CSC311H1 Assignmnet 2

Xiaoyu Zhou

September 10, 2020

1. (a) To show that $h_{arg}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$, which is $m = \frac{1}{n}\sum_{i=1}^{n} Y_i$, we want to find the minimum of $\frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2$.
   Let $L(m) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - 2Y_i m + m^2)$
   $L'(m) = \frac{1}{n}\sum_{i=1}^{n}(-2Y_i + 2m)$
   $= \frac{2}{n}\sum_{i=1}^{n}(m - Y_i)$
   Let $L'(m) = 0$, we can have that :
   $\frac{2}{n}\sum_{i=1}^{n}(m - Y_i) = 0$
   $\frac{2}{n}\sum_{i=1}^{n}(m) - \frac{2}{n}\sum_{i=1}^{n} Y_i = 0$
   $2m = \frac{2}{n}\sum_{i=1}^{n} Y_i$
   $m = \frac{1}{n}\sum_{i=1}^{n} Y_i$
   Thus, $h_{arg}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the solution.

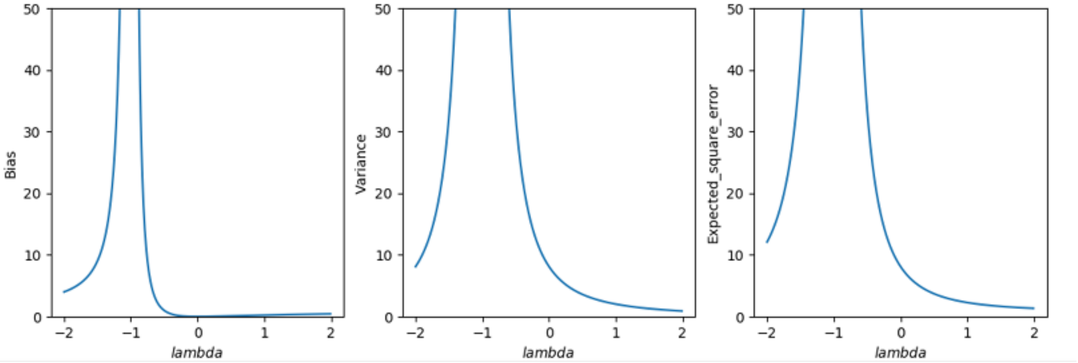
   (b) from part a, we can know that $h_{arg}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
   So the bias of $h_{arg}(\mathcal{D})$ is:
   $|\mathbb{E}_{\mathcal{D}}[h_{arg}(\mathcal{D})] - \mu|^2$
   $= |\mathbb{E}_{\mathcal{D}}[\frac{1}{n}\sum_{i=1}^{n} Y_i] - \mu|^2$
   $= |\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[Y_i] - \mu|^2$
   $= |\frac{1}{n}\sum_{i=1}^{n} \mu - \mu|^2$
   $= 0$
   The variance of the $h_{arg}(\mathcal{D})$ is:
   $\mathbb{E}_{\mathcal{D}}[|h_{arg}(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_{arg}(\mathcal{D})]|^2]$
   $= \mathbb{E}_{\mathcal{D}}[|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}_{\mathcal{D}}[\frac{1}{n}\sum_{i=1}^{n} Y_i]|^2]$
   $= \mathbb{E}_{\mathcal{D}}[|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[Y_i]|^2]$
   $= \mathbb{E}_{\mathcal{D}}[|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} \mu|^2]$
   $= \mathbb{E}_{\mathcal{D}}[|\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)|^2]$
   $= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}}[(Y_i - \mu)^2]$
   $= \frac{1}{n^2}\sum_{i=1}^{n} Var[(Y_i - \mu)^2]$
   $= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2$
   $= \frac{\sigma^2}{n}$

   (c) Let $f(m) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2$
   $f(m) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2$
   $= \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - 2Y_i m + m^2) + \lambda m^2$
   $= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{2m}{n}\sum_{i=1}^{n} Y_i + m^2 + \lambda m^2$
   $f'(m) = -\frac{2}{n}\sum_{i=1}^{n} Y_i + 2m + 2\lambda m$
   $= -2h_{arg}(\mathcal{D}) + 2m(1 + \lambda)$

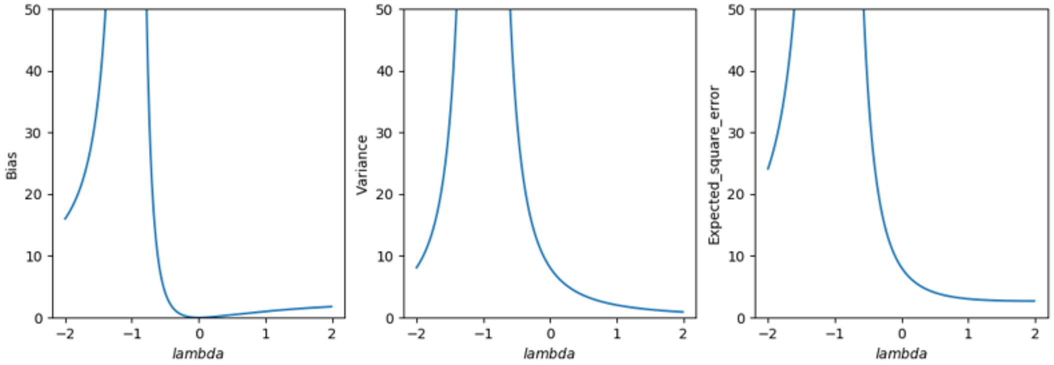
To find the estimator, let $f'(m) = 0$
So $-2h_{arg}(\mathcal{D}) + 2m(1+\lambda) = 0$
$2h_{arg}(\mathcal{D}) = 2m(1+\lambda)$
$m = \frac{h_{arg}(\mathcal{D})}{1+\lambda}$
So, the mean estimator $h_\lambda(\mathcal{D}) = m = \frac{h_{arg}(\mathcal{D})}{1+\lambda}$

(d) From part(c), we can know that $h_\lambda(\mathcal{D}) = \frac{h_{arg}(\mathcal{D})}{1+\lambda}$.
So the bias of $h_\lambda(\mathcal{D})$ is:
$|\mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})] - \mu|^2$
$= |\mathbb{E}_{\mathcal{D}}[\frac{h_{arg}(\mathcal{D})}{1+\lambda}] - \mu|^2$
$= |\frac{1}{1+\lambda}\mathbb{E}_{\mathcal{D}}[h_{arg}(\mathcal{D})] - \mu|^2$
$= |\frac{1}{1+\lambda}\mu - \mu|^2$ (from part(b), we can know that $\mathbb{E}_{\mathcal{D}}[h_{arg}(\mathcal{D})] = \mu$)
$= |\frac{1-1-\lambda}{1+\lambda}\mu|^2$
$= \frac{\lambda^2}{(1+\lambda)^2}\mu^2$
The variance of the $h_\lambda(\mathcal{D})$ is:
$\mathbb{E}_{\mathcal{D}}[|h_\lambda(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[h_\lambda(\mathcal{D})]|^2]$
$= \mathbb{E}_{\mathcal{D}}[|\frac{h_{arg}(\mathcal{D})}{1+\lambda} - \mathbb{E}_{\mathcal{D}}[\frac{h_{arg}(\mathcal{D})}{1+\lambda}]|^2]$
$= \mathbb{E}_{\mathcal{D}}[|\frac{1}{1+\lambda}h_{arg}(\mathcal{D}) - \frac{1}{1+\lambda}\mathbb{E}_{\mathcal{D}}[h_{arg}(\mathcal{D})]|^2]$
$= \mathbb{E}_{\mathcal{D}}[|\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\mu}{1+\lambda}|^2]$
$= \mathbb{E}_{\mathcal{D}}[|\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\lambda}\mu|^2]$
$= \mathbb{E}_{\mathcal{D}}[|\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)|^2]$
$= \mathbb{E}_{\mathcal{D}}[\frac{1}{(1+\lambda)^2 n^2}\sum_{i=1}^{n}(Y_i - \mu)^2]$
$= \frac{1}{(1+\lambda)^2 n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}[(Y_i - \mu)^2]$
$= \frac{1}{(1+\lambda)^2 n^2}\sum_{i=1}^{n} Var[Y_i]$
$= \frac{1}{(1+\lambda)^2 n^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}}[(Y_i - \mu)^2]$
$= \frac{1}{(1+\lambda)^2 n^2} n\sigma^2$
$= \frac{\sigma}{(1+\lambda)^2 n}$

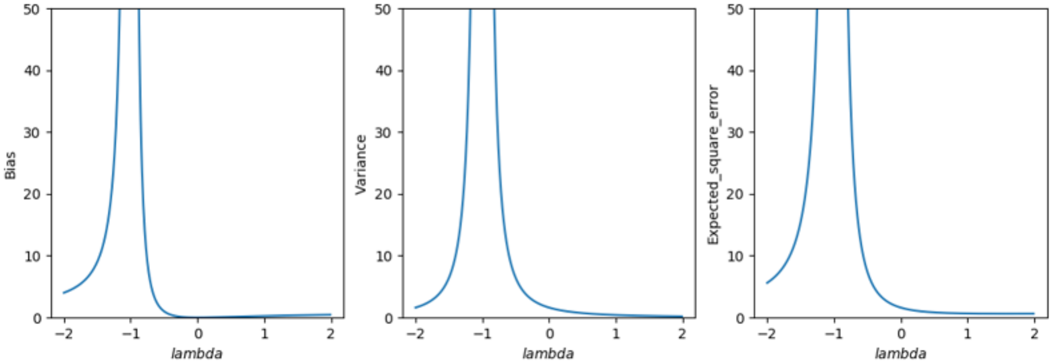
(e) For $\mu = 1, \sigma^2 = 9, n = 10$



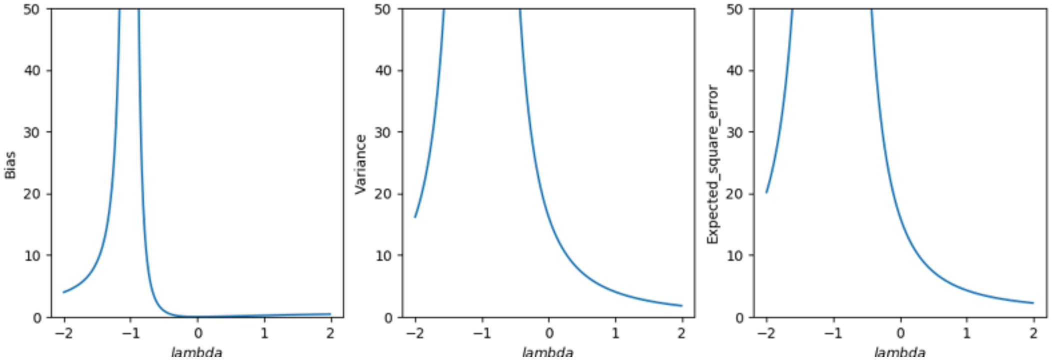
For $\mu = 2, \sigma^2 = 9, n = 10$



For $\mu = 1, \sigma^2 = 4, n = 10$



For $\mu = 1, \sigma^2 = 9, n = 5$

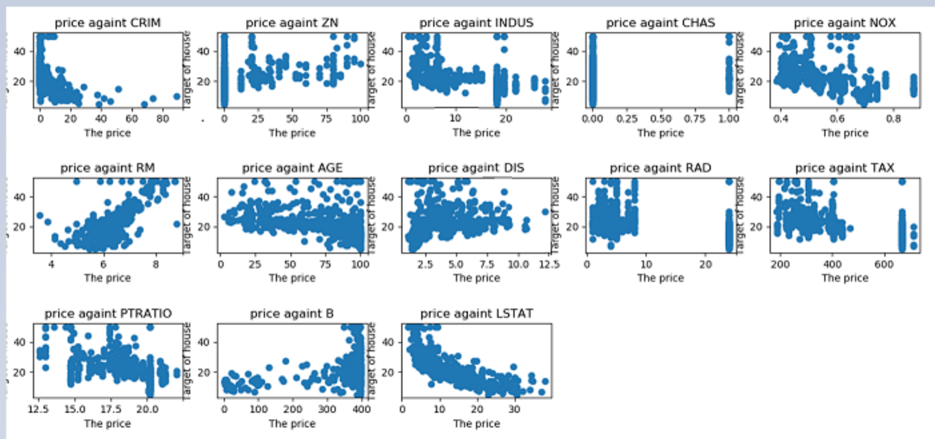


(f) Since the expected squared error is the sum of bias and variance, from the picture, we can clearly see that the expected squared error has a similar shape with the bias and the variance, where are all closed to the infinity near $\lambda = 1$ and tend to 0 when λ is big enough.
Also, we can find that near $\lambda = 0$, the shape of the error are more similar with the picture of variance and when λ tends to infinity, the error's shape are more similar with the bias. That means that when λ is small, the variance influences the expected square error more while the λ is big enough, the bias will influence more.

2. (a) See code

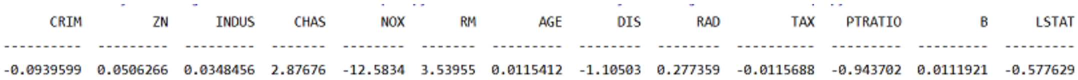
(b) The Boston data set is a 3D dataset, which includes 506 data points. Every data point includes 14 data depends on different features. The target of this data set is the houses' price to distinguish different houses.

(c)



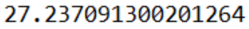
(d) See code

(e)


```
      CRIM         ZN      INDUS     CHAS        NOX       RM        AGE       DIS       RAD         TAX    PTRATIO          B      LSTAT
----------  ---------  ---------  -------  --------  -------  ---------  --------  --------  ----------  ---------  ---------  ---------
-0.0939599  0.0506266  0.0348456  2.87676  -12.5834  3.53955  0.0115412  -1.10503  0.277359  -0.0115688  -0.943702  0.0111921  -0.577629
```

The sign of the INDUS in the table is positive. Actually, it does not fit my expection because on the picture of the INDUS, we can see that the tendency of the picture is going down. So my expected sign is negative. However, the table is the weights based on the training data, which is only 70% of the whole data. So it is possible that the training data's weight is positive and the weight is really a small number.

(f) Here is the result of the Mean Square Error for the test set.


```
27.237091300201264
```

(g) The error measurement that I used are L2 regularization and root mean square error. The reason that I chose them are they are the measurements that the professor mentioned during the lecture and they are closely related to the mean square error. Both of them can reflect whether the algorithm is reliable or not.

Here is the result for L2 regularization: 4112.800786330391

Here is the result for root mean square error: 5.2189166787946775

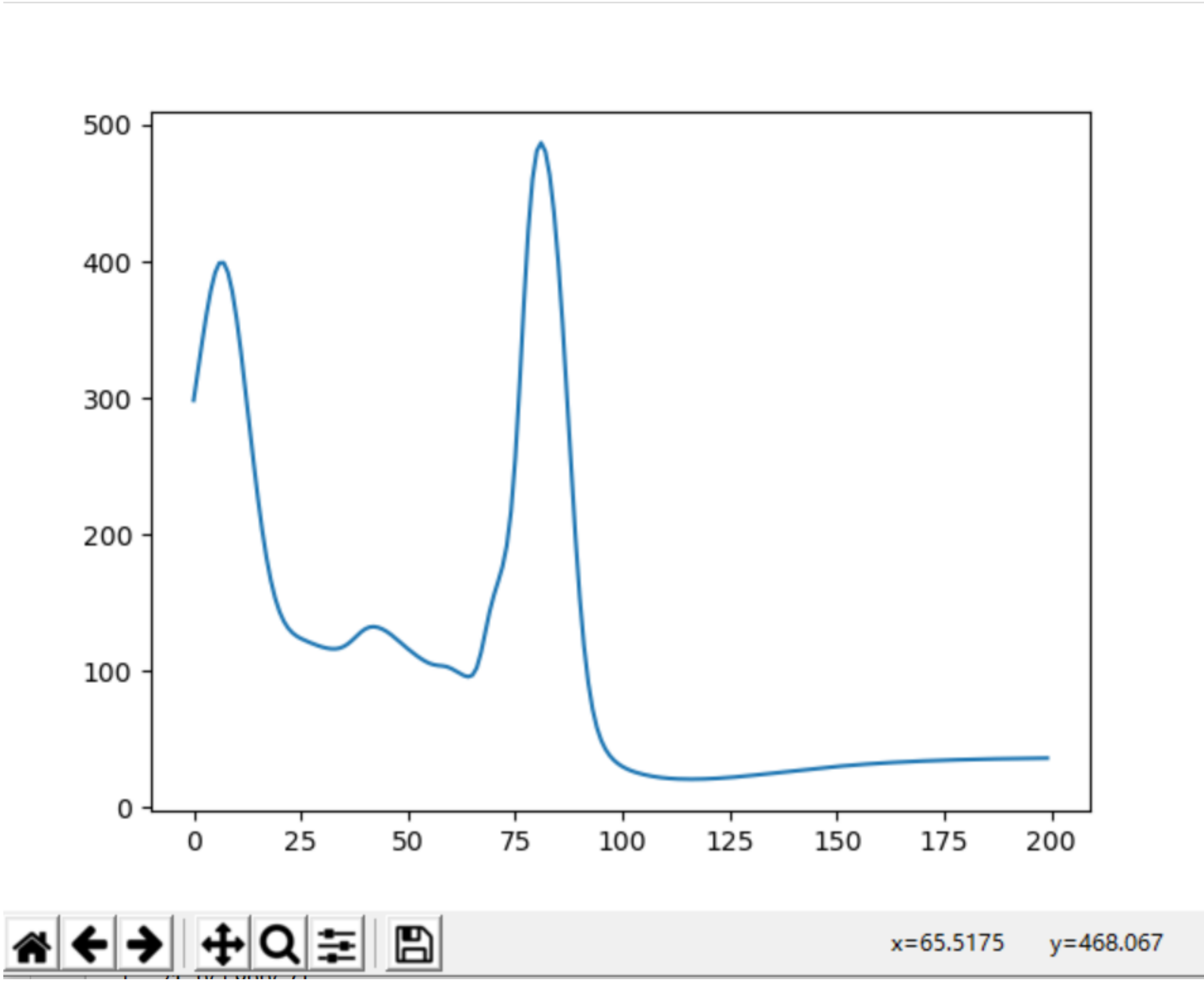
(h) I think the most significant feature is the NOX because its weight is much bigger that others. That means that the price would change dramatically for different houses depends on NOX.

Moreover, some other features like LSTAT and RM are also useful to predict the price since from the picture that we plot, we can easily see that the points are forming a clear line. That means there exists a closely relationship between the feature and the price, which can also help us to predict the price.

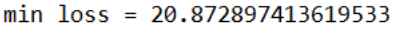
3. (a) Since $w^* = argmin\frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y^{(i)} - w^T x^{(i)})^2$, we can know that:
$L(w) = \frac{1}{2}A||y - wX||^2$ (since $A_{ii} = a^{(i)}$ and X is the design matrix.)
$= \frac{1}{2}(y - wX)^T A(y - wX)$
$= \frac{1}{2}y^T A(y - wX) - \frac{1}{2}w^T X^T A(y - wX)$
$= \frac{1}{2}y^T Ay - \frac{1}{2}y^T AwX - \frac{1}{2}w^T X^T Ay - \frac{1}{2}w^T X^T AwX$
Take the derivative of $L(w)$, we can have:
$\nabla_w L(w) = -X^T Ay + X^T AXw$
Let $\nabla_w L(w) = 0$.
So $-X^T Ay + X^T AXw^* = 0$
$X^T Ay = X^T AXw^*$
$w^* = (XTAX)^{-1}XTAy$

(b) See code

(c) Here is the loss values for each choice of τ



Here is the min loss:



```
min loss = 20.872897413619533
```

(d) For $\tau \to \infty$, this algorithm will behave as the linear regression. And we can predict the data more accuracy and the algorithm has more generalization at this time.
For $\tau \to 0$, the result will change a lot depends on the different points and its neighbourhood. So we cannot find any pattern to predict the value at this time.

(e) Advantage: it is useful for the data that are discrete; it fits more to the data instead of just a long line that the ordinary linear regression produces.
Disadvantage: it is more complicates; it need a lot of time and steps, even for the simple data.