CSC311 HOMEWORK 1

Q1

(a)

We want to compute $\underset{m\in R}{argmin}\ \frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2$.

To obtain max/min, we need to find the m which let the derivate of the function to be 0.

$$\frac{d}{dm}\frac{1}{\mathrm{n}}\sum_{i=1}^{n}|Y_i - m|^2 = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}-2(Y_i - m)$$

Therefore,

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n}-2(Y_i - h_{avg}(D)) = 0$$

$$\sum_{i=1}^{n}(Y_i - h_{avg}(D)) = 0$$

$$\sum_{i=1}^{n}Y_i - nh_{avg}(D) = 0$$

$$h_{avg}(D) = \frac{1}{n}\sum_{i=1}^{n}Y_i$$

To prove it obtain minimum, we take the second derivative,

$$\frac{d}{dm}\frac{1}{\mathrm{n}}\sum_{i=1}^{n}-2(Y_i - m) = \frac{1}{n} * 2n = 2 > 0$$

So it obtain minimum when m = $\frac{1}{n}\sum_{i=1}^{n}Y_i$

Therefore, the sample average estimator $\mathrm{h_{avg}} = \frac{1}{n}\sum_{i=1}^{n}Y_i$ is the solution of the optimize problem:

$$h_{avg}(D) \leftarrow \underset{m\in R}{argmin}\ \frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2$$

(b)

The bias of $h_{avg}(D)$ is

$$\left|\mathrm{E_D}\left[h_{avg}(D)\right] - \mu\right|^2 = \left|\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}Y_i\right] - \mu\right|^2 = \left|\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Y_i] - \mu\right|^2 = \left|\frac{1}{n}\sum_{i=1}^{n}\mu - \mu\right|^2 = |\mu - \mu|^2 = 0$$

The variance of $h_{avg}(D)$ is

$$\mathrm{E_D}\left[\left|h_{avg}(D) - E_D\left[h_{avg}(D)\right]\right|^2\right]$$

$$= E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right|^2\right] = E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]\right|^2\right]$$

$$= E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i]\right|^2\right] = E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} \mu\right|^2\right]$$

$$= E\left[\left|\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu)\right|^2\right]$$

Let $\mathrm{Z_i} = Y_i - \mu$

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{2} (Y_i - \mu)\right|^2\right] = E\left[\left|\frac{1}{n}(Z_1 + Z_2)\right|^2\right] = \frac{1}{n^2}E[Z_1^2 + Z_2^2 + 2Z_1Z_2]$$

Since $\mathrm{Y_i}$ are independent, $\mathrm{Z_i}$ are also independent, hence $\mathrm{E}[\mathrm{Z_1}Z_2] = E[Z_1]E[Z_2] = E[Y_1 - \mu]E[Y_2 - \mu] = (\mathrm{E}[\mathrm{Y_1}] - \mu)(\mathrm{E}[\mathrm{Y_2}] - \mu) = 0$

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{2} (Y_i - \mu)\right|^2\right] = \frac{1}{n^2}E[Z_1^2 + Z_2^2 + 2Z_1Z_2] = \frac{1}{n}(E[Z_1^2] + E[Z_2^2]) = \frac{1}{n^2}\sum_{i=1}^{2} E[|Y_i - \mu|^2\ ]$$

Similarly, $E\left[\left|\frac{1}{n}\sum_{i=1}^{2}(Y_i - \mu)\right|^2\right] = = \frac{1}{n^2}\sum_{i=1}^{n} E[|Y_i - \mu|^2\ ]$.

Hence,

$$\mathrm{E_D}\left[\left|h_{avg}(D) - E_D\left[h_{avg}(D)\right]\right|^2\right] = E\left[\left|\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu)\right|^2\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} E[|Y_i - \mu|^2\ ] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

(c)

We want to compute $\underset{m\in R}{argmin}\ \frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2 + \lambda|m|^2$.

To obtain max/min, we need to find the m which let the derivate of the function to be 0.

$$\frac{d}{dm}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i - m|^2 + \lambda|m|^2\right) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} -2(Y_i - m) + 2\lambda m$$

Therefore,

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n} -2(Y_i - h_\lambda(D)) + 2\lambda h_\lambda(D) = 0$$

$$\lambda h_\lambda(D) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} (Y_i - h_\lambda(D))$$

$$\lambda h_\lambda(D) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} Y_i - \frac{1}{\mathrm{n}}\sum_{i=1}^{n} h_\lambda(D)$$

$$(\lambda+1)h_\lambda(D) = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} Y_i$$

$$h_\lambda(D) = \frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i = \frac{1}{\lambda+1}h_{avg}$$

To prove it obtain minimum, we take the second derivative,

$$\frac{d}{dm}\frac{1}{\mathrm{n}}\sum_{i=1}^{n} -2(Y_i - m) + 2\lambda m = \frac{1}{n} * 2n + 2\lambda = 2 + 2\lambda > 0 \text{ since } \lambda \geq 0$$

So it obtain minimum when m = $\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i$

Therefore, the estimator

$$h_\lambda(D) = \frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i = \frac{1}{\lambda+1}h_{avg}$$

is the solution of the optimize problem:

$$h_\lambda(D) \leftarrow \underset{m\in R}{argmin}\ \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2$$

(d)

The bias of $h_\lambda(D)$ is

$$|\mathrm{E}_\mathrm{D}[h_\lambda(D)] - \mu|^2 = \left|\mathrm{E}\left[\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i\right] - \mu\right|^2 = \left|\frac{1}{n(\lambda+1)}\sum_{i=1}^{n} \mathrm{E}[Y_i] - \mu\right|^2$$

$$= \left|\frac{1}{n(\lambda+1)}\sum_{i=1}^{n} \mu - \mu\right|^2 = \left|\frac{1}{(\lambda+1)}\mu - \mu\right|^2 = \left(\frac{\lambda}{\lambda+1}\mu\right)^2$$

The variance of $h_{avg}(D)$ is

$$\mathrm{E}_\mathrm{D}[|h_\lambda(D) - E_D[h_\lambda(D)]|^2]$$

$$= \mathrm{E}\left[\left|\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i - E\left[\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i\right]\right|^2\right]$$

$$= \mathrm{E}\left[\left|\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} Y_i - \frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} E[Y_i]\right|^2\right]$$

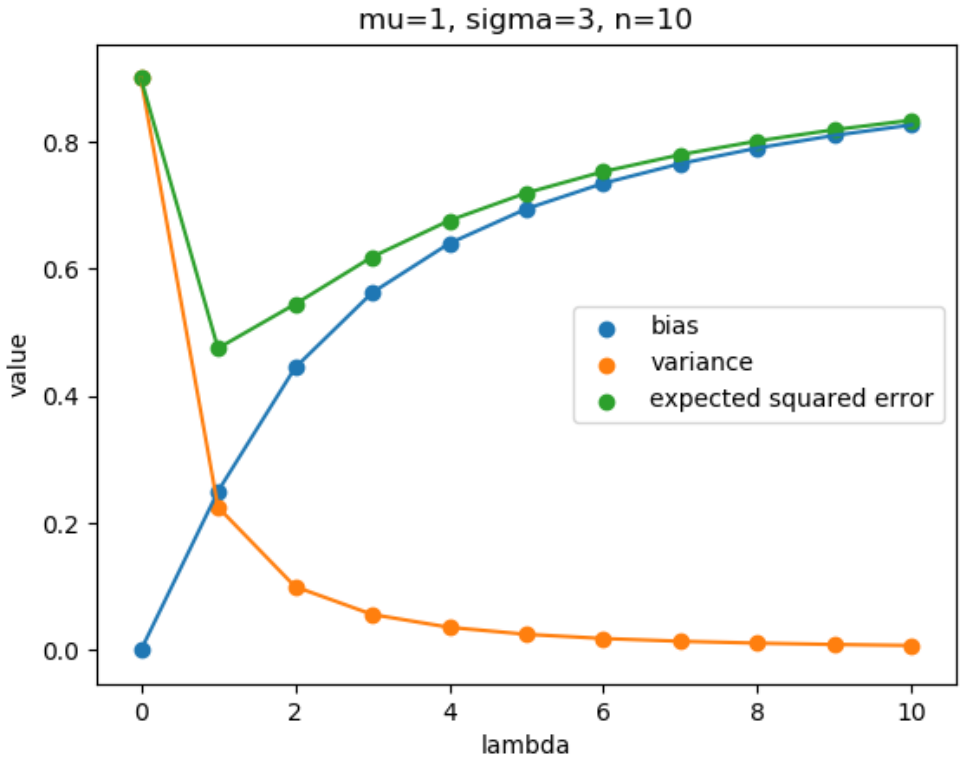
$$= \mathrm{E}\left[\left|\frac{1}{\mathrm{n}(\lambda+1)}\sum_{i=1}^{n} (Y_i - \mu)\right|^2\right]$$

Similar to question(b),

$$= \frac{1}{\mathrm{n}^2(\lambda+1)^2}\sum_{i=1}^{n} \mathrm{E}[(Y_i - \mu)^2]$$

$$= \frac{\sigma^2}{n(\lambda+1)^2}$$

(e)



(f)

Both bias and variance contribute to the expected squared error. When bias is equal to variance, the expected squared error obtains minimum.

The expected squared error converges to μ^2 as bias increases and variance decreases. And the converge speed decreases as bias increases and variance decreases.

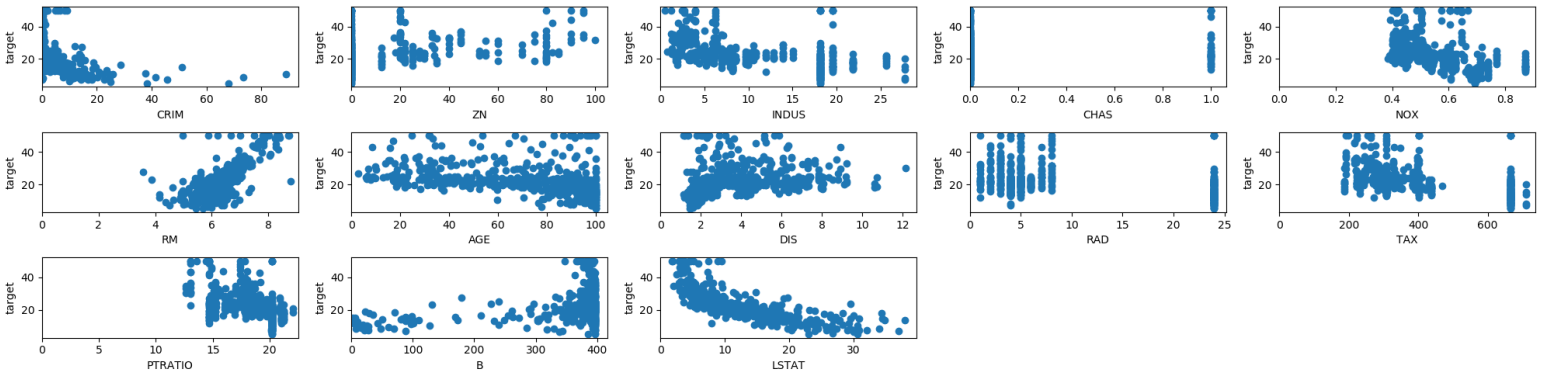
Q2

(b)

Number of data points	506

Dimensions	13
Target	Median value of owner-occupied homes in $1000's. Real number between 5. and 50.
Features	1. CRIM - per capita crime rate by town 2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft. 3. INDUS - proportion of non-retail business acres per town. 4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise) 5. NOX - nitric oxides concentration (parts per 10 million) 6. RM - average number of rooms per dwelling 7. AGE - proportion of owner-occupied units built prior to 1940 8. DIS - weighted distances to five Boston employment centres 9. RAD - index of accessibility to radial highways 10. TAX - full-value property-tax rate per $10,000 11. PTRATIO - pupil-teacher ratio by town 12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town 13. LSTAT - % lower status of the population All values are non-negative real number.

(c)



(e)

feature	weight
BIAS	24.12582032461337
CRIM	-8.985410101854086
ZN	3.9958249833540758
INDUS	2.05625773874197
CHAS	0.2645639721583636
NOX	-6.969685841771517
RM	25.224381733300913
AGE	-0.722526784049925

DIS -14.590029086830917
RAD 6.290396309153057
TAX -6.835056369043815
PTRATIO -8.992283980450706
B 2.547903262627145
LSTAT -17.633324805565643

The sign of the weight of INDUS is positive, meaning that as INDUS increases, the median value of owner-occupied homes in $1000's also increases.

The sign match what I expected as the proportion of non-retail business acres per town increases, people tend to be richer so the median value of owner-occupied homes in $1000's will increase.

(f)
The mean square error is 28.405854810508153.

(g)
Error measurement 1: Mean Absolute Error

The formula is $\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$.

This error measurement is directly calculating prediction error on each data point and take the average of all errors to measure the average error of the fitted model.
The mean absolute error is 3.691362677116224.

Error measurement 2: Mean Absolute Percentage Error

The formula is $\mathrm{MPE} = \frac{100\%}{\mathrm{n}}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right|$.

This error measurement is "the computed average of absolute percentage errors by which forecasts of a model differ from actual values of the quantity being forecast." (In *Wikipedia*. From https://en.wikipedia.org/wiki/Mean_percentage_error)
The mean absolute error is 19.699996167172944%.

(h)
The further the weight is away from 0, the more significant the feature is. This is because the weight represents the change of target as one unit increase in the feature.
From the list of weights, we can see the most 3 significant features are RM, DIS and LSTAT.

Q3
(a)

We want to compute $\underset{w}{argmin}\ \frac{1}{2}\sum_{i=1}^{n} a^{(i)}(y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)})^2 = \underset{w}{argmin}\ \frac{1}{2}(\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X})^T\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{X})$

where **X** is the design matrix and **A** is a diagonal matrix where $\mathbf{A}_{ii} = a^{(i)}$.
To obtain max/min, we need to find the w which let the derivate of the function to be 0.

$$\frac{d}{dw}\frac{1}{2}(\boldsymbol{y}-\boldsymbol{w}^T\boldsymbol{X})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{w}^T\boldsymbol{X}) = \frac{d}{dw}\frac{1}{2}(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{w}^T\boldsymbol{X}-\boldsymbol{X}^T\boldsymbol{w}\boldsymbol{A}\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{w}\boldsymbol{A}\boldsymbol{w}^T\boldsymbol{X})$$

$$= \frac{1}{2}(-\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}-\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}+2\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}) = -\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}$$

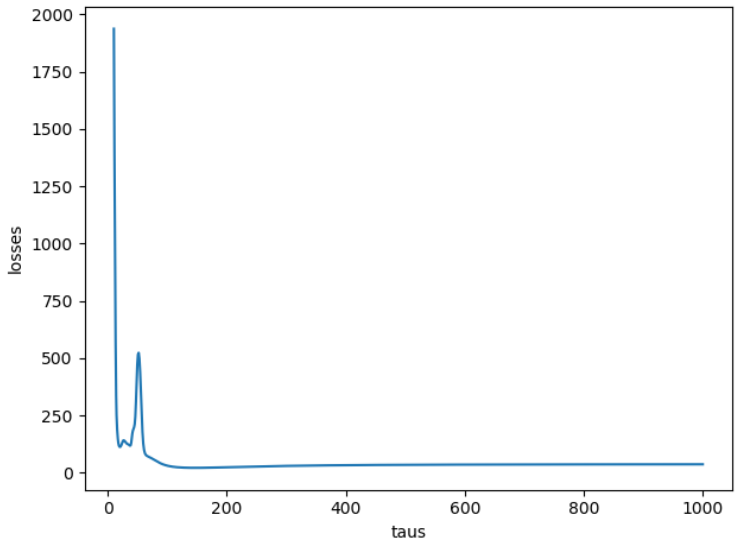
Since $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X} = \sum_{i=1}^{n} y^{(i)}a^{(i)}x^{(i)} = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}$,

$$\frac{d}{dw}\frac{1}{2}(\boldsymbol{y}-\boldsymbol{w}^T\boldsymbol{X})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{w}^T\boldsymbol{X}) = -\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w} = -\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}$$

Therefore,

$$-\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}+\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w}^* = 0$$
$$\boldsymbol{w}^* = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}$$

(c)



(d)

As $\tau \to \infty$, the weight for each data point $\lim_{\tau\to\infty} \mathrm{a}^{(i)} = \lim_{\tau\to\infty} \frac{\exp(-||x-x^{(i)}||^2/2\tau^2)}{\sum_j \exp(-||x-x^{(j)}||^2/2\tau^2)} = \frac{\exp(0)}{\sum_j \exp(0)} = \frac{1}{n}$.

So, the weight for each data point is equal. This means the regression becomes ordinary linear regression. So, the loss $\to$ mean squared error in ordinary linear regression.

As $\tau \to 0$, the weight for each data point is influenced by the distance between x and $\mathrm{x}^{(i)}$ heavily.

As for the case that x and $\mathrm{x}^{(i)}$ are similar, the weight $\to \frac{1}{\mathrm{n}}$ while if x and $\mathrm{x}^{(i)}$ differs lot, the weight $\to 0$. So, the algorithm performs bad and the loss $\to \infty$.

(e)

Advantages:

1. For data sets that are not linear, locally weighted linear regression fits better.
2. We can put less care into selecting the features in order to avoid overfitting.

Disadvantages:

1. It's much more computationally expensive compared to the simple linear regression.

2. It does not produce a regression function that is easily represented by linear regression.