THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN

DDA 3020

MACHINE LEARNING

Assignment 3 Report

Author:

Ma Kexuan

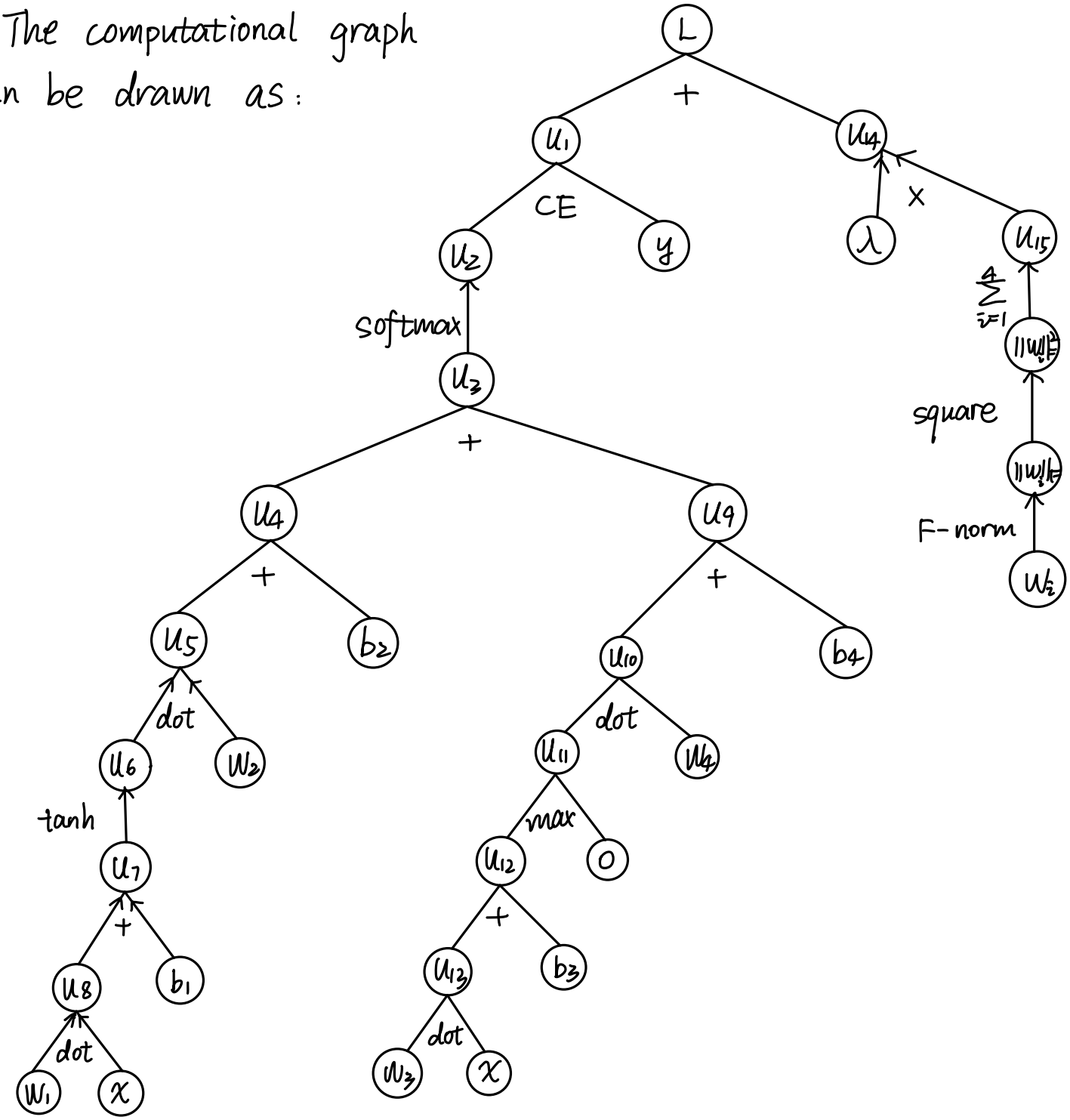
Student Number:

ID 120090651

November 28, 2022

1 Written Problems

1. The computational graph can be drawn as:



$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial u_{14}} \cdot \frac{\partial u_{14}}{\partial u_{15}} \cdot \frac{\partial u_{15}}{\partial \|W\|_F^2} \cdot 2W_1 + \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_4} \cdot \frac{\partial u_4}{\partial u_5} \cdot \frac{\partial u_5}{\partial u_6}$$

$$\cdot \frac{\partial u_6}{\partial u_7} \cdot x$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial u_{14}} \cdot \frac{\partial u_{14}}{\partial u_{15}} \cdot \frac{\partial u_{15}}{\partial \|W\|_F^2} \cdot 2W_2 + \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_4} \cdot \frac{\partial u_4}{\partial u_5} \cdot \frac{\partial u_5}{\partial W_2}$$

$$\frac{\partial L}{\partial W_3} = \frac{\partial L}{\partial u_{14}} \cdot \frac{\partial u_{14}}{\partial u_{15}} \cdot \frac{\partial u_{15}}{\partial \|W\|_F^2} \cdot 2W_3 + \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_9} \cdot \frac{\partial u_9}{\partial u_{10}} \cdot \frac{\partial u_{10}}{\partial u_{11}} \cdot \frac{\partial u_{11}}{\partial u_{12}} \cdot x$$

$$\frac{\partial L}{\partial W_4} = \frac{\partial L}{\partial u_{14}} \cdot \frac{\partial u_{14}}{\partial u_{15}} \cdot \frac{\partial u_{15}}{\partial \|W\|_F^2} \cdot 2W_4 + \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_9} \cdot \frac{\partial u_9}{\partial u_{10}} \cdot \frac{\partial u_{10}}{\partial W_4}$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_4} \cdot \frac{\partial u_4}{\partial u_5} \cdot \frac{\partial u_5}{\partial u_6} \cdot \frac{\partial u_6}{\partial u_7}$$

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_4}$$

$$\frac{\partial L}{\partial b_3} = \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_9} \cdot \frac{\partial u_9}{\partial u_{10}} \cdot \frac{\partial u_{10}}{\partial u_{11}} \cdot \frac{\partial u_{11}}{\partial u_{12}}$$

$$\frac{\partial L}{\partial b_4} = \frac{\partial L}{\partial u_1} \cdot \frac{\partial u_1}{\partial u_2} \cdot \frac{\partial u_2}{\partial u_3} \cdot \frac{\partial u_3}{\partial u_9}$$

2. (1) For $Conv_1$: $N = 100$, $F = 5$, stride $= 3$

We need $N + 2P_1 - F$ / stride to be an integer

$(95 + 2P_1) \bmod 3 = 0 \Rightarrow P_1 = 2$.

For $Maxpool_1$: $N = (100 + 4 - 5)/3 + 1 = 34$, $F = 2$, stride $= 2$

$(34 + 2P_2 - 2) \bmod 2 = 0 \Rightarrow P_2 = 0$

For $Conv_2$: $N = (34 - 2)/2 + 1 = 17$, $F = 3$, stride $= 2$

$(17 + 2P_3 - 3) \bmod 2 = 0 \Rightarrow P_3 = 0$

For $Maxpool_2$: $N = (17 - 3)/2 + 1 = 8$, $F = 3$, stride $= 3$

$(8 + 2P_4 - 3) \bmod 3 = 0 \Rightarrow P_4 = 2$

(2) For $Conv_1$: $(100 + 4 - 5)/3 + 1 = 34$

Shape: $34 \times 34 \times 8$

Parameters: $(5 \times 5 \times 3 + 1) \times 8 = 608$

For $MaxPool_1$: $(34 + 0 - 2)/2 + 1 = 17$

Shape: $17 \times 17 \times 8$

Parameters: 0

For $Conv_2$: $(17 + 0 - 3)/2 + 1 = 8$

Shape: $8 \times 8 \times 16$

Parameters: $(3 \times 3 \times 8 + 1) \times 16 = 1168$

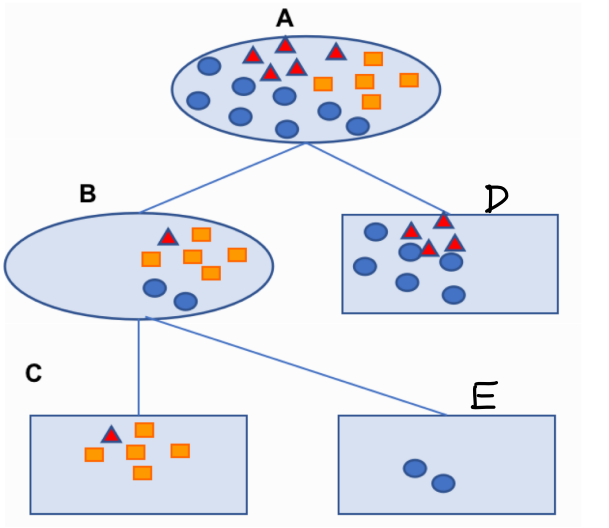
For $MaxPool_2$: $(8+2\times2-3)/3+1=4$

Shape: $4\times4\times16$

Parameters: 0

Total number of parameters: $608+0+1168+0=1776$

3.



A: $P_\triangle=\frac{5}{18}$ $P_\square=\frac{5}{18}$ $P_\circ=\frac{8}{18}=\frac{4}{9}$

Gini index:

$\varphi(p)=1-(\frac{5}{18})^2-(\frac{5}{18})^2-(\frac{4}{9})^2=0.648$

Entropy:

$\varphi(p)=-(\frac{5}{18}\log_2\frac{5}{18}+\frac{5}{18}\log_2\frac{5}{18}+\frac{4}{9}\log_2\frac{4}{9})$

$=1.547$

Classification Error:

$\varphi(p)=1-\max(p_i)=1-\frac{4}{9}=0.556$

B: $P_\triangle=\frac{1}{8}$, $P_\square=\frac{5}{8}$, $P_\circ=\frac{1}{4}$

Gini index:

$\varphi(p)=1-(\frac{1}{8})^2-(\frac{5}{8})^2-(\frac{1}{4})^2=0.531$

Entropy:

$\varphi(p)=-(\frac{1}{8}\log_2\frac{1}{8}+\frac{5}{8}\log_2\frac{5}{8}+\frac{1}{4}\log_2\frac{1}{4})$

$=1.299$

Classification Error:

$\varphi(p)=1-\max(p_i)=1-\frac{5}{8}=0.375$

C: $P_\triangle=\frac{1}{6}$ $P_\square=\frac{5}{6}$

Gini index:

$\varphi(p)=1-(\frac{1}{6})^2-(\frac{5}{6})^2=0.278$

Entropy:

$\varphi(p)=-(\frac{1}{6}\log_2\frac{1}{6}+\frac{5}{6}\log_2\frac{5}{6})$

$=0.650$

Classification Error:

$\varphi(p)=1-\max(p_i)=0.167$

D: $P_\triangle=\frac{4}{10}=\frac{2}{5}$, $P_\circ=\frac{3}{5}$

Gini index:

$\varphi(p)=1-(\frac{2}{5})^2-(\frac{3}{5})^2=0.48$

Entropy:

$\varphi(p)=-(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5})$

$=0.971$

Classification Error:

$\varphi(p)=1-\max(p_i)=\frac{2}{5}$

E: $P_\circ=1$

Gini index: $1-1^2=0$

Entropy: $-1\log_2 1=0$

Classification Error: $1-1=0$

4. (a) $\hat{MSE} = \frac{1}{10}(1^2+1^2+2^2+2^2+2^2+2^2+3^2+1^2+2^2+3^2) = 4.1$

Avg prediction $= \frac{1}{10}(6+8+9+5+9+5+4+8+9+4) = 6.7$

$\text{Bias}^2 = (\bar{h}_D(x) - t(x))^2 = (6.7-7.2)^2 = 0.25$

$\text{Variance} = \frac{1}{10}(0.7^2+1.3^2+2.3^2+1.7^2+2.3^2+1.7^2+2.7^2+1.3^2+2.3^2+2.7^2)$

$= 4.01$

(b) $\hat{MSE}(x,y) = \frac{1}{10}\sum_{i=1}^{10}(h_{D_i}(x)-y)^2$

$= \frac{1}{10}\sum_{i=1}^{10}(h_{D_i}(x) - \bar{h}(x) + \bar{h}(x) - y)^2$

$= \frac{1}{10}\sum_{i=1}^{10}\left[(h_{D_i}(x)-\bar{h}(x))^2 + 2(h_{D_i}(x)-\bar{h}(x))(\bar{h}(x)-y) + (\bar{h}(x)-y)^2\right]$

$= \frac{1}{10}\left[\sum_{i=1}^{10}(h_{D_i}(x)-\bar{h}(x))^2 + 2(h(x)-y)\sum_{i=1}^{10}(h_{D_i}(x)-\bar{h}(x)) + 10\times(\bar{h}(x)-y)^2\right]$

$= \frac{1}{10}\sum_{i=1}^{10}(h_{D_i}(x)-\bar{h}(x))^2 + (\bar{h}(x)-y)^2$

$= \frac{1}{10}\sum_{i=1}^{10}(h_{D_i}(x)-\bar{h}(x))^2 + (\bar{h}(x)-t(x))^2 + 2(\bar{h}(x)-t(x))(t(x)-y) +$

$(t(x)-y)^2$

$= \text{Variance} + \text{Bias}^2 + \varepsilon^2 + 2(\bar{h}(x)-t(x))(t(x)-y)$

Since $\hat{MSE} = \text{Variance} + \text{Bias}^2 + \varepsilon^2 + 2(\bar{h}(x)-t(x))(t(x)-y)$,

$\varepsilon^2 + 2(\bar{h}(x)-t(x))(t(x)-y) = 0.04 - 0.2 = -0.16 \neq 1 = \sigma^2$

So we conclude that $\hat{MSE} \neq \text{Variance} + \text{Bias}^2 + \sigma^2$ under these 10 models.

5. $\sigma(a) = \frac{1}{1+e^{-a}}$, $\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$

$1 - 2\sigma(a) = \frac{1+e^{-a}}{1+e^{-a}} - \frac{2}{1+e^{-a}} = \frac{e^{-a}-1}{1+e^{-a}} = -\frac{1-e^{-a}}{1+e^{-a}}$

$$= -\frac{e^{\frac{a}{2}} - e^{-\frac{a}{2}}}{e^{\frac{a}{2}} + e^{-\frac{a}{2}}} = -\tanh\left(\frac{a}{2}\right)$$

Thus, we obtain that $1 - 2\sigma(a) = -\tanh\left(\frac{a}{2}\right)$

$$\Rightarrow \tanh(a) = 2\sigma(2a) - 1$$

$$\hat{y}_k(x, \hat{w}) = \sigma\left(\sum_{j=1}^{M} \hat{w}_{kj}^{(2)} \tanh\left(\sum_{i=1}^{D} \hat{w}_{ji}^{(1)} x_i + \hat{w}_{j0}^{(1)}\right) + \hat{w}_{k0}^{(2)}\right)$$

$$= \sigma\left(\sum_{i=1}^{M} \hat{w}_{kj}^{(2)} \cdot \left[2h\left(2\sum_{i=1}^{D} \hat{w}_{ji}^{(1)} x_i + 2\hat{w}_{j0}^{(1)}\right) - 1\right] + \hat{w}_{k0}^{(2)}\right)$$

$$= \sigma\left(\sum_{i=1}^{M} 2\hat{w}_{kj}^{(2)} h\left(\sum_{i=1}^{D} 2\hat{w}_{ji}^{(1)} x_i + 2\hat{w}_{j0}^{(1)}\right) - \sum_{i=1}^{M} \hat{w}_{kj}^{(2)} + \hat{w}_{k0}^{(2)}\right)$$

Compare it with the original $y_k(x, w)$, we can get:

$$w_{kj}^{(2)} = 2\hat{w}_{kj}^{(2)}$$

$$w_{ji}^{(1)} = 2\hat{w}_{ji}^{(1)}$$

$$w_{j0}^{(1)} = 2\hat{w}_{j0}^{(1)}$$

$$w_{k0}^{(2)} = \hat{w}_{k0}^{(2)} - \sum_{i=1}^{M} \hat{w}_{kj}^{(2)}$$

Thus, there exists linear transformation between these w, $\hat{w}$, that enable $y_k(x, w) = \hat{y}_k(x, \hat{w})$ for all x.

2 Programming

2.1 Decision Tree

In this question, we implement Decision Tree, Bagging of Trees and Random Forests to predict the sale of the Carseats dataset. All the algorithms can be implemented by sklearn, the loss will be set as MSE.

2.1.1 Data Statistics

There are 11 columns of the dataset (including sales), 8 of them are numeric type, 3 of them are string type ('ShelveLoc', 'Urban', 'US').

The details of the numerical data are as follows:

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	7.496325	124.975000	68.657500	6.635000	264.840000	115.795000	53.322500	13.900000
std	2.824115	15.334512	27.986037	6.650364	147.376436	23.676664	16.200297	2.620528
min	0.000000	77.000000	21.000000	0.000000	10.000000	24.000000	25.000000	10.000000
25%	5.390000	115.000000	42.750000	0.000000	139.000000	100.000000	39.750000	12.000000
50%	7.490000	125.000000	69.000000	5.000000	272.000000	117.000000	54.500000	14.000000
75%	9.320000	135.000000	91.000000	12.000000	398.500000	131.000000	66.000000	16.000000
max	16.270000	175.000000	120.000000	29.000000	509.000000	191.000000	80.000000	18.000000

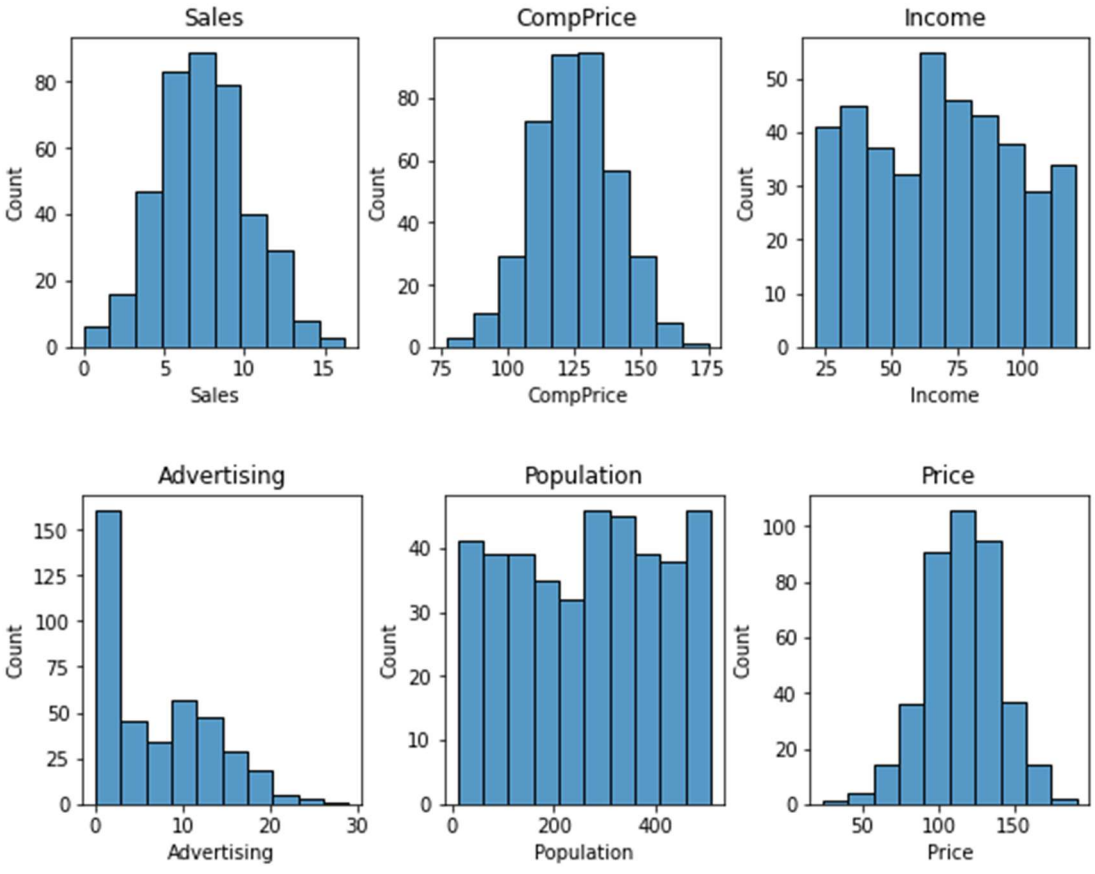
The details of the categorical data are as follows:

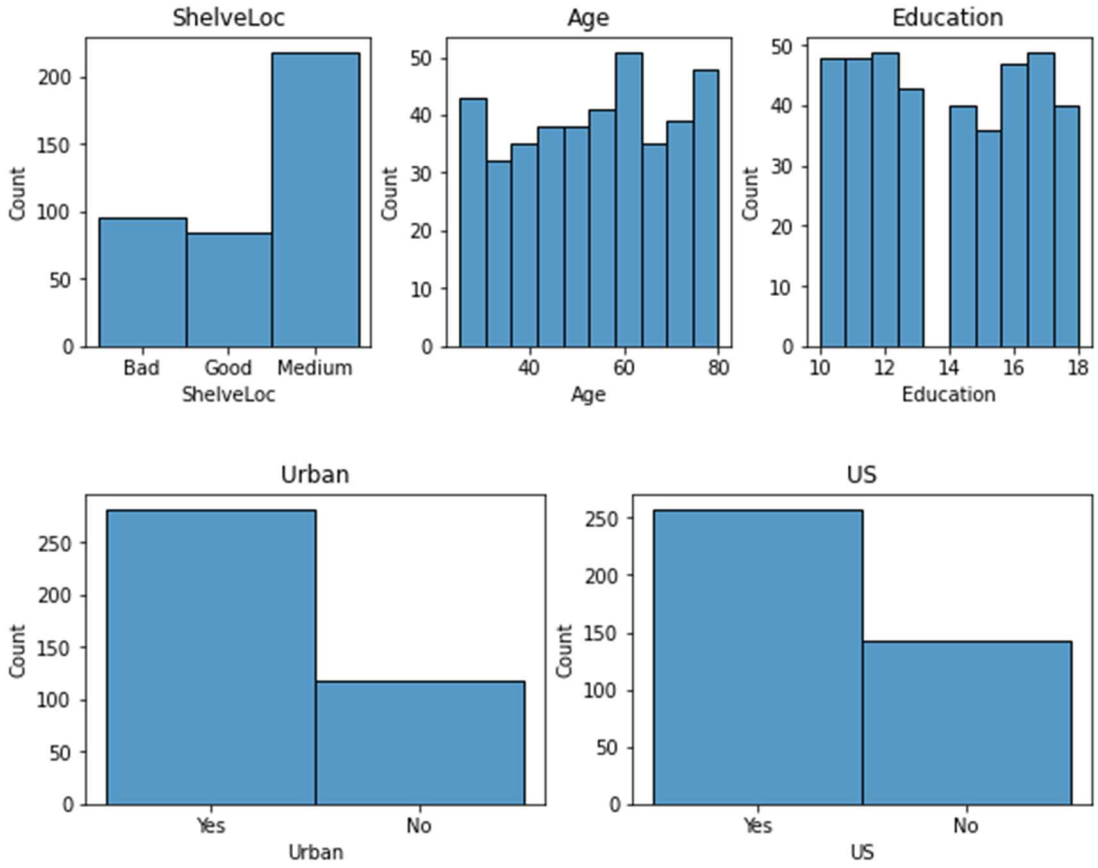
```
Medium    219
Bad        96
Good       85
Name: ShelveLoc, dtype: int64
```

```
Yes    282
No     118
Name: Urban, dtype: int64
```

```
Yes    258
No     142
Name: US, dtype: int64
```

Also, I draw the histograms of each column, and use seaborn to visualize them.





Data Preprocessing:

1. Since there're 3 categorical variables, we cannot use it directly to do regressions. It's straightforward to think about one-hot encoding in this scenario since each of the three categorical variables has less than or equal to 3 categories. Thus, I use one-hot encoding to change the variable. There's also no need to do normalization since decision tree is not sensitive to the scale of the numerical data.
2. For the training and testing data, I choose the first option, that is, I use the first 300 rows as the training set, and the remaining 100 rows as the testing set.

2.1.2 Decision Tree

First, I use GridSearchCV offered by sklearn to search for the best maximum depth and least node size. The result is as follows:

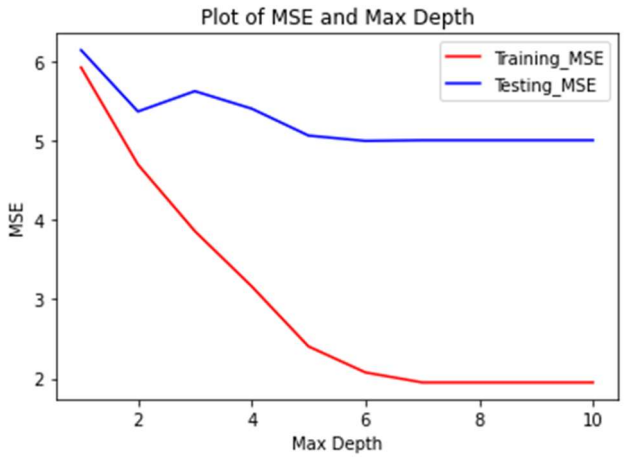
```
Best Parameter Pairs:  {'max_depth': 8, 'min_samples_leaf': 9}
```

Then, I define a function decision_tree (maxdepth, minleafsamples), which use DecisionTreeRegressor in sklearn to do the task and return its training and testing mse by metrics.mean_squared_error ().

Result Analysis:

1. First, I change maximum depth from 1 to 10, while maintaining least node sizes as 9, according to the GridSearchCV.

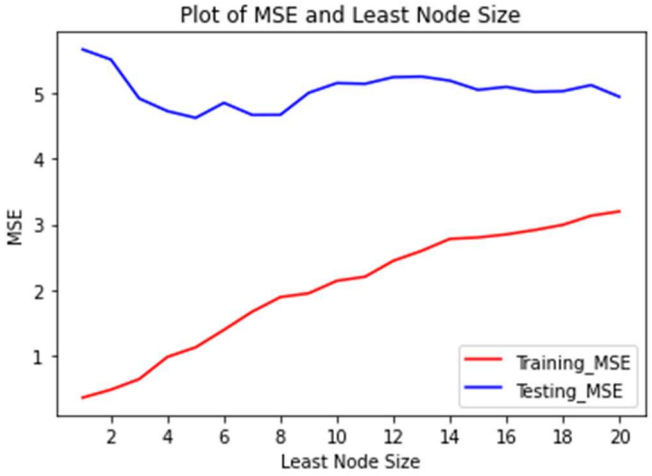
The result is shown below:



From this figure, we find that with the increase of max depth, which means that model complexity is increasing, the train error always decreases. The test error decrease at first and start be stable when the depth is around 5. It proves that when the complexity is too small, the model will be underfitted, both training and testing error is high. When the complexity is too large, it tends to be overfitted. The training error is smaller when increasing the max depth, but the testing error is stable.

2. Then I change least node size from 1 to 20, while maintaining maximum depth as 8, according to the GridSearchCV.

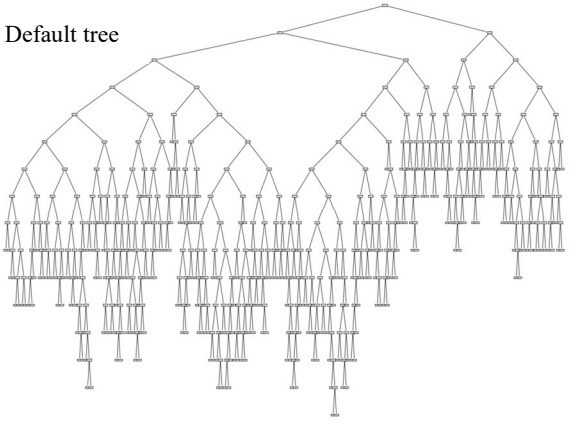
The result is shown below:



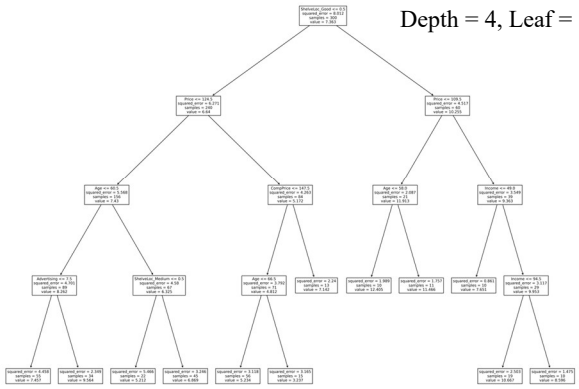
From this figure, we find that with the increase of node size, which means that model complexity is decreasing, the training error always increases. The testing error decreases first and then increases. This proves that when the node size is small, the model tends to be overfitted, when the node size is large, the model tends to be underfitted.

Below are some of the Decision Tree plots, each of which has different set of parameters. (If the plot shown below is hard to recognize, you can directly find them in the assignment folder, whose names are p1.png, p2.png, etc. They will be much clearer.)

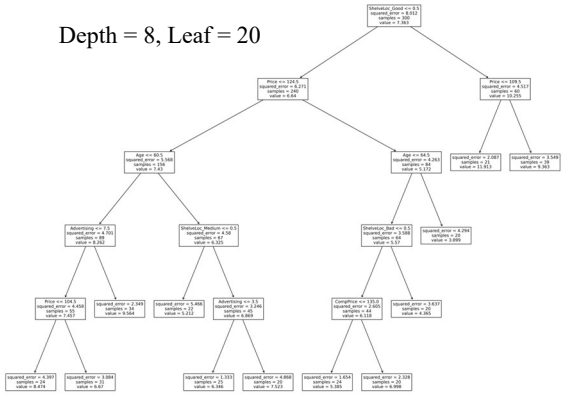
Default tree



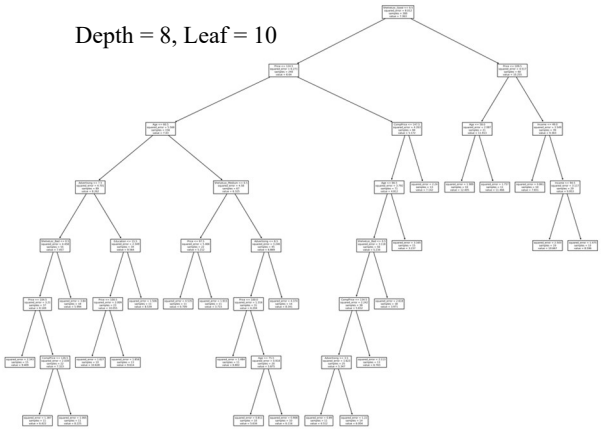
Depth = 4, Leaf = 10



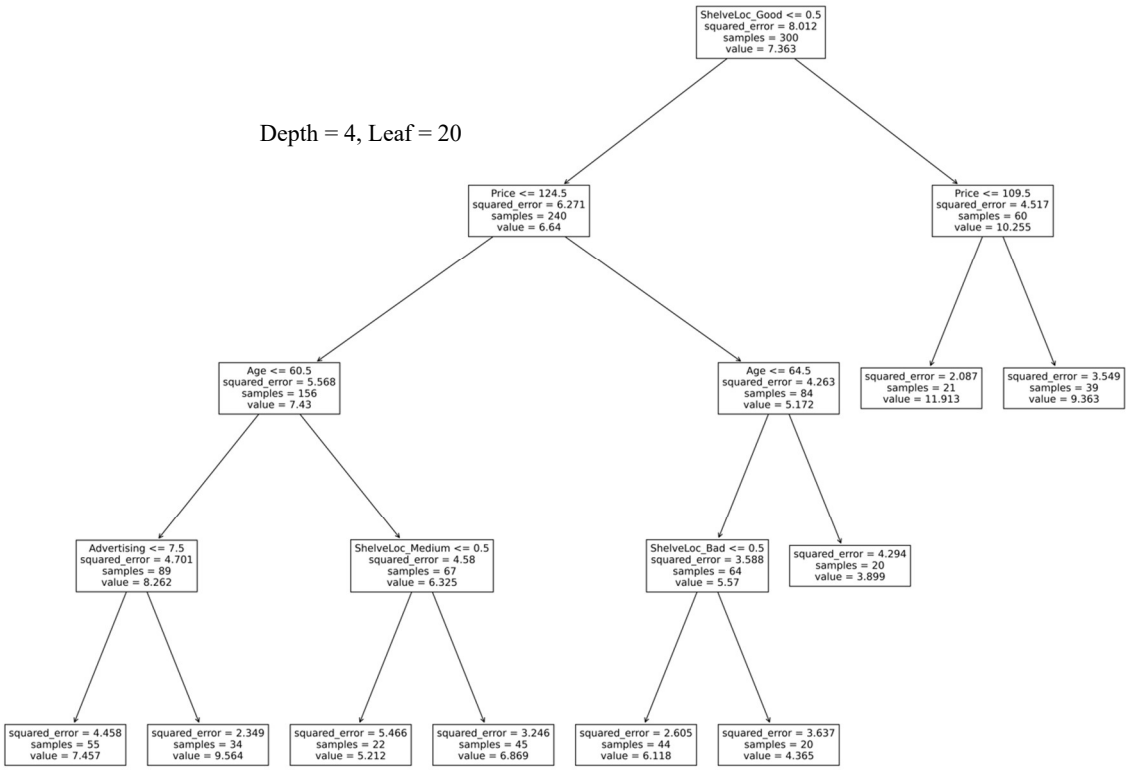
Depth = 8, Leaf = 20



Depth = 8, Leaf = 10



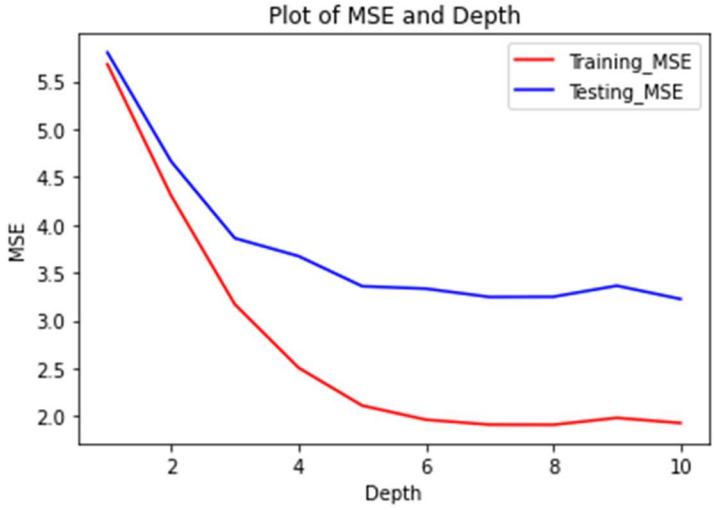
Depth = 4, Leaf = 20



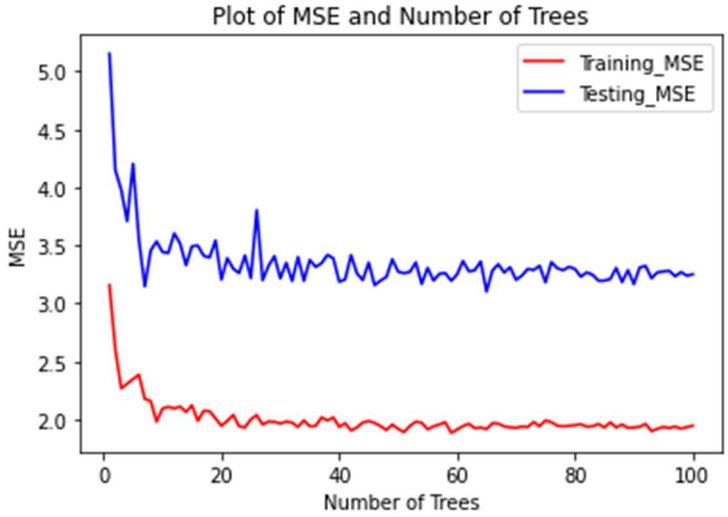
2.1.3 Bagging of Trees

First, I define a function bagging (depth, numtrees), which use BaggingRegressor in the outer space and DecisionTreeRegressor in the inner space from sklearn to do the task and

return its training and testing mse by metrics.mean_squared_error ().
Result Analysis:
1. First, I change maximum depth from 1 to 10, while maintaining the number of trees as 50. The result is shown below:



From this figure, we find that with the increase of depth, which means that model complexity is increasing, the training error always decreases. The testing error decreases at first and start to be stable when the depth is larger than 6. It proves that when the depth is too small, the model will be underfitted, when the depth is too large, the model might be overfitted. (E.g. Depth = 100)
2. Then, I change the number of trees from 1 to 100, while maintaining the maximum depth as 6. The result is shown below:



From this figure, we find that when the number of trees increase, the training error and testing error start to decrease at first and then be stable. The number of trees means data-level randomness, has no relationship with complexity since each tree may have roughly the same depth.

2.1.4 Random Forests

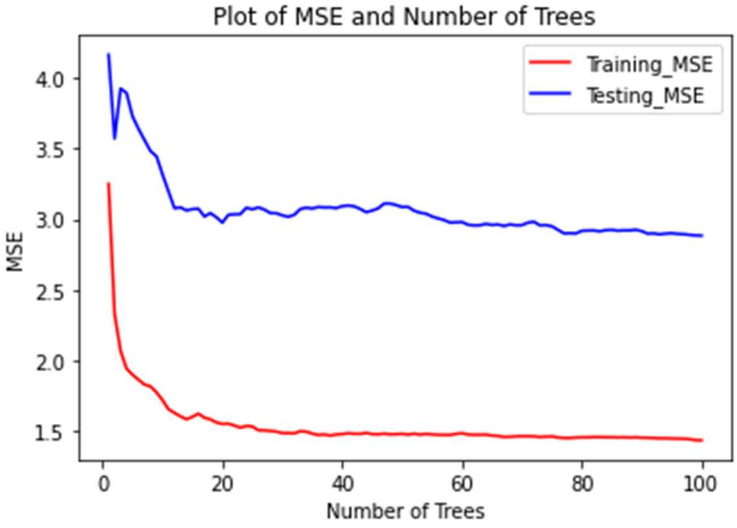
First, I use GridSearchCV offered by sklearn to search for the best number of trees, value of m, maximum depth and least node size. The result is as follows:

```
Best Parameter Pairs:  {'max_depth': 7, 'max_features': 7, 'min_samples_leaf': 5, 'n_estimators': 20}
```

Then, I define a function random_forests (numtrees, m), which use RandomForestRegressor in sklearn to do the task and return its training and testing mse by metrics.mean_squared_error ().

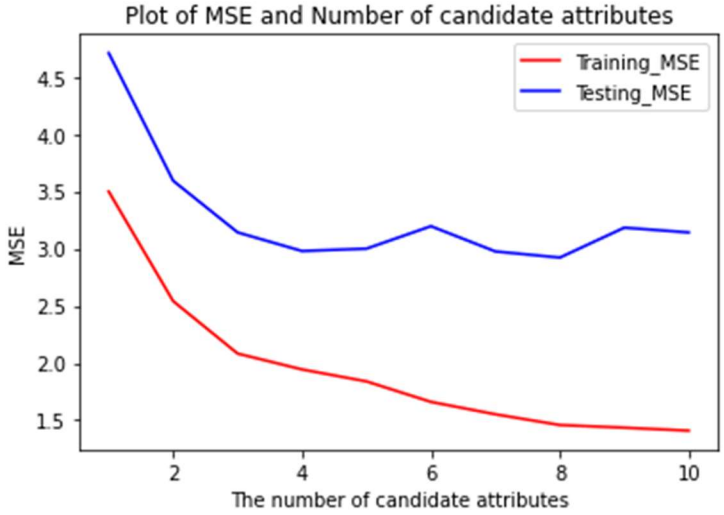
Result Analysis:

1. First, I change the number of trees from 1 to 100, while setting max_depth as 7, m as 7, min_samples_leaf as 5, according to the GridSearchCV. The result is shown below:



From this figure, we find that when the number of trees increase, the training error and testing error start to decrease at first and then be stable. The number of trees means data-level randomness, has no relationship with complexity since each tree may have roughly the same depth in the whole random forest, for example, they may have the depth of only 2 or 3.

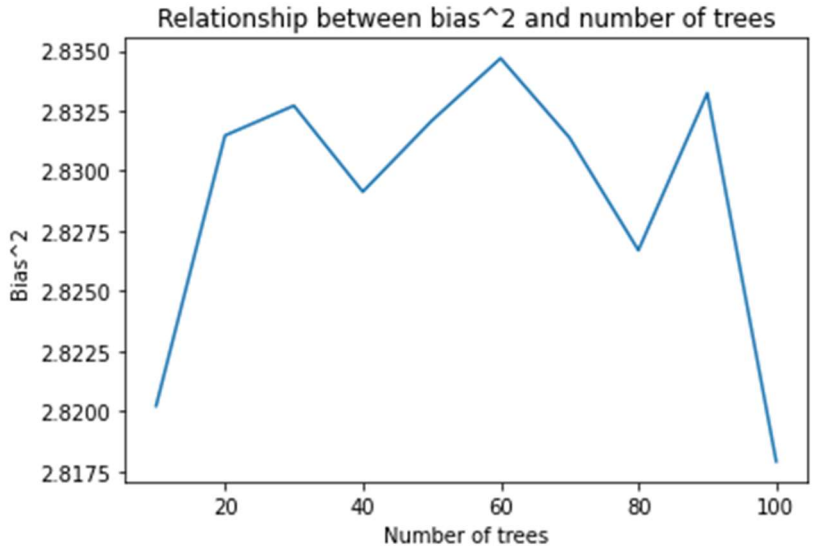
2. Then, I change value of m from 1 to 100, while setting max_depth as 7, number of trees as 20, min_samples_leaf as 5, according to the GridSearchCV. The result is shown below:



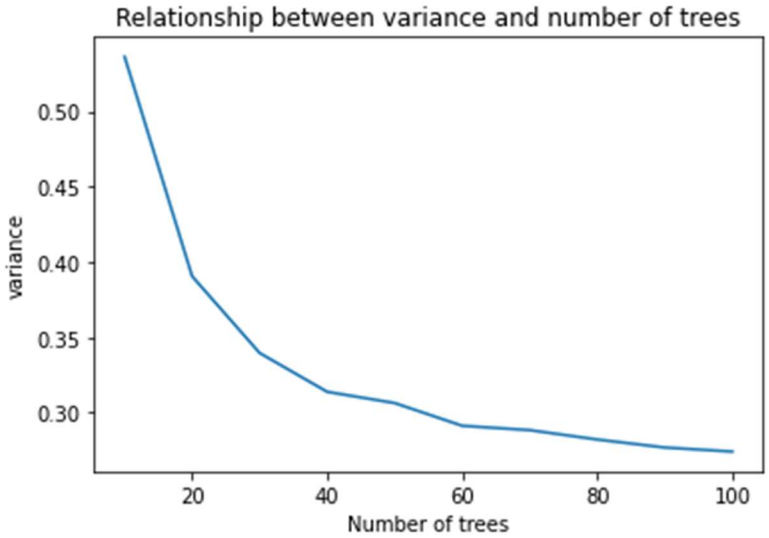
From this figure, we find that with the increase of candidate attributes, the training error is always decreasing. The testing error decreases at first and start to be stable when m = 3 or 4, which is same as we discussed in lecture slides that m = N/3 is always a good choice.

2.1.5 Bias^2 and Variance

In this question, I use bias_variance_decomp () provided by mlxtend to get the average bias and average variance throughout the whole dataset.
Below is the curve of bias w.r.t the number of trees:



From the figure, we can obtain that there's no obvious relationship between bias and the number of trees, which is consistent with the statement provided in lecture slides that different trees in the random forests are independent and the overall model complexity is not increased, so we can't expect increasing the number of trees can reduce bias.
However, for the curve of variance and the number of trees, it's different:



From the figure, we can see that increasing the number of trees can indeed decrease the variance. Since random forests introduce both data-level randomness and model-level randomness (From random feature selection), so there is a guarantee that variance will be reduced.
Mathematical Proof:
We can write a random forest as the formula below:

$$T(x) = \frac{1}{B}\sum_{i=1}^{B} T_{i,Z_i}(x)$$

where B is the number of trees, T_{i,Z_i} is the i-th decision tree, Z_i is the training sample.

We can easily obtain that for a certain x, $T_{i,Z_i}(x)$ follows the same distribution, so increasing the number of trees has no contribution to reduce the model bias.

However, $T_{i,Z_i}(x)$ is dependent to each other, since the samples are drawn from the same dataset, and they use the same algorithm to get each decision tree. If we set the correlation coefficient to be ρ, and variance of each tree to be σ^2, we can get the variance of the whole random forest:

$$Var(T(x)) = [\frac{1}{B}, \cdots, \frac{1}{B}] \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{B} \\ \vdots \\ \frac{1}{B} \end{bmatrix} \sigma^2$$
$$= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

From the formula derived, we can see that if we increase the number of trees B, we can significantly reduce the variance of the random forest. In conclusion, our conclusions from processing the dataset and mathematical formulation are the same.

2.2 Handwritten Digit Recognition

In this question, I use load_mnist.py to load the train/test data, and use MLPClassifier from sklearn to construct a Multi-Layer Perceptron neural network, with hyperparameters: (solver = 'adam', activation = 'relu', alpha = 0.05, random_state = 3020, max_iter = 100, verbose = False, learning_rate_init = 3e-4, batch_size = 512), which are figured out by GridSearchCV. The result is shown below:

	n hidden nodes	n hidden layers	score
0	50.0	1.0	0.9456
1	200.0	1.0	0.9617
2	784.0	1.0	0.9718
3	50.0	2.0	0.9478
4	200.0	2.0	0.9524
5	784.0	2.0	0.9689
6	50.0	3.0	0.9577
7	200.0	3.0	0.9714
8	784.0	3.0	0.9747

From the chart, we can see that if we fix the number of hidden nodes, with the increase of the number of hidden layers, we may not have a higher prediction accuracy. In contrast, if we fix the number of hidden layers, with the increase of the number of hidden nodes, we would have a higher prediction accuracy.