbf{w}$ as:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

Therefore, $\mathbf{y}^T\mathbf{X}\mathbf{w}$ can be written as $\sum_{i=1}^{h} y_i x_{i1} w_1 + \sum_{i=1}^{h} y_i x_{i2} w_2 + \cdots + \sum_{i=1}^{h} y_i x_{id} w_d$, which is a scalar.
Then, we can get that

$$\frac{d(\mathbf{y}^T\mathbf{X}\mathbf{w})}{d\mathbf{w}} = \frac{d(\sum_{i=1}^{h} y_i x_{i1} w_1 + \sum_{i=1}^{h} y_i x_{i2} w_2 + \cdots + \sum_{i=1}^{h} y_i x_{id} w_d)}{d\mathbf{w}}$$

Therefore,

$$\frac{d(\mathbf{y}^T\mathbf{X}\mathbf{w})}{d\mathbf{w}} = \begin{bmatrix} \sum_{i=1}^{h} y_i x_{i1} \\ \sum_{i=1}^{h} y_i x_{i2} \\ \vdots \\ \sum_{i=1}^{h} y_i x_{id} \end{bmatrix} = \mathbf{X}^T\mathbf{y}$$

(2) We can denote $\mathbf{w}$ as:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

Therefore, $\mathbf{w}^T\mathbf{w}$ can be written as $w_1^2 + w_2^2 + \cdots + w_d^2$, which is a scalar.
Then, we can get that

$$\frac{d\mathbf{w}^T\mathbf{w}}{d\mathbf{w}} = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_d \end{bmatrix} = 2\mathbf{w}$$

(3) We can denote $\mathbf{w}$ and $\mathbf{X}$ as:

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{d1} & \cdots & x_{dd} \end{bmatrix}$$

Therefore, $\mathbf{w}^T\mathbf{X}\mathbf{w}$ can be written as $\sum_{j=1}^{d}\sum_{i=1}^{d} w_i w_j x_{ij}$, which is a scalar.

$$\frac{d(\mathbf{w}^T\mathbf{X}\mathbf{w})}{d\mathbf{w}} = \frac{d(\sum_{j=1}^{d}\sum_{i=1}^{d} w_i w_j x_{ij})}{d\mathbf{w}}$$

Then, we can get that

$$\frac{d(\mathbf{w}^T\mathbf{X}\mathbf{w})}{d\mathbf{w}} = \begin{bmatrix} \sum_{i=1}^{d} w_i x_{i1} + \sum_{j=1}^{d} w_j x_{1j} \\ \sum_{i=1}^{d} w_i x_{i2} + \sum_{j=1}^{d} w_j x_{2j} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{id} + \sum_{j=1}^{d} w_j x_{dj} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{d} x_{i1} + \sum_{j=1}^{d} x_{1j} \\ \sum_{i=1}^{d} x_{i2} + \sum_{j=1}^{d} x_{2j} \\ \vdots \\ \sum_{i=1}^{d} x_{id} + \sum_{j=1}^{d} x_{dj} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = (\mathbf{X} + \mathbf{X}^T)\mathbf{w}$$

1.2 Question 2

(1) Let's add an 1 on the head of each $\mathbf{x}$ and add a b on the head of each $\mathbf{w}$. So, $f_{\mathbf{w},b}(\boldsymbol{x}_i)$ will be converted to $f_{\mathbf{w}}(\boldsymbol{x}_i)$.

Denote $\mathbf{X}$ as:

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix}$$

Denote $\boldsymbol{y}$ as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}$$

To solve $\min_{\mathbf{w}} \sum_{i=1}^{N} (f_{\mathbf{w}}(\boldsymbol{x}_i) - y_i)^2 + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}$

$$\frac{\partial}{\partial \mathbf{w}} (\sum_{i=1}^{N} (f_{\mathbf{w}}(\boldsymbol{x}_i) - y_i)^2 + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}) = 0$$

.

$$\frac{\partial}{\partial \mathbf{w}} (\sum_{i=1}^{N} ((\boldsymbol{x}_i^T \mathbf{w}) - y_i)^2 + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}) = 0$$

$$\frac{\partial}{\partial \mathbf{w}} ((\mathbf{X}\mathbf{w} - \boldsymbol{y})^T (\mathbf{X}\mathbf{w} - \boldsymbol{y}) + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}) = 0$$

$$2\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + 2\lambda \hat{\mathbf{I}}_d \mathbf{w} = 0$$

$$\mathbf{X}^T \mathbf{X}\mathbf{w} + \lambda \hat{\mathbf{I}}_d \mathbf{w} = \mathbf{X}^T y$$

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \hat{\mathbf{I}}_d)^{-1} \mathbf{X}^T \mathbf{y}$$

So $[\hat{b};\ \hat{\mathbf{w}}]^T = (\mathbf{X}^T \mathbf{X} + \lambda \hat{\mathbf{I}}_d)^{-1} \mathbf{X}^T \mathbf{y}$

(2) Denote $\sum_{i=1}^{N} (f_{\mathbf{w}}(\boldsymbol{x}_i) - y_i)^2 + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}$ as $J(\mathbf{w})$.

Solve $\min_{\mathbf{w}} J(\mathbf{w})$ by gradient descent algorithm,

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}$$

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y}) + 2\lambda \hat{\mathbf{I}}_d \mathbf{w}$$

α is called learning rate.

1.3 Question 3

(1) Since $f(x) = x^2$, the first order derivative of $f(x)$ is $2x$. Then, its second order derivative is 2. Since $2 > 0$, we can draw a conclusion that $f(x) = x^2$ is convex.

(2) Since $f(x) = ax + b$, the first order derivative of $f(x)$ is a. Then, its second order derivative is 0. Since $0 = 0$, we can draw a conclusion that $f(x)$ is convex but not strictly convex.

(3)

$$f'(x) = \begin{cases} 1, & x > 0 \\ -1, & x \leq 0 \end{cases}$$

$\forall x_0 \leq 0 \;\; x_1 \leq 0, f(x_1) - (f(x_0) + f'(x_0)(x_1 - x_0)) = -x_1 - (-x_0 - x_1 + x_0) = 0$
$\forall x_0 \geq 0 \;\; x_1 \geq 0, f(x_1) - (f(x_0) + f'(x_0)(x_1 - x_0)) = x_1 - (x_0 + x_1 - x_0) = 0$
$\forall x_0 \leq 0 \;\; x_1 \geq 0, f(x_1) - (f(x_0) + f'(x_0)(x_1 - x_0)) = x_1 - (-x_0 - x_1 + x_0) = 2x_1 \geq 0$
$\forall x_0 \geq 0 \;\; x_1 \leq 0, f(x_1) - (f(x_0) + f'(x_0)(x_1 - x_0)) = -x_1 - (x_0 + x_1 - x_0) = -2x_1 \geq 0$
So, in all cases, $f(x_1) \geq (f(x_0) + f'(x_0)(x_1 - x_0))$ holds true, which means $f(x)$ is convex.
But when $x_0 = -1$ and $x_1 = -2$,

$$f(\theta x_0 + (1 - \theta x_1)) = \theta f(x_0) + (1 - \theta) f(x_1)$$

So $f(x)$ is not strictly convex.

1.4 Question 4

The pdf of Laplace distribution is $f(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$.
Denote $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$.
So the MLE can be structed by $L(\mu, b|\boldsymbol{x}) = \frac{1}{(2b)^N} e^{\sum_{i=1}^{N} -\frac{|x_i-\mu|}{b}}$.
Take a logarithmic, and we can get $l(\mu, b|\boldsymbol{x}) = Nlog(\frac{1}{2b}) + \sum_{i=1}^{N} (-\frac{|x_i-\mu|}{b})$.
For b,

$$\left.\frac{\partial l(\mu, b|\mathbf{x})}{\partial b}\right|_{b=b^*} = 0$$

$$-\frac{N}{b^*} + \sum_{i=1}^{N} (\frac{|x_i - \mu|}{{b^*}^2}) = 0$$

$$\frac{N}{b^*} = \sum_{i=1}^{N} (\frac{|x_i - \mu|}{{b^*}^2})$$

$$b^* = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

Check the second order derivative,

$$\left.\frac{\partial^2 l(\mu, b|\mathbf{x})}{\partial b^2}\right|_{b=b^*} = \frac{N}{{b^*}^2} - \sum_{i=1}^{N} \frac{2|x_i - \mu|}{{b^*}^3} = -\frac{N^3}{(\sum_{i=1}^{N} |x_i - \mu|)^2} < 0$$

So b^* is the MLE of b.
For μ,

Firstly, let's proof the derivative of an absolute function:

$$|f| = |f| \Rightarrow |f|^2 = f^2 \Rightarrow 2|f| \cdot |f|' = 2f \cdot f' \Rightarrow |f|' = f' \cdot \frac{f}{|f|}$$

So,

$$\left.\frac{\partial l(\mu, b|\mathbf{x})}{\partial \mu}\right|_{\mu=\mu^*} = 0$$

$$\frac{\sum_{i=1}^{N} \frac{x_i - \mu^*}{|x_i - \mu^*|}}{b} = 0$$

$$\sum_{i=1}^{N} sgn(\frac{x_i - \mu^*}{|x_i - \mu^*|}) = 0$$

So, the MLE of μ is $\mu^* = median(\boldsymbol{x})$.

2 Programming

2.1 Select features

After checking, this dataset without na/null. The following is the statistical summary:

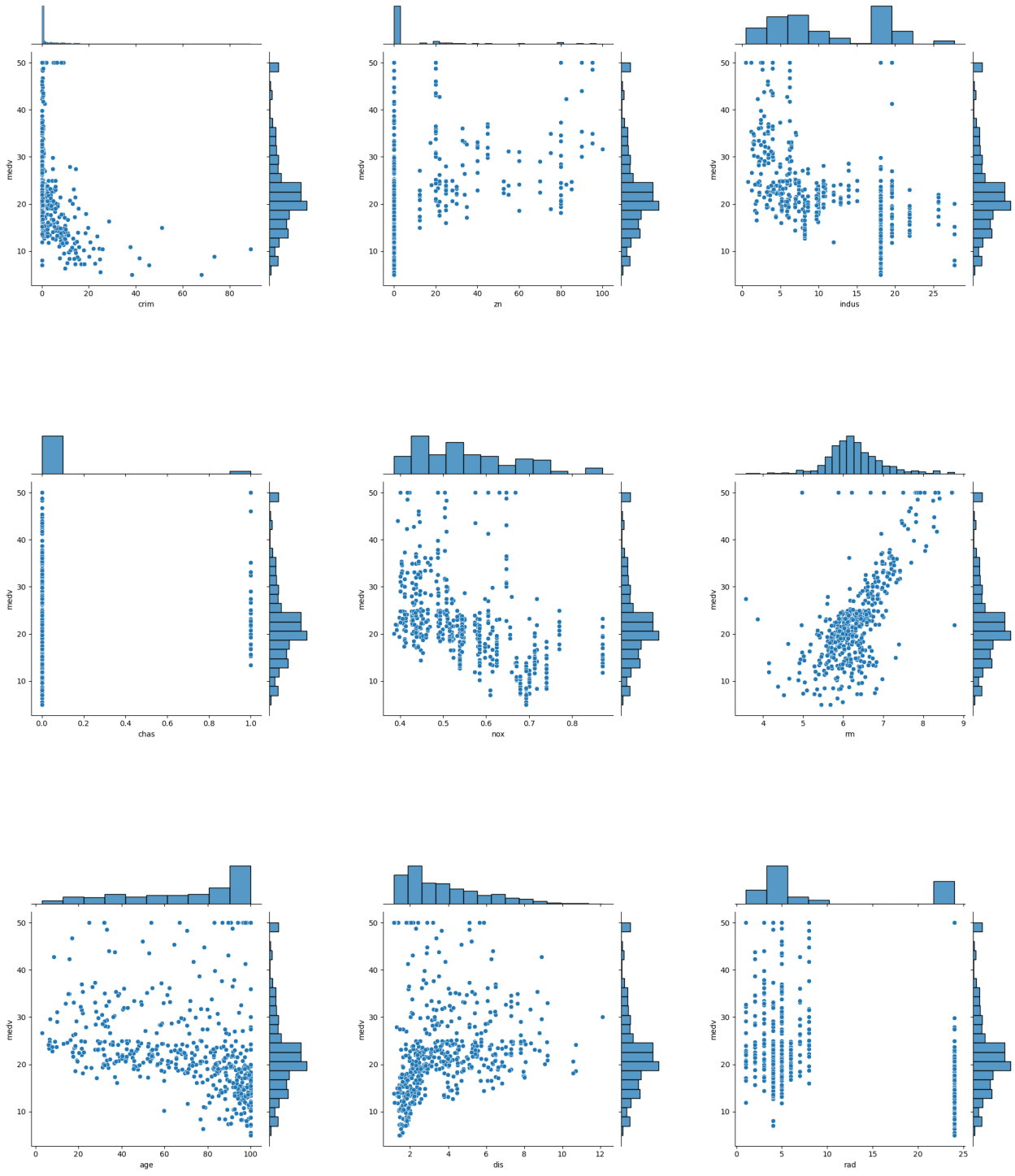
	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

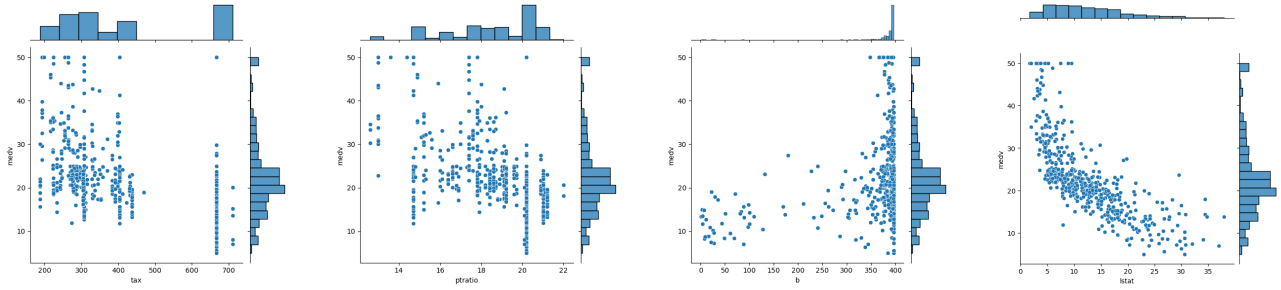
Table 1: The statistical summary of all varialbes

Findings:

- The variable crim has a large std based on the mean of it, and the max of it is 88.98, but the 75
- The varialbe zn is similar to crim.
- Chas is a dummy varialbe. From the mean of it, we can see most homes do not at the tract bounds river.
- The std of variable rm and ptratio is small. These variables are very stable.
- The 75% quantile of tax is the same as max of it.

So I guess crim, zn, tax are the most related variables to MEDV.
But, after I plot the jointplots of all variables to MEDV.
I notice that indus, nox, rm, dis, lstat are the most related variables to MEDV.


medv
crim
zn
indus
chas
nox
rm
age
dis
rad



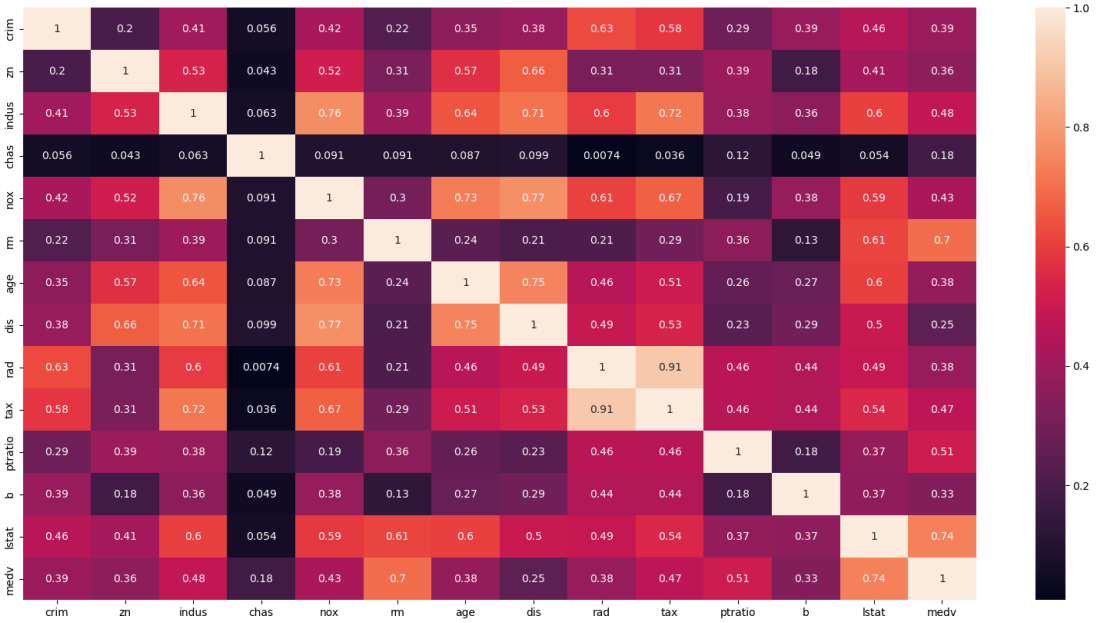


Figure 1: heatmap

Then I make a heatmap as Figure 2. From the heatmap we can see that mehv has strong correlation with indus, nox, rm, tax, ptratio and lstat. They are good indications of using as predictors.

2.2 Linear regression

2.2.1 OLS plot

The following is the OLS with 95% confidence interval.

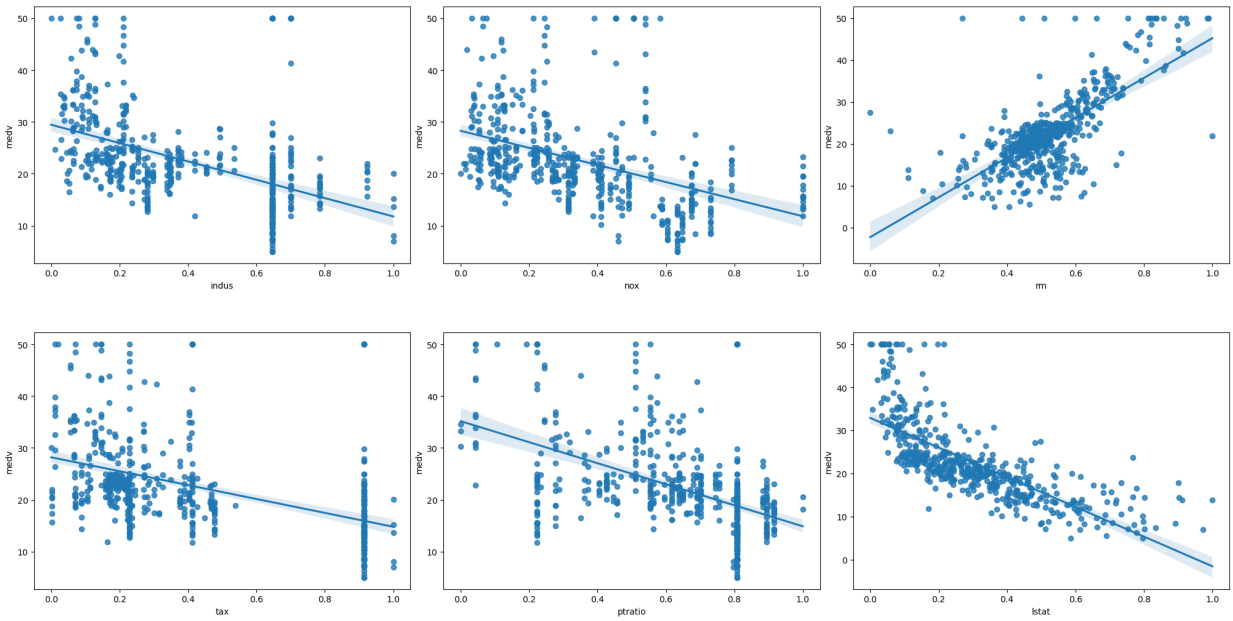


Figure 2: regplot

2.2.2 Linear model description

Form the analysis in the ipynb file, I select indus, nox, rm, tax, ptratio and lstat to construct a linear model to predict medv:

$$medv = w_0 + w_1 indus + w_2 nox + w_3 rm + w_4 tax + w_5 ptratio + w_6 lstat$$

As there are six features, so the training dataset $\mathbf{X} \in \mathbb{R}^{404\times 7}$ and $\mathbf{y} \in \mathbb{R}^{404\times 1}$. We can set the coefficients vector $\mathbf{w} \in \mathbb{R}^{7\times 1}$. The linear model will be $\mathbf{y} = \mathbf{X}\mathbf{w}$.

2.2.3 Loss function and gradient descent

Denote m as the number of observation.
We can write the loss function based on $\mathbf{e} = \mathbf{X}\mathbf{w} - \mathbf{y}$:

$$J(\mathbf{w}) = \frac{1}{2m}\mathbf{e}^T\mathbf{e} = \frac{1}{2m}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

Then we can use gradient descent to find the optimal $\mathbf{w}$.
Choosig a initial $\mathbf{w}$, then repeat num_iters times:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{m}\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}}$$

α is the learning rate.

2.2.4 Hyperparameter settings

As we only test for ten times, we can not use heuristic algorithm to find the appropriate hyperparameters. So I set α to be one of $[0.005, 0.01, 0.015, 0.02, 0.025]$ and num_iters to be one of $[10, 20]$.

2.2.5 RMSE equation

$$RMSE = \sqrt{\frac{1}{m}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})}$$

2.3 The result of first training

In the first training, I choose α to 0.01 and num_iters to be 20. The following is the loss plot and the RMSE of training set and testing set.

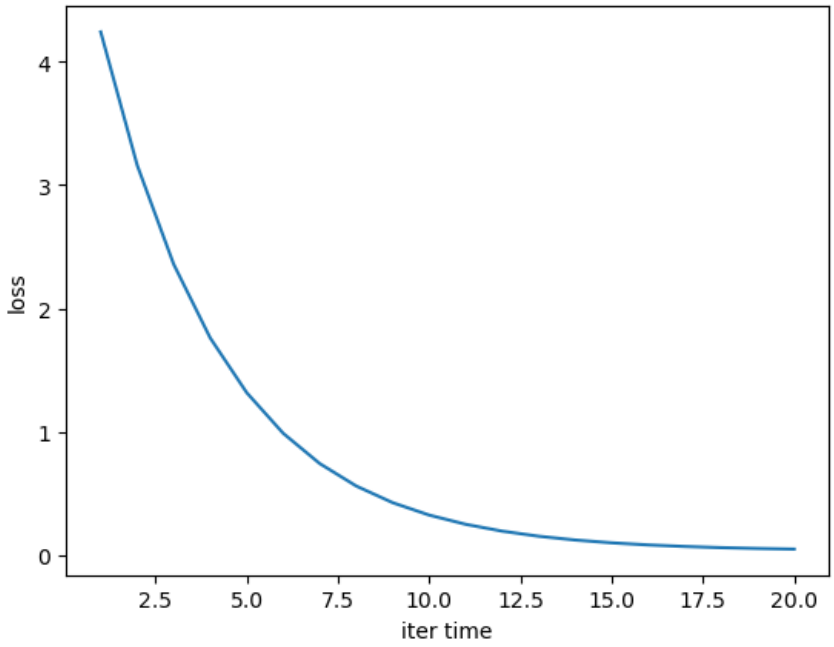


Figure 3: The first gradient descent (loss is $J(\mathbf{w})$)

	RMSE
training set	0.3177675935748315
testing set	0.3407474271659464

Table 2: RMSE of first gradient descent (RMSE)

2.3.1 Different parameters

	0.005	0.01	0.015	0.02	0.025
10	1.67 and 1.72	0.81 and 0.84	0.43 and 0.45	0.30 and 0.32	0.27 and 0.29
20	0.85 and 0.88	0.32 and 0.34	0.27 and 0.29	0.27 and 0.28	0.27 and 0.28

Table 3: RMSE of different parameters (the left is training dataset and the right is testing)

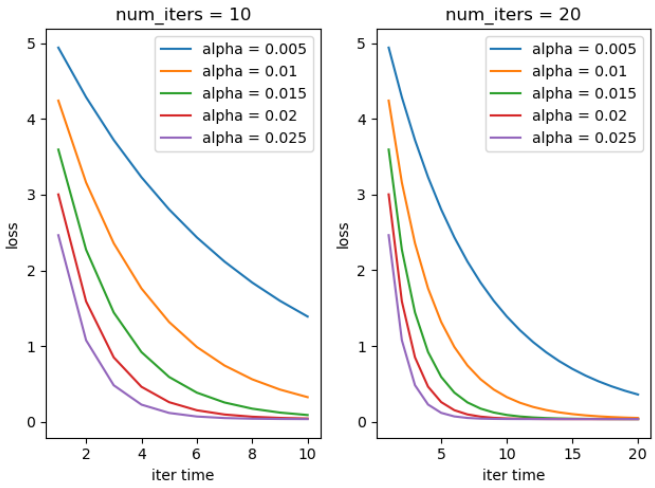


Figure 4: gradient descent (loss is $J(\mathbf{w})$)

From the above table and figure, we can see that when the learning rate α and iter tiems are too small. The gradient descent method may not be able to let $\mathbf{w}$ rearch the optimal $\mathbf{w}$. So, the RMSE is higher than the other parameters. In an appropriate interval, the higher the parameters are, the less the RMSE of training or testing are. However, beyond a certain threshold of them, the RMSEs will not decrease and further increasing in these parameters yield negligible improvements. Basically, the RMSE of testing dataset is higher than the training dataset.