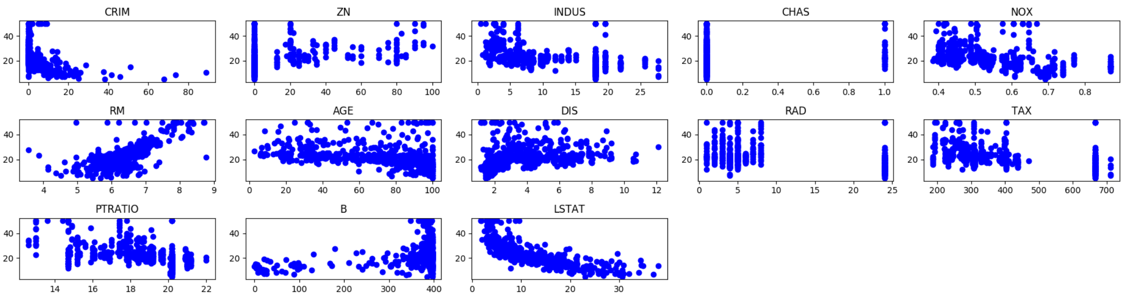
CSC411 A1

Shiyuan Lin

October 2017

Question 1 *Tabulate Features*



BIAS	39.7053268485
CRIM	-0.10285526522556665
ZN	0.051674620968530635
INDUS	0.02469877192394195
CHAS	2.7512818986117176
NOX	-19.209000026395376
RM	4.068989771643805
AGE	-0.007482183946514214
DIS	-1.5813690378086012
RAD	0.29299147745687015
TAXL	-0.012852687494588852
PTRATIO	-0.8807973330344827
B	0.007898140994755908
LSTAT	-0.46393957984444784

Calculate the Mean Square Error
MSE: 20.4099209442

Two more error measurement metrics
I chooses RMSE and MAD for error measurement metrics. As their values are

smaller than the MSE's value and more sensible for human to understand the result. What's more, RMSE and MAD are easier to compile based on the MSE code computed

Feature Selection

Based on my result, I would say NOX and RM are the most significant features that best predict the price. Since both two features have the greatest absolute values for their w as well as their graph show that they are continous.

Question 2

2.1

$w^* = argmin\frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y^{(i)} - w^T x^{(i)})^2 + \frac{\lambda}{2}||w||^2$

$\frac{dw^*}{dw} = -2 \cdot \frac{1}{2}argmin\sum_{i=1}^{N} a^{(i)}(y^{(i)} - w^T x^{(i)})^2 \cdot x^{(i)} + 2 \cdot \frac{\lambda}{2}||w||$

Let $\frac{dw^*}{dw} \rightarrow 0$, *we get minimum value for w*

$0 = -\sum_{i=1}^{N} a^{(i)}y^{(i)}x^{(i)} + \sum_{i=1}^{N} a^{(i)}w^T x^{(i)}x^{(i)} + \lambda w$

$0 = -\sum_{i=1}^{N} a^{(i)}y^{(i)}x^{(i)^T} + \sum_{i=1}^{N} a^{(i)}x^{(i)}x^{(i)^T}w + \lambda w$

$0 = -\sum_{i=1}^{N} a^{(i)}y^{(i)}x^{(i)^T} + (\sum_{i=1}^{N} a^{(i)}x^{(i)}x^{(i)^T} + \lambda)w$

$\sum_{i=1}^{N} a^{(i)}y^{(i)}x^{(i)^T} = (\sum_{i=1}{}_{N} a^{(i)}x^{(i)}x^{(i)^T} + \lambda) \cdot w$

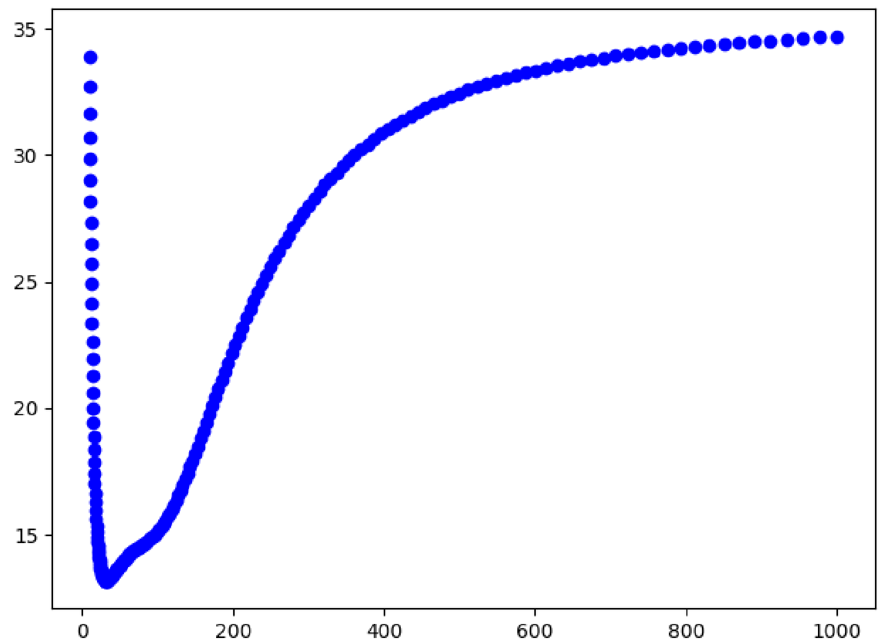
$X^T Ay = (X^T AX + \lambda I)w$

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

2.2

min loss = 13.159575940105318

2.3



2.4
From graph,
$\tau \to 0, loss \to \infty$
$\tau \to \infty, loss \to k, where \exists K \in N$

Question 3
3.1
$E_\imath[\frac{1}{m}\sum_{i\in\imath} a_i] = \frac{1}{n}\sum_{i=\imath}^{n} a_i$
$E_\imath[\frac{1}{m}\sum_{i\in\imath} a_i] = E_\imath[\overline{a}] = \frac{1}{m}E_\imath[\sum_{i\in\imath} a_i] = \frac{1}{m}(\frac{1}{\binom{n}{m}}\sum_{i=1}^{\binom{n}{m}}[\sum_{i\in\imath} a_i]) = \frac{1}{m}(\frac{1}{\binom{n}{m}}\binom{n-1}{m-1}\sum_{i=1}^{n} a_i) = \frac{1}{m}(\frac{m}{n}\sum_{i=1}^{n} a_i) = \frac{1}{n}\sum_{i=1}^{n} a_i$

3.2
$E_\imath[\nabla L_\imath(x,y,\theta)] = E_\imath[\frac{1}{m}\sum_{i\in\imath}\nabla\ell(x^{(i)},y^{(i)},\theta) = \frac{1}{n}\sum_{i=1}^{n}\nabla\ell(x^{(i)},y^{(i)},\theta) = \nabla L(x,y,\theta)$

3.3
Using this result, we only need to calculate the gradient of one mini-batch instead of the whole set of data.

3.4
From the question, we know that $\ell(x,y,\theta) = (y - w^T x)^2$ *then,*
$\nabla\ell(x,y,\theta) = -2(y - w^T x)\cdot x$
Hence that,

$\nabla L(x,y,\theta) = \frac{1}{n}\nabla \ell(x,y,\theta) = \frac{-2}{n}(y - w^T x)x$

3.5
Cosine similarity: 0.99940281309
Square matrix distance: 411085359.563
I think cosine similarity is a more meaningful measure. Since it is easier for human to understand and the value is more stable during the several times running. As for the square matrix distance, its value varies largely through couple of times running.

3.6
For w_0, log $\tilde{\sigma}_j$ against log m