CSC311 A2

Wilson Wei-Sheng Hsu

February 2020

Question 1

Part (a)

In order to find solution to $h_{avg}(D) \leftarrow argmin_{m \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - m|^2$ We need to find the critical point.
Take derivative w.r.t m:

$$\begin{aligned}
\frac{\partial f}{\partial m} &= \frac{\partial}{\partial m} \frac{1}{n} \sum_{i=1}^{n} (Y_i - m)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} -2(Y_i - m) \\
0 &= \frac{-2}{n} \sum_{i=1}^{n} (Y_i - m) && \text{Set to 0 to find critical point} \\
&= \frac{-2}{n} (\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} m) \\
&= \frac{-2}{n} (\sum_{i=1}^{n} Y_i - nm) \\
&= \frac{-2}{n} \sum_{i=1}^{n} Y_i + 2m \\
2m &= \frac{2}{n} \sum_{i=1}^{n} Y_i \\
m &= \frac{1}{n} \sum_{i=1}^{n} Y_i
\end{aligned}$$

Check second derivative w.r.t. m, to see if critical point is a minimum:

$$\begin{aligned}
\frac{\partial^2 f}{\partial^2 m} &= \frac{\partial}{\partial m} \frac{-2}{n} (\sum_{i=1}^{n} Y_i - nm) \\
&= \frac{\partial}{\partial m} (\frac{-2}{n} \sum_{i=1}^{n} Y_i + 2m) \\
&= 2 && \text{Hence, our critical point is a minimum}
\end{aligned}$$

and therefore we have shown that $h_{avg} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is the solution to this optimization problem.

Part (b)

$$
\begin{aligned}
Bias: |E_D[h(D)] - \mu|^2 &= |E[\frac{1}{n}\sum_{i=1}^{n} Y_i] - \mu|^2 \\
&= |\frac{1}{n}\sum_{i=1}^{n} E(Y_i) - \mu|^2 \\
&= |\frac{1}{n}\sum_{i=1}^{n} \mu - \mu|^2 && \text{We know } E(Y_i) = \mu \\
&= |\mu - \mu|^2 \\
&= 0 \\
Variance: E[|h(D) - E_D[h(D)]|^2] &= E[(\frac{1}{n}\sum_{i=1}^{n} Y_i - E[\frac{1}{n}\sum_{i=1}^{n} Y_i])^2] \\
&= E[(\frac{1}{n}\sum_{i=1}^{n} Y_i - m)^2] \\
&= \frac{1}{n^2} E[\sum_{i=1}^{n} (Y_i - m)^2] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} E[(Y_i - m)^2] \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

So our $bias = 0, Variance = \frac{\sigma^2}{n}$

Part (c)

In order to find solution to $h_{avg}(D) \leftarrow argmin_{m\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2$ We need to find the critical point. Take derivative w.r.t m:

$$\begin{aligned}
\frac{\partial f}{\partial m} &= \frac{\partial}{\partial m}\frac{1}{n}\sum_{i=1}^{n} |Y_i - m|^2 + \lambda|m|^2 \\
&= \frac{-2}{n}\sum_{i=1}^{n}(Y_i - m) + 2\lambda m \\
0 &= 2(\frac{-1}{n}\sum_{i=1}^{n}(Y_i - m) + \lambda m) && \text{Set to 0 to find critical point} \\
&= \frac{-1}{n}\sum_{i=1}^{n}(Y_i - m) + \lambda m \\
\lambda m &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - m) \\
\lambda m &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} m \\
\lambda m &= \frac{1}{n}\sum_{i=1}^{n} Y_i - m \\
m &= \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\lambda + 1} \\
m &= \frac{h_{avg}}{\lambda + 1}
\end{aligned}$$

Check second derivative w.r.t. m, to see if critical point is a minimum:

$$\begin{aligned}
\frac{\partial^2 f}{\partial^2 m} &= \frac{\partial}{\partial m}(\frac{-1}{n}\sum_{i=1}^{n}(Y_i - m) + \lambda m) \\
&= \frac{\partial}{\partial m}(\frac{-1}{n}\sum_{i=1}^{n} Y_i + \frac{1}{n}\sum_{i=1}^{n} m + \lambda m) \\
&= \frac{\partial}{\partial m}(\frac{-1}{n}\sum_{i=1}^{n} Y_i + m + \lambda m) \\
&= 1 + \lambda && \text{Since } \lambda \geq 0 \text{ by assumption, our critical point is a minimum}
\end{aligned}$$

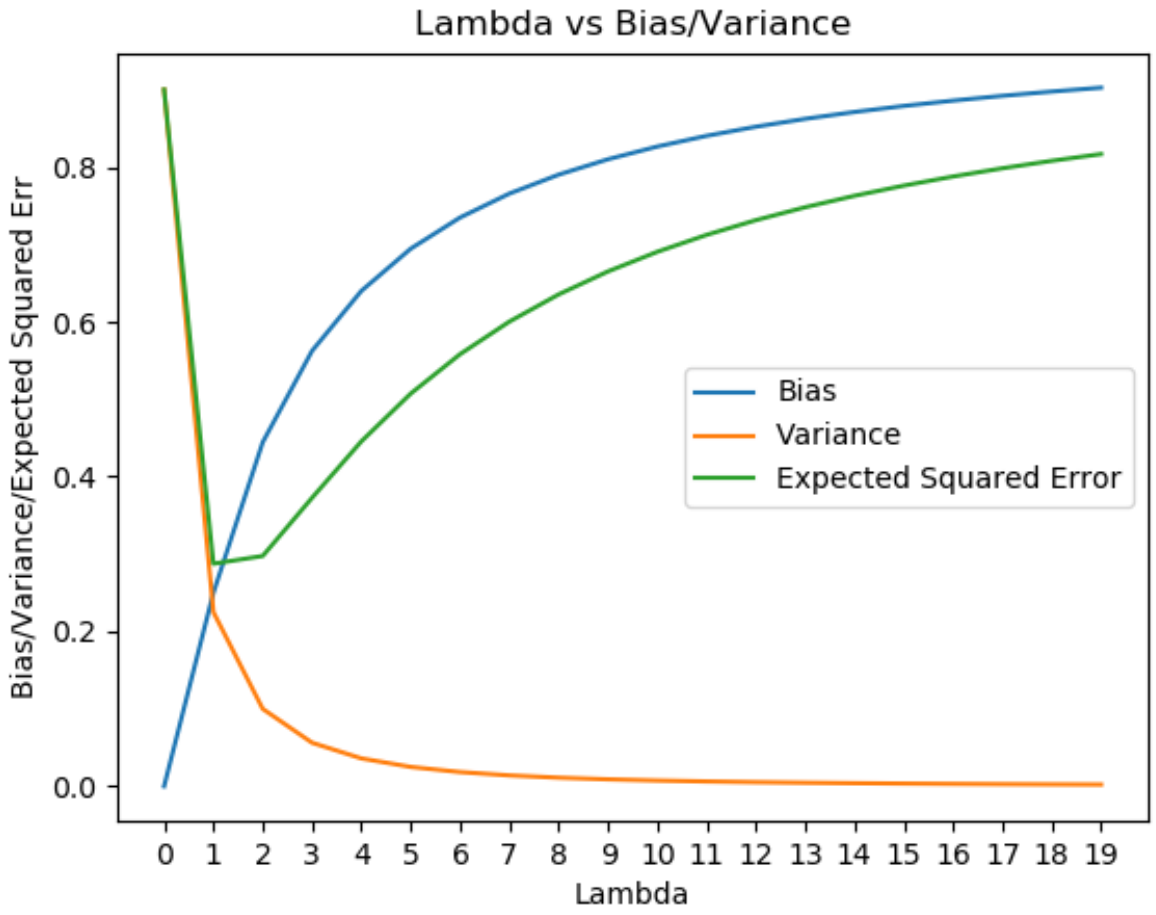
and therefore we found the solution to this optimization problem.

Part (d)

$$
\begin{aligned}
Bias: |E_D[h(D)] - \mu|^2 &= |E[\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\lambda+1} - \mu|^2 \\
&= |\frac{1}{n(\lambda+1)}\sum_{i=1}^{n} E(Y_i) - \mu|^2 \\
&= |\frac{n\mu}{n(\lambda+1)} - \mu|^2 && \text{We know } E(Y_i) = \mu \\
&= |\frac{\lambda\mu}{\lambda+1}|^2 \\
&= \frac{\lambda^2\mu^2}{(\lambda+1)^2} \\
Variance: E[|h(D) - E_D[h(D)]|^2] &= E[\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\lambda+1} - E[\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i}{\lambda+1}])^2] \\
&= E[(\frac{1}{n(\lambda+1)}\sum_{i=1}^{n} Y_i - \frac{1}{n(\lambda+1)}E[\sum_{i=1}^{n} Y_i])^2] \\
&= \frac{1}{n^2(\lambda+1)^2}E[(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} m)^2] \\
&= \frac{1}{n^2(\lambda+1)^2}E[(\sum_{i=1}^{n} Y_i - m)^2] \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^{n} E(Y_i - m)^2 \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^{n} E(Y_i - m)^2 && \text{We know } E(Y_i - m)^2 = \sigma^2 \\
&= \frac{1}{n^2(\lambda+1)^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{\sigma^2}{n(\lambda+1)^2}
\end{aligned}
$$

So our $bias = \frac{\lambda^2\mu^2}{(\lambda+1)^2}, Variance = \frac{\sigma^2}{n(\lambda+1)^2}$

Part (e)



Part (f)

We know the expected squared error formula is $E_D(|h(D) - \mu|^2) = bias^2 + variance$, therefore bias would bring a greater effect on expected squared error than variance. Which as we can see in our plot, that our expected squared error decreases first because of dramatic decrease in variance, but then grows with the increasing value of bias afterward.

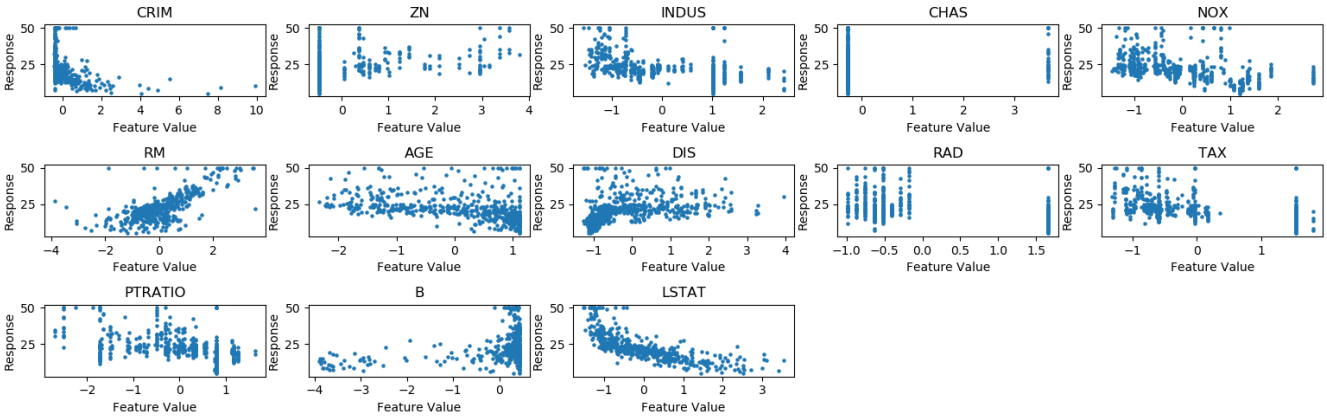
Question 2

Part (b)

```
Features: ['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'
 'B' 'LSTAT']
Number of data points: 506
Dimension of Input data: (506, 13)
Dimension of Input target: (506,)
```

There are 506 inputs, each has 13 features.

Part (c)



Part (e)

Feature	Weight
CRIM	-0.49571412919539803
ZN	0.8078009184917263
INDUS	0.23317507916479385
CHAS	0.788197187601953
NOX	-2.3235337388526793
RM	3.044421867666562
AGE	0.23436663857158768
DIS	-2.7718624579878153
RAD	2.061620608129455
TAX	-1.7324922837606795
PTRATIO	-1.8485609514198187
B	0.8855317251663194
LSTAT	-3.4799377078486544

Yes, the sign matches what I expected. Since INDUS represents proportion of non-retail business acres per town, to me it makes sense that housing price and INDUS will have a positive associative. As higher proportion of non-retail business acres per town (Possibly downtown area) , would lead to higher housing price.

Part (f)

```
Mean Squared Error: 22.688543055804033
Mean Absolute Percentage Error: 16.178473937715655 %
Mean Percentage Error: 0.012197749857354665
```

Part (g)

I chose Mean Absolute Percentage Error and Mean Percentage Error as another two error measuring metric.
Mean absolute percentage error is a good measurement, because it is easy for human to interpret, also because it is robust to outliers.
Mean percentage error is a good measurement, because it is easy for us to interpret and also by using this along with Mean absolute percentage error, we can make a comparison, and know that our underestimate and overestimation actually cancels out pretty well in our model, leaving us with 1.22% mean error.

Part (h)

Feature LSTAT and RM are the most significant features that best describe the price, because they have the highest magnitude out of all. With RM at 3.0444, and LSTAT at -3.4799. Based on our prediction formula $y = \sum_j w_j x_j + b$,

we know that the higher the magnitude the weight w is, the more effect a feature will have on our predicted value y. Therefore LSTAT, RM has the highest effect on our predicted value, and so they are the most significant features out of all.

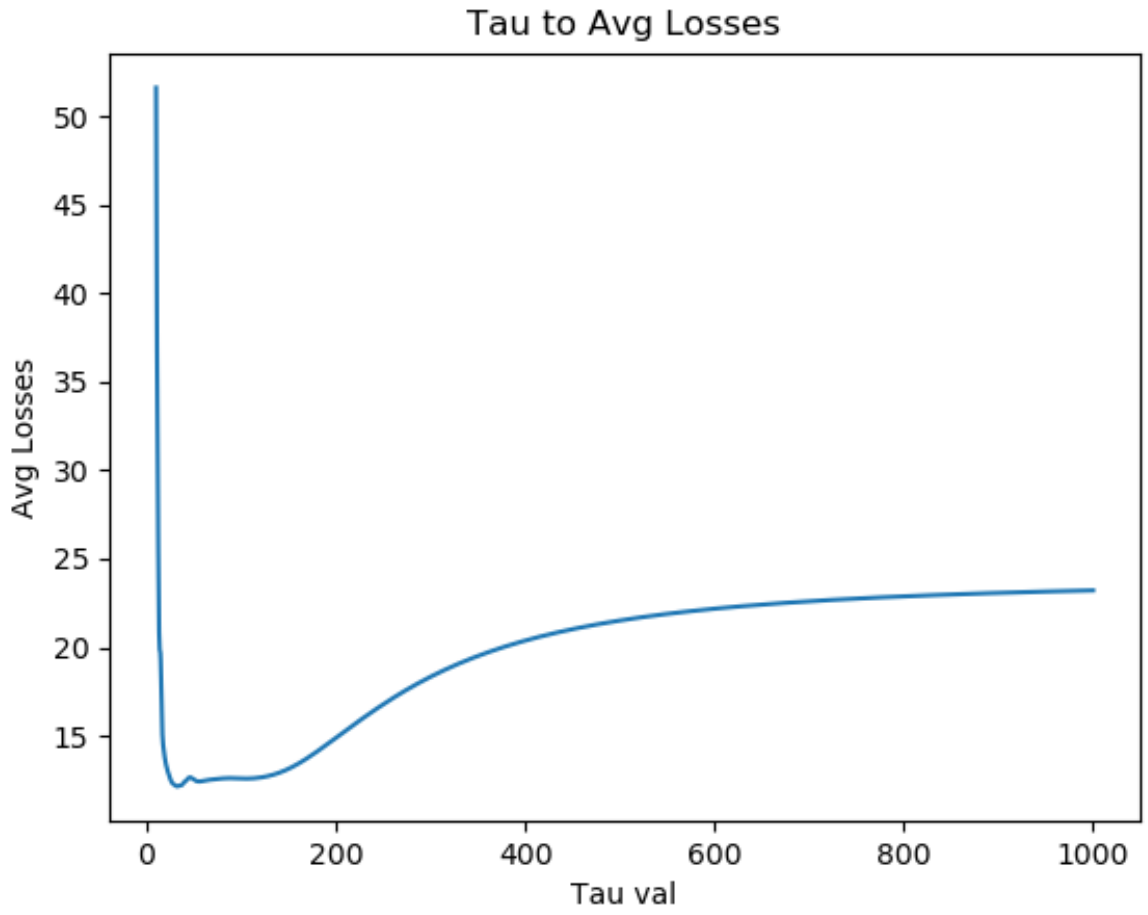
Question 3

Part (a)

To solve $\mathbf{w}^* = argmin \frac{1}{2} \sum_{i=1}^{N} a^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$
We take derivative of it w.r.t. w:

$$\begin{aligned}
\frac{\partial}{\partial w}\frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2 &= \frac{1}{2}\sum_{i=1}^{N} -2\mathbf{x}^{(i)}a^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{N} -\mathbf{x}^{(i)}a^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}) \\
&= \sum_{i=1}^{N} -\mathbf{x}^{(i)}a^{(i)}y^{(i)} + \sum_{i=1}^{N} \mathbf{x}^{(i)}a^{(i)}\mathbf{w}^T\mathbf{x}^{(i)} && \text{Set to 0 for critical point} \\
\sum_{i=1}^{N} \mathbf{x}^{(i)}a^{(i)}y^{(i)} &= \sum_{i=1}^{N} \mathbf{x}^{(i)}a^{(i)}(\mathbf{x}^{(i)})^T\mathbf{w} && \mathbf{w}^T\mathbf{x}^{(i)} == (\mathbf{x}^{(i)})^T\mathbf{w} \\
(\mathbf{X}^T\mathbf{A}\mathbf{X})\mathbf{w} &= \sum_{i=1}^{N} \mathbf{x}^{(i)}a^{(i)}y^{(i)} && \mathbf{x}^{(i)}a^{(i)}(\mathbf{x}^{(i)})^T) \text{ is equivalent to } \mathbf{X}^T\mathbf{A}\mathbf{X} \text{ in matrix form} \\
(\mathbf{X}^T\mathbf{A}\mathbf{X})\mathbf{w} &= \mathbf{X}^T\mathbf{A}\mathbf{y} && \mathbf{x}^{(i)}a^{(i)}\mathbf{y}^{(i)} \text{ is equivalent to } \mathbf{X}^T\mathbf{A}\mathbf{y} \text{ in matrix form} \\
\mathbf{w} &= (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{y} && \text{solve for } \mathbf{w}
\end{aligned}$$

Part (c)



Part (d)

As $\tau \to 0$, we look at points closer and closer to our test datum. And eventually, it would be like kNN algorithm when $k = 1$, our test datum's prediction would be based on its closest neighbour. Which normally would have a very high loss.
As $\tau \to \infty$, it is like kNN algorithm when $k \to \infty$. Which also mean that, our test datum's prediction would be based on every point in training data.
In our experience with kNN algorithm, we know there $k = 1$ overfits, and $k \to \infty$ will underfit. This is the same for τ in our Locally weight least squared regression algorithm.

Part (e)

Advantages of Locally weighted linear regression over Ordinary linear regression:

1. We have more control over our algorithm, we have hyperparameter τ to manipulate in order to find the lowest loss rate.

2. If the relation cannot be captured by a straight line (has a complicated relation between features), locally weighted linear regression would be a better algorithm to use over ordinary linear regression.

Advantages of Ordinary linear regression over Locally weighted linear regression:

1. Ordinary linear regression is less computationally heavy than locally weighted linear regression.

2. Does not require tuning any hyperparameter. Straight forward, easy to understand algorithm.