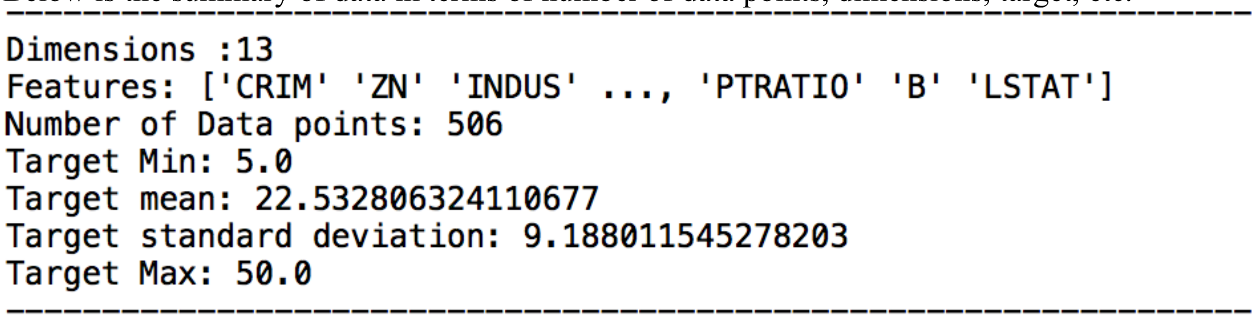
CSC411A1
Guanxiong Liu
1002077726

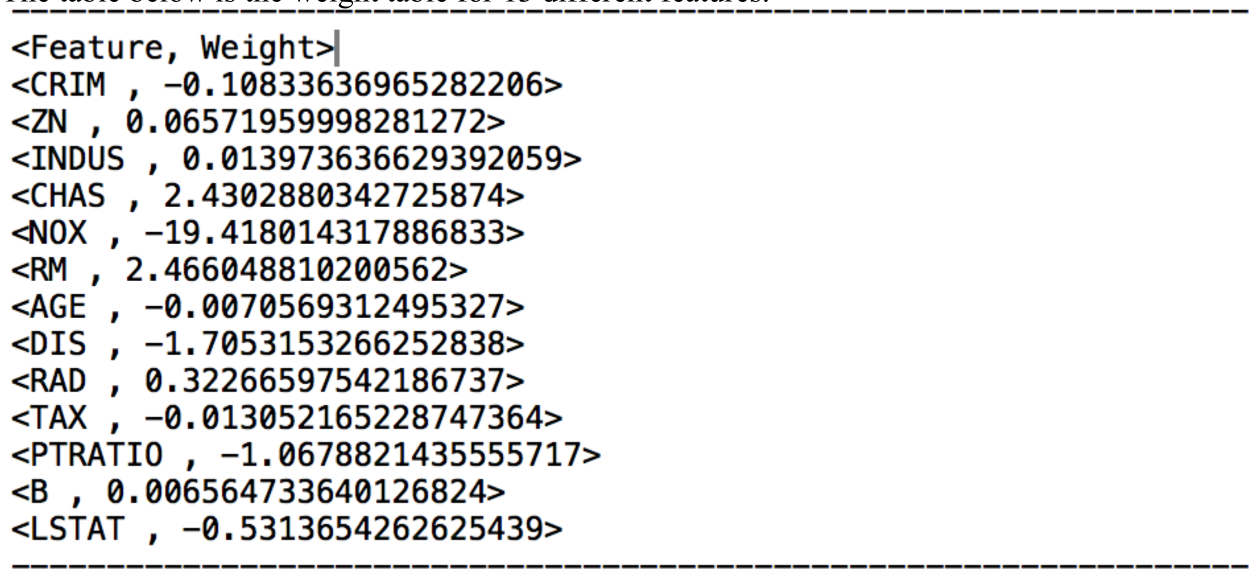
Q1:

Below is the summary of data in terms of number of data points, dimensions, target, etc.

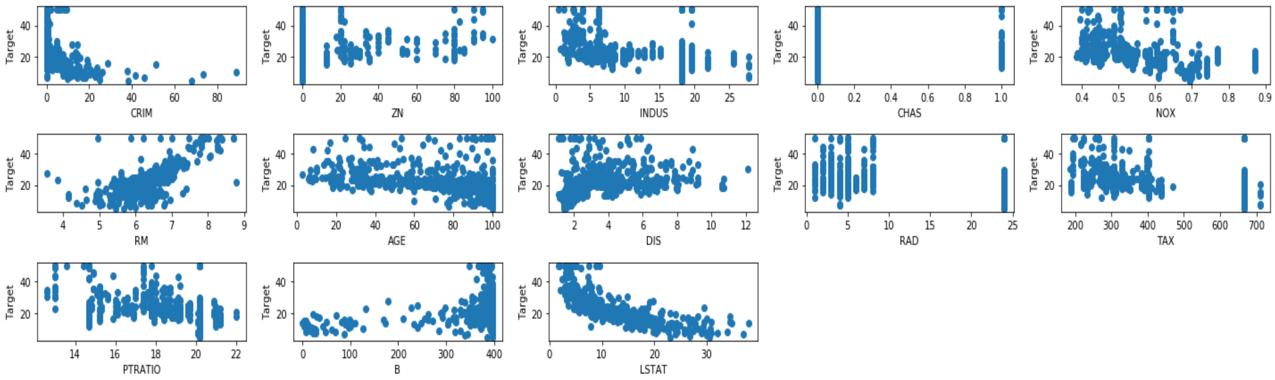


```
------------------------------------------------------------------
Dimensions :13
Features: ['CRIM' 'ZN' 'INDUS' ..., 'PTRATIO' 'B' 'LSTAT']
Number of Data points: 506
Target Min: 5.0
Target mean: 22.532806324110677
Target standard deviation: 9.188011545278203
Target Max: 50.0
------------------------------------------------------------------
```

The table below is the weight table for 13 different features:



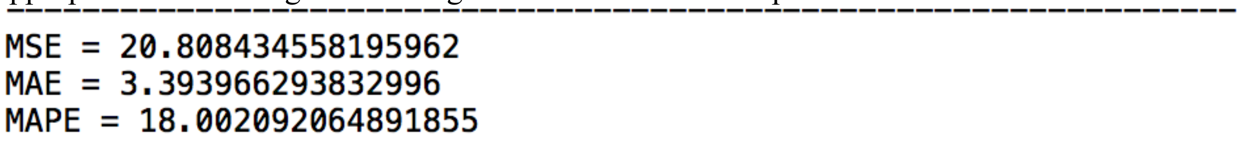
```
------------------------------------------------------------------
<Feature, Weight>
<CRIM , -0.10833636965282206>
<ZN , 0.06571959998281272>
<INDUS , 0.013973636629392059>
<CHAS , 2.4302880342725874>
<NOX , -19.418014317886833>
<RM , 2.466048810200562>
<AGE , -0.0070569312495327>
<DIS , -1.7053153266252838>
<RAD , 0.32266597542186737>
<TAX , -0.013052165228747364>
<PTRATIO , -1.0678821435555717>
<B , 0.006564733640126824>
<LSTAT , -0.5313654262625439>
------------------------------------------------------------------
```



As we can see, INDUS has weight 0.013973. The sign is positive. That means feature, INDUS, has positive but a little effect on Boston's house price. It is what I expected. If a place has prosperous industry. It will have more people live in that area. The price will then go up. But people don't like living in the pollution area, so the positive effect on price is very little.

The MSE is 29.274835388128697.

The table below is three error measurements. I think mean absolute percentage error is more appropriate because it gives a strong sense of how much the prediction differs from the actual.



```
MSE = 20.808434558195962
MAE = 3.393966293832996
MAPE = 18.002092064891855
```

The nitric oxides concentration will be the best prediction. Since it has the largest weight in magnitude, it affects house price a lot. The more nitric oxides the less the house price because it has negative effect on house price.

Q2:

1.

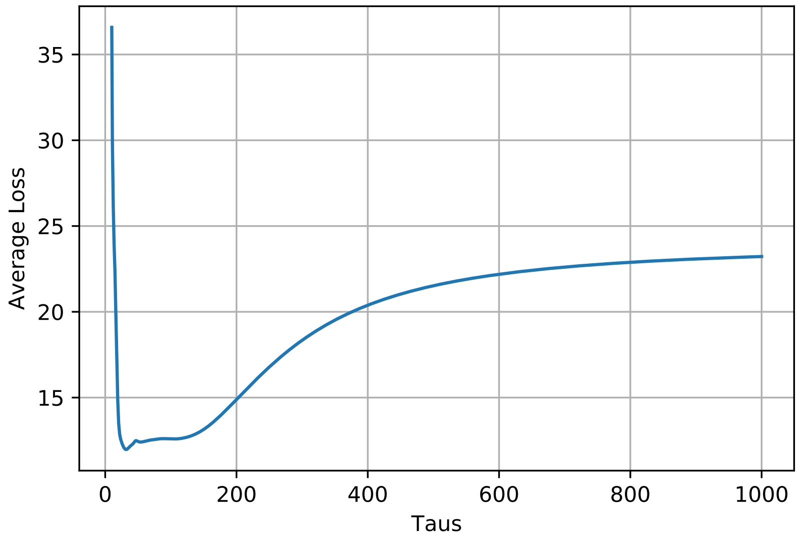
$$\mathrm{L(w)} = \frac{1}{2}(y - Xw)^T A(y - Xw) = \frac{1}{2}(y^T Ay - y^T AXw - w^T X^T Ay + w^T X^T AXw)$$

$$\nabla \mathrm{L}(w) = \frac{1}{2}(-2X^T Ay + 2X^T AXw) + \lambda w = -X^T Ay + X^T AXw + \lambda w = 0$$

$$(X^T Ax + \lambda I)w = X^T Ay$$

$$w = (X^T AX + \lambda I)^{-1} X^T Ay$$

2. See the code attached.
3.



4. This algorithm makes the average loss go to infinity as tau goes to 0, and as tau goes to infinity, the average loss is close to 23.

Q3:

1.

$$\begin{aligned} E_{\mathcal{J}}\left[\frac{1}{m}\sum_{i\in\mathcal{J}} ai\right] &= E_{\mathcal{J}}\left[\frac{1}{m}(a_{i_1}+a_{i_2}+\cdots+a_{i_m})\right] \\ &= \frac{1}{m}\left[E(a_{i_1})+E(a_{i_2})+\cdots+E(a_{i_m})\right] \qquad \# E(a_{i_1}) = \frac{1}{n}\sum_{i=1}^{n} a_i \text{ by CLT} \\ &= \frac{1}{m}\left[m\frac{1}{n}\sum_{i=1}^{n} a_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} a_i \end{aligned}$$

2.

$$\begin{aligned} E_{\mathcal{J}}[\nabla L_{\mathcal{J}}(\boldsymbol{x},y,\theta)] &= E_{\mathcal{J}}\left[\nabla(\frac{1}{m}\sum_{i\in\mathcal{J}} l(\boldsymbol{x}^i,y^i,\theta))\right] && \text{\#by equation (3)} \\ &= \nabla\left[E_{\mathcal{J}}(\frac{1}{m}\sum_{i\in\mathcal{J}} l(\boldsymbol{x}^i,y^i,\theta))\right] \\ &= \nabla\left[\frac{1}{n}\sum_{i=1}^{n}\ l(\boldsymbol{x}^i,y^i,\theta)\right] && \text{\#prove in question 1} \\ &= \nabla L(\boldsymbol{x},y,\theta) && \text{\#given in the question} \end{aligned}$$

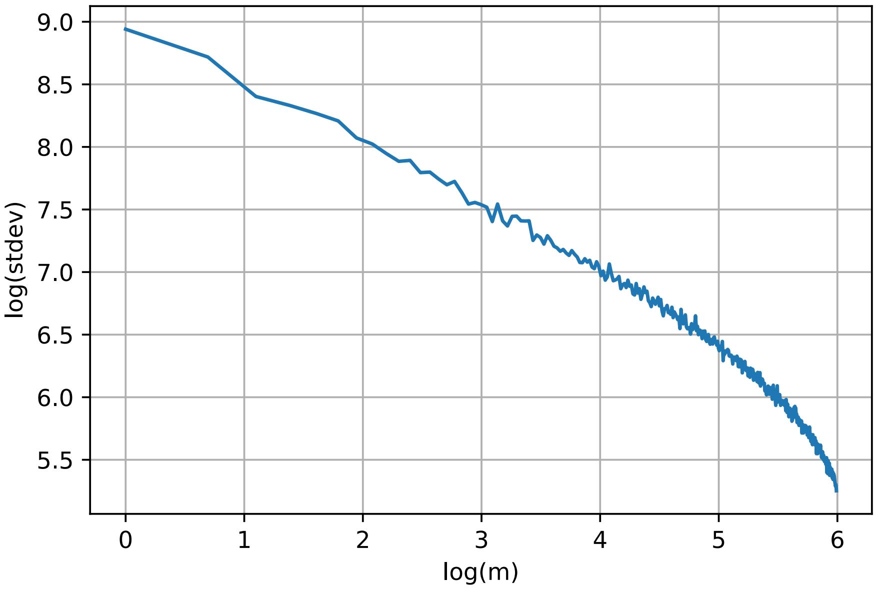
3. Mini-batch SGD Gradient Estimator is unbiased.

4.

$$\begin{aligned} L(\boldsymbol{x},y,\theta) &= \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{x}^i,y^i,\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}(y-\boldsymbol{w}^T\boldsymbol{x})^2 \\ \nabla L(x,y,\theta) &= \frac{1}{n}\nabla\sum_{i=1}^{n}(y-\boldsymbol{w}^T\boldsymbol{x})^2 \\ &= \frac{1}{n}(2\boldsymbol{X}^T\mathbf{X}\mathbf{w}-2\mathbf{X}^{\mathrm{T}}\mathrm{y}) \end{aligned}$$

5. cosine similarity is more import because it can predict the direction. Squared distance I calculated were different every time. So it is not worth for referencing.

6.


log(stdev)
log(m)