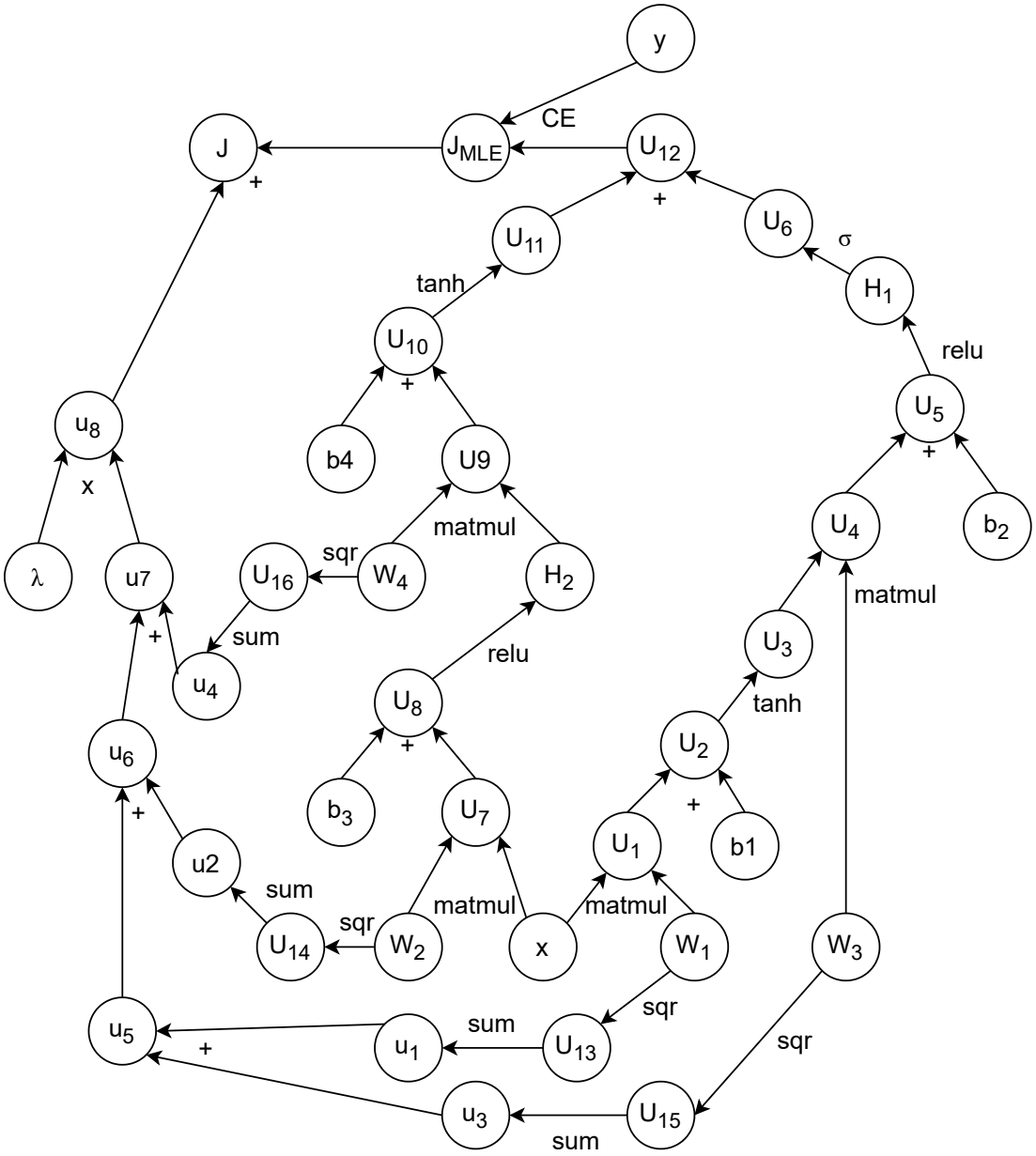
Assignment3 Report

Author: 余永琦 Student Number: ID 120090761

Contents

Part1: Written Questions

Problem1
computational graph



1. Update procedure:

Forward pass: Starting from the lowest level x, we calculate its parent function, according to the graph, we need to do:

$V_1 = W_1 x$, $V_2 = V_1 + b_1$, $V_3 = \tanh V_2$ .... keep moving forward and we can get our cost function $J = J_{MLE} + U_8$.

This is the forward pass step.

Back ward pass: Starting from J, we calculate the node's derivative with respect to each of its children node, un-til we caver all nodes. and store them.

In this graph, we need to calculate:

$\frac{\partial J}{\partial U_8}, \frac{\partial J}{\partial J_{MLE}}, \frac{\partial J_{MLE}}{\partial y}, \frac{\partial J_{MLE}}{\partial V_{12}}, \dots, \frac{\partial V_7}{\partial x}, \frac{\partial V_1}{\partial x}, \frac{\partial V_1}{\partial W_1}$, cover all nodes.

Then, we now know all the derivatives of each node w.r.t. its children nodes.

By the chain rule, we can get:

~~$\frac{\partial L}{\partial W_1} = \frac{\partial J}{\partial J_{MLE}} \cdot \frac{\partial J_{MLE}}{\partial V_{12}} \cdot \frac{\partial V_{12}}{\partial V_{11}} \cdot \frac{\partial V_{11}}{\partial V_{10}} \cdot$~~

$\frac{\partial L}{\partial W_1} = \frac{\partial J}{\partial J_{MLE}} \cdot \frac{\partial J_{MLE}}{\partial V_{12}} \cdot \frac{\partial V_{12}}{\partial V_6} \cdot \frac{\partial V_6}{\partial H_1} \cdot \frac{\partial H_1}{\partial V_5} \cdot \frac{\partial V_5}{\partial V_4} \cdot \frac{\partial V_4}{\partial V_3} \cdot \frac{\partial V_3}{\partial V_2} \cdot \frac{\partial V_2}{\partial V_1} \cdot \frac{\partial V_1}{\partial W_1} + \frac{\partial J}{\partial U_8} \cdot \frac{\partial U_8}{\partial U_7} \cdot \frac{\partial U_7}{\partial U_6} \cdot \frac{\partial U_6}{\partial U_5} \cdot \frac{\partial U_5}{\partial U_4} \cdot \frac{\partial U_4}{\partial V_{13}} \cdot \frac{\partial V_{13}}{\partial W_1}$

Since we've known all these derivatives, we can just mutiply them to get $\frac{\partial L}{\partial W_1}$.

By doing the same way, we can easily get $\frac{\partial L}{\partial W_2}, \frac{\partial L}{\partial W_3}, \frac{\partial L}{\partial W_4}, \frac{\partial L}{\partial b_1} \dots \frac{\partial L}{\partial b_4}$

Then, we know the gradient of L, use gradient descent to update the parameters.

2.

Layer 1: $\frac{63+2\times2-5}{2}+1=32$

Shape: $32\times32\times10$, parameters: W: $5\times5\times3\times10=750$, bais: 10, total: 760.

computational cost: one time: 75-dimension dot product + bais: $75+74+1=150$

total: $150\times32\times32\times10=1536000$

Layer 2: $\frac{32+0-2}{3}+1=11$

Shape: $11\times11\times10$, parameters: none.

computational cost: one time: $2\times2=4$.

total: $11\times11\times10\times4=4840$.

Layer 3: $\frac{11+2\times1-3}{2}+1=6$

Shape: $6\times6\times20$. parameters: W: $3\times3\times10\times20=1800$, bias: 20, total: 1820.

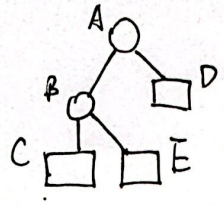
computational cost: $(90+89+1)\times6\times6\times20=129600$.

Layer 4: $\frac{6+1\times2-2}{2}+1=4$

Shape: $4\times4\times20$. parameters: none.

computational cost: $4\times4\times4\times20=1280$

3.



A: Entropy: $-\frac{4}{9}\log\frac{4}{9} - \frac{5}{18}\log\frac{5}{18} - \frac{5}{18}\log\frac{5}{18} \approx 1.5466$

Gini index: $1-(\frac{4}{9})^2-(\frac{5}{18})^2-(\frac{5}{18})^2 \approx 0.6481$

Misclassification error: $1-\frac{4}{9}=\frac{5}{9}\approx 0.5556$

B: Entropy: $-\frac{5}{8}\log\frac{5}{8} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} \approx 1.0884$

Gini index: $1-(\frac{5}{8})^2-(\frac{1}{4})^2-(\frac{1}{8})^2 \approx 0.5313$

Misclassification error: $1-\frac{5}{8}=\frac{3}{8}=0.375$

C: Entropy: $-\frac{5}{6}\log\frac{5}{6} - \frac{1}{6}\log\frac{1}{6} \approx 0.2837$

Gini index: $1-(\frac{5}{6})^2-(\frac{1}{6})^2 \approx \frac{5}{18}\approx 0.2778$

Misclassification error: $1-\frac{5}{6}=\frac{1}{6}\approx 0.1667$

D: Entropy: $-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} \approx 0.9710$

Gini index: $1-(\frac{3}{5})^2-(\frac{2}{5})^2=\frac{12}{25}=0.48$

Misclassification error: $1-\frac{3}{5}=\frac{2}{5}=0.4$

E: Entropy: $-1\log 1=0$

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

Misclassification error: $1-1=0$.

4.

(a) $\bar{h}(x) = \frac{1}{10}(6+8+9+5+10+5+4+8+9+3) = 6.7.$

$\widehat{MSE}(x,y) = \frac{1}{10}[(6-7)^2+(8-7)^2+(9-7)^2+(5-7)^2+(10-7)^2+(5-7)^2+(4-7)^2+(8-7)^2+(9-7)^2+(3-7)^2]$

$= 5.3.$

$Bias^2 = (6.7-6.7)^2 = 0$

$Variance = \frac{1}{10}[(6-6.7)^2+(8-6.7)^2+(9-6.7)^2+(5-6.7)^2+(10-6.7)^2+(5-6.7)^2+(4-6.7)^2+(8-6.7)^2+(9-6.7)^2+(3-6.7)^2]$

$= 5.21$

(b) In this question:

$$\widehat{MSE}(x,y) = \frac{1}{10}\sum_{i=1}^{10}(h_D(x)-y)^2 = \frac{1}{10}\sum_{i=1}^{10}(h_D(x)-\bar{h}(x)+\bar{h}(x)-y)^2$$

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

Here, we know $\frac{1}{10}\sum_{i=1}^{10} h_D(x)-\bar{h}(x)=0$, and $\bar{h}(x)-y=-0.3$, a constant

$\therefore \frac{1}{10}\sum_{i=1}^{10}2(h_D(x)-\bar{h}(x))(\bar{h}(x)-y)=0$

$\therefore \widehat{MSE}(x,y) = \frac{1}{10}\sum_{i=1}^{10}(h_D(x)-\bar{h}(x))^2 + (\bar{h}(x)-y)^2$.

For the term $\frac{1}{10}\sum_{i=1}^{10}(\bar{h}(x)-y)^2$, we can do the transformation:

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

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

To achieve the equation $MSE = Bias^2 + Variance + \epsilon^2$,

We need to make two assumptions here:

① $\frac{1}{10}\sum_{i=1}^{10}2(\bar{h}(x)-t(x))(t(x)-y)$ by letting $E_{y|x}(y)=t(x)$.

~~However, it doesn't equals to 0 here.~~ However, they doesn't equal here.

But luckily, $\bar{h}(x)-t(x)=0$ here, so the term is still 0,

② $E[\epsilon^2] = \sigma^2$.

However, in this question, $\frac{1}{10}\sum_{i=1}^{10}\epsilon^2 = 0.09 \neq \sigma^2 = 0.5$.

The instantiated ϵ can not well represent the distribution here.

∴ So it cause $\widehat{MSE}(x,y) \neq Bias^2 + Variance + \epsilon^2$,

~~Instead~~ instead, in this question, $\widehat{MSE}(x,y) = Bias^2 + Variance + \epsilon^2$.

$$5.3 = 0 + 5.21 + 0.09.$$

5. $\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} = \frac{1 - e^{-2a}}{e^{-2a} + 1} = \frac{2}{e^{-2a}+1} - \frac{1+e^{-2a}}{e^{-2a}+1} = \frac{2}{e^{-2a}+1} - 1$

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

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

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

We can find a linear transformation between w and $\hat{w}$:

$\hat{w}_{ji}^{(1)} = \frac{1}{2} w_{ji}^{(1)}$ for all j, i. $\hat{w}_{j0}^{(1)} = \frac{1}{2} w_{j0}$ for all j.

$\hat{w}_{kj}^{(2)} = \frac{1}{2} w_{kj}^{(2)}$ for all j and $\hat{w}_{k0}^{(2)} = w_{k0}^{(2)} + \sum_{j=1}^{M}\hat{w}_{kj}^{(2)}$,

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

$= y_k(x, w)$

So we've proved that there exists linear transformation between these w and $\hat{w}$, that enable $y_k(x, w) = \hat{y}_k(x, \hat{w})$ for all x.

Part2: Programming Questions

Data Preprocessing

In the dataset, there are 3 features “ShelveLoc”, “Urban” and “US” that are categorical data that can not directly be used for the models. So we need to encode them to the one we can use.

For “Urban” and “US”, there are only two kind of data, “Yes” and “No”. So we can transform them into a binary set: “Yes” = 1, “No” = -1.

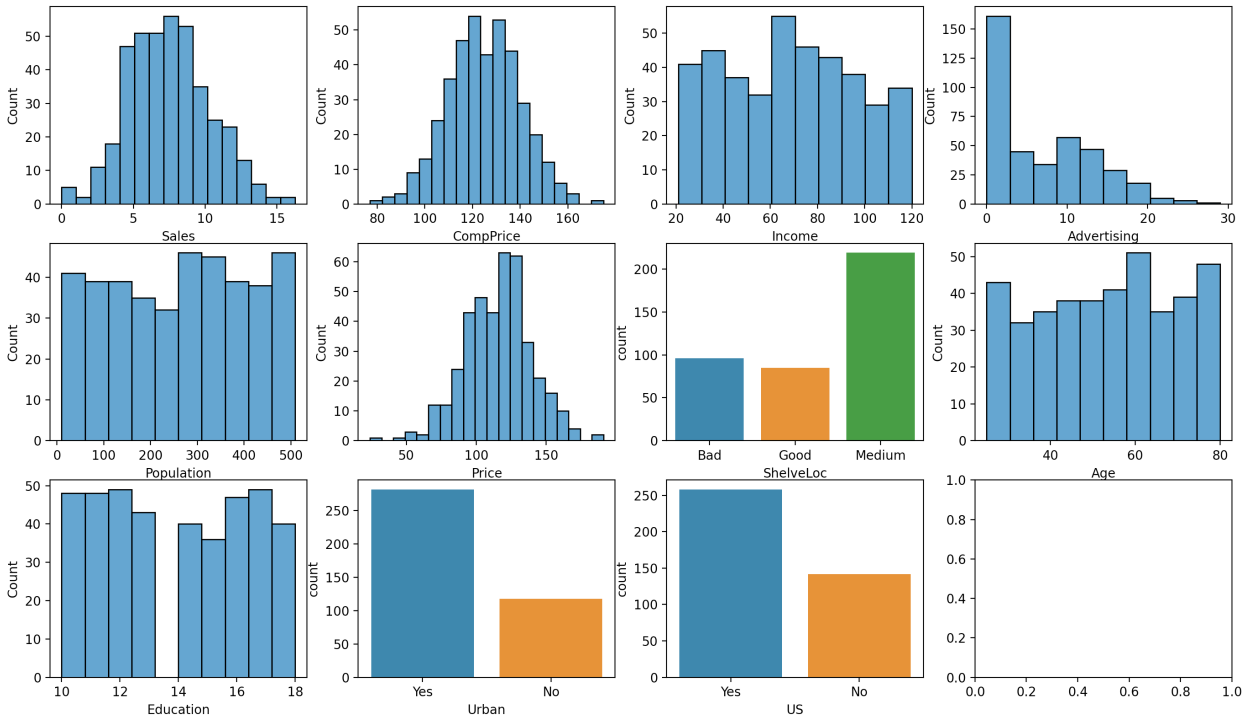
For the feature “ShelveLoc”, there are three kind of data, so consider the one hot encoding. That is: “Good” = [1,0,0], “Medium” = [0,1,0], “Bad” = [0,0,1].

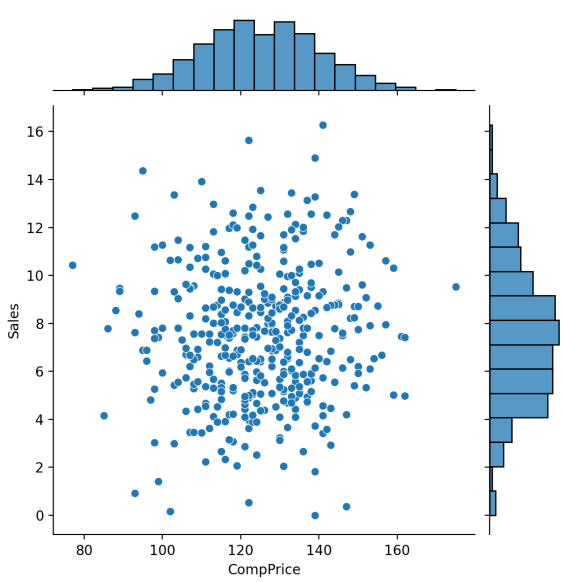
*** To make the result reproducible, set all random_state to 12.**

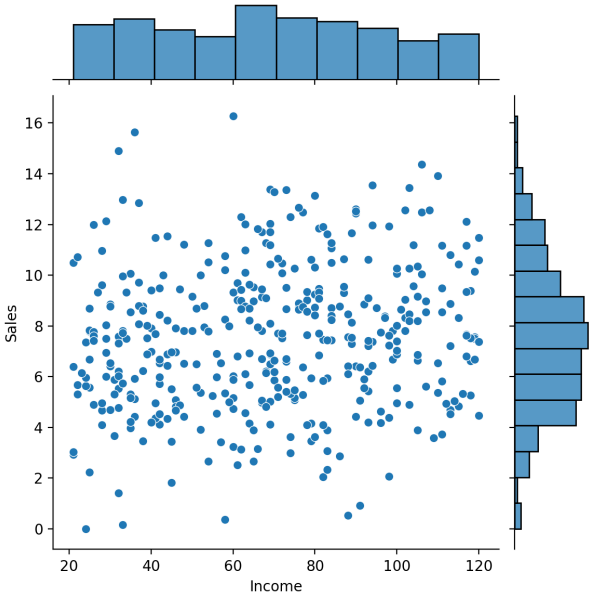
Train/Test Split

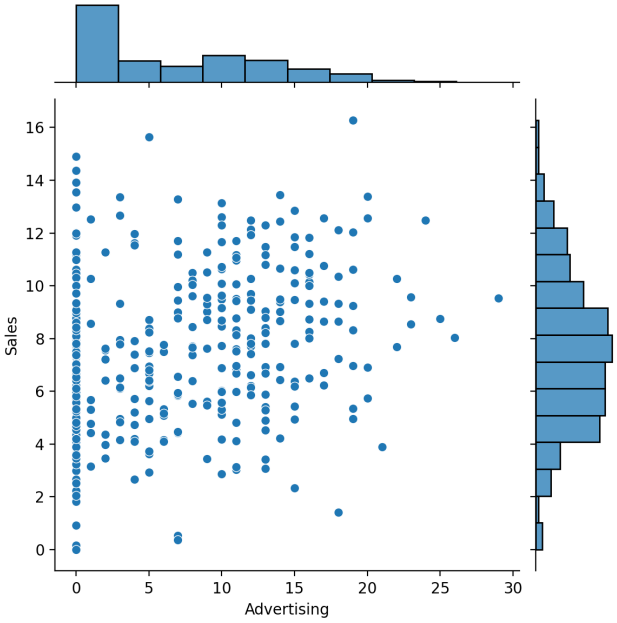
I simply set the first 300 rows as the training set, and the remaining 100 rows as the testing set.

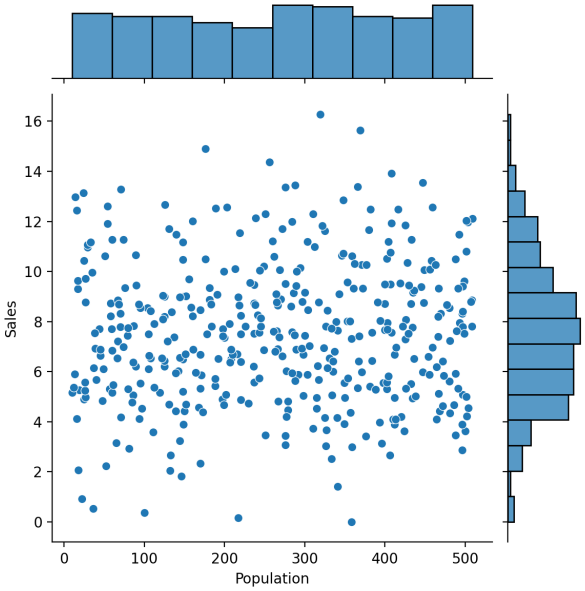
Data Statistics

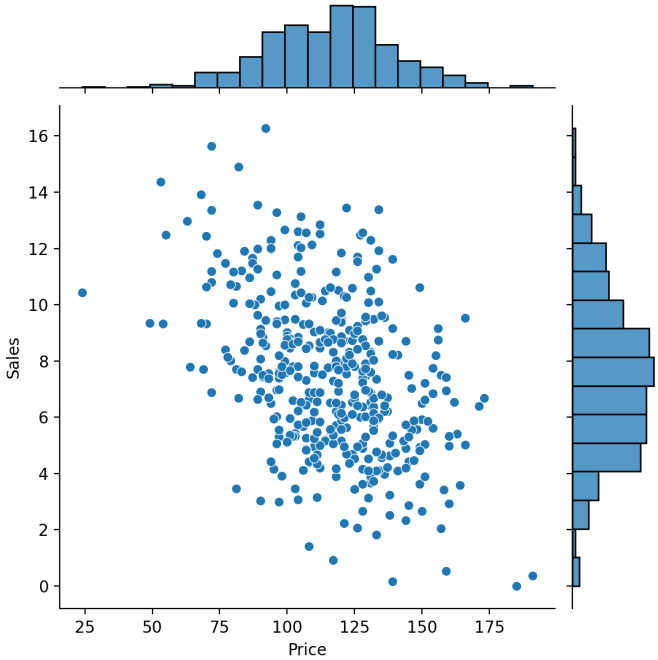


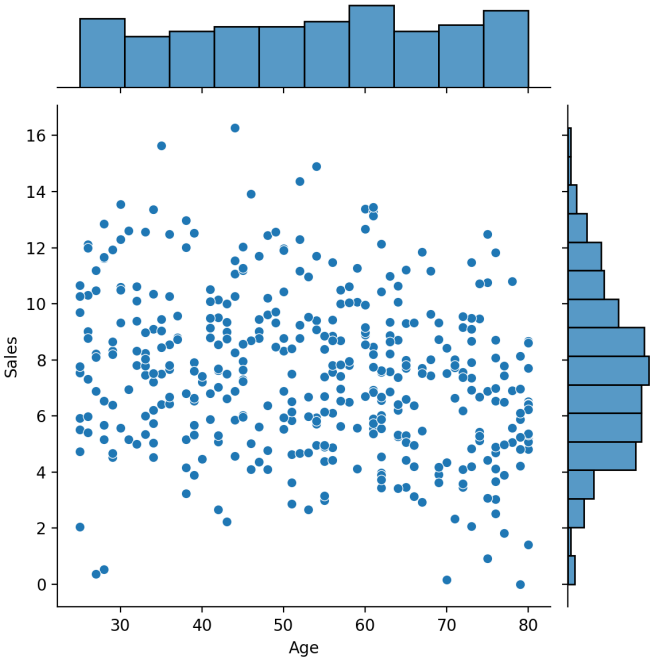

Sales
CompPrice

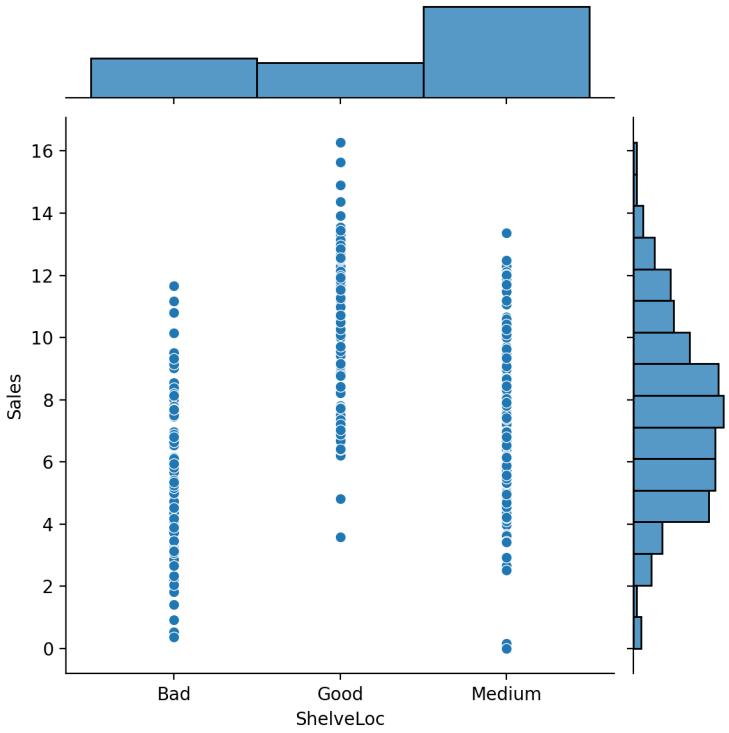

Sales
Income

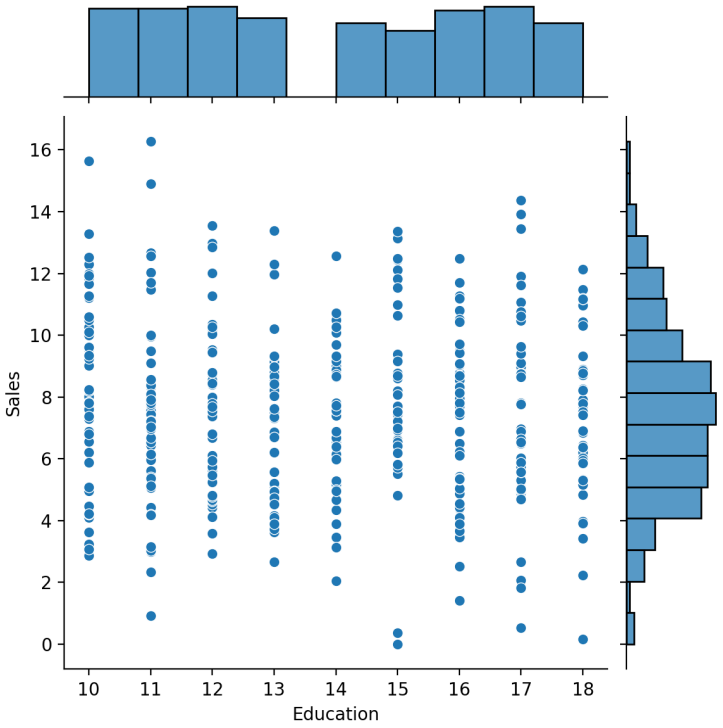

Sales
Advertising

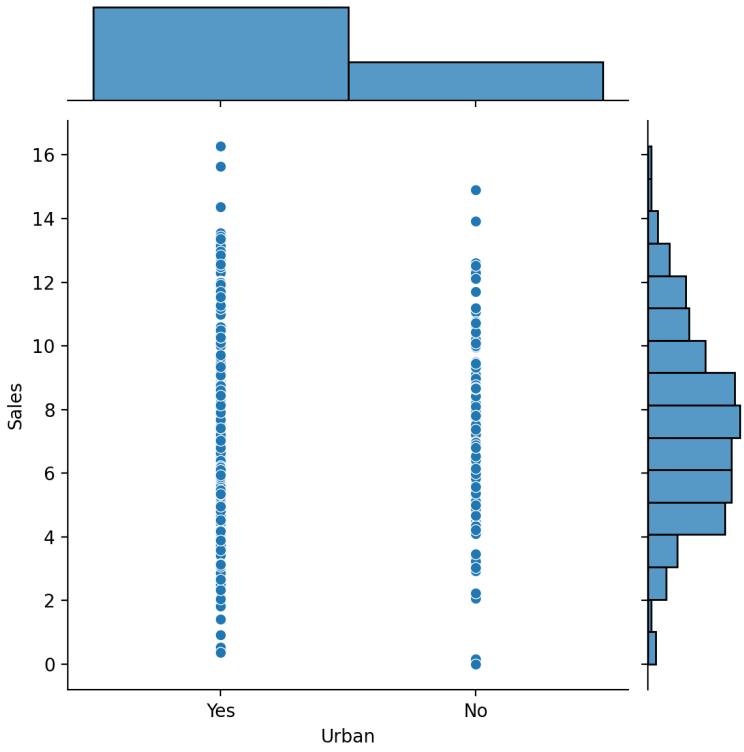

Sales
Population

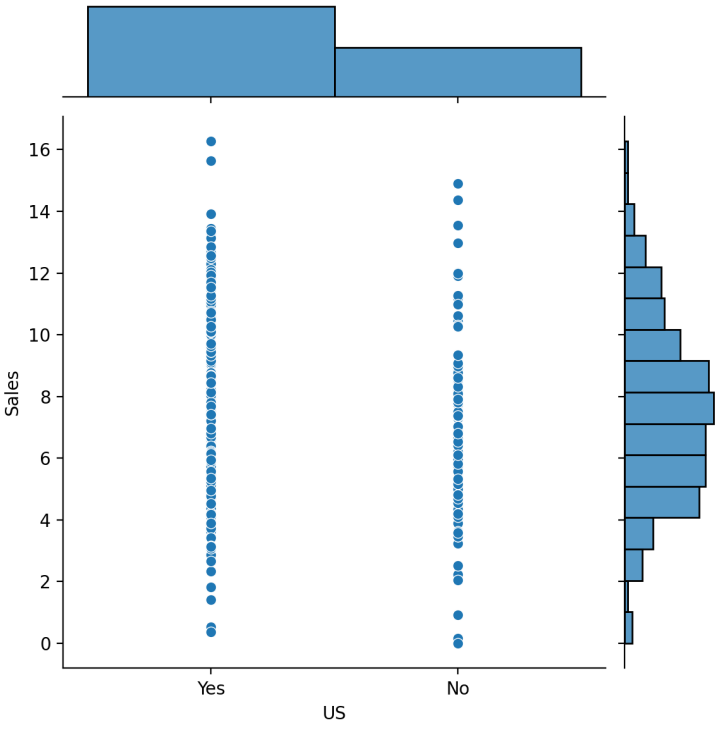

Sales
Price


Sales
Age


Sales
ShelveLoc
Bad
Good
Medium


Sales
Education


Sales
Urban
Yes
No

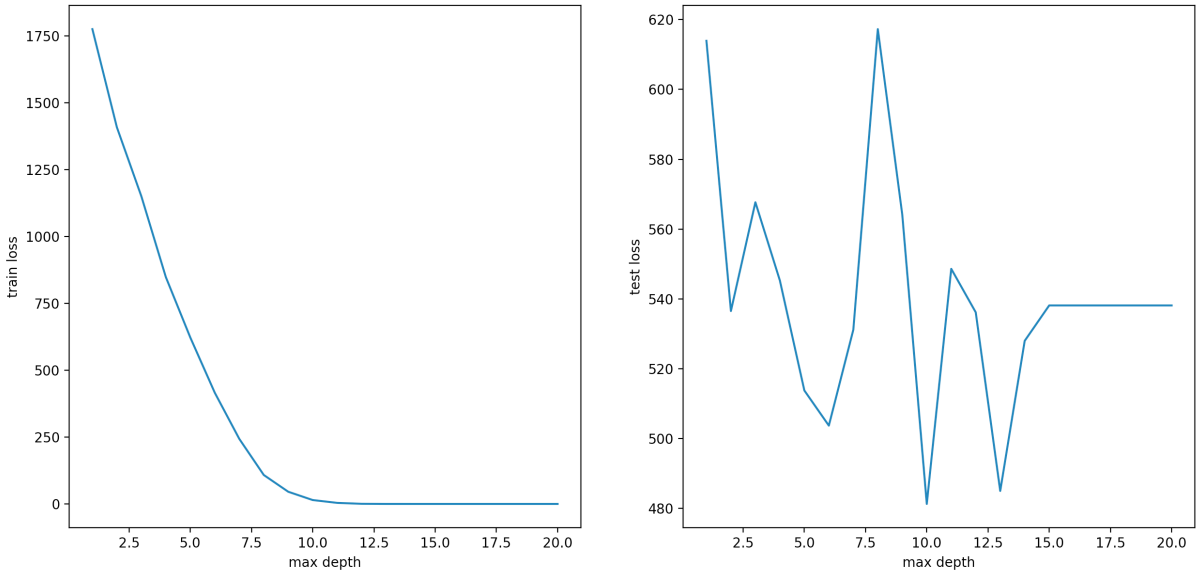

Sales
US
Yes
No

Decision tree

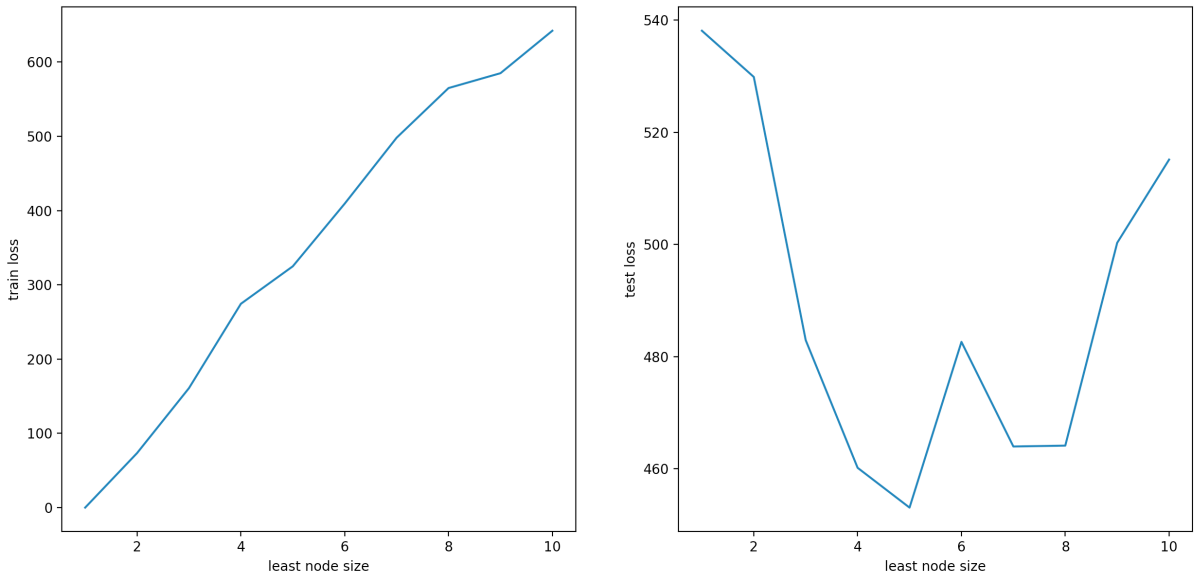
Regression of test data(Sales):

```
test loss of fit with default parameters(decision tree):538.1256999999999
Regression of Sales with default parameters(decision tree):
[12.01  8.98  1.42 10.59  6.67  6.71  5.08  5.05  7.23 10.07  5.4   8.25
  4.96  7.71  8.77  6.95 11.99  7.49 12.49  7.23  8.25  8.78 11.27  9.49
  3.24 10.14  7.32  8.67  2.23 12.98  4.36  7.8   6.5   5.61  8.41  8.73
  5.31  6.56  8.78  8.77  9.14  5.27  9.32  8.32 12.29  9.09  8.54  7.8
 12.13 12.3   7.7   8.73 11.17 10.21  5.08  6.67  8.87 11.91  8.01  6.87
 11.7  10.59  4.19 13.91  9.48  5.93  7.23 11.82 10.    9.7   7.82  6.2
  8.19  5.07  5.64  6.43 12.13  8.54  7.44  4.38  3.47  4.1   8.73  7.71
  8.21  6.87  4.97  4.11  6.9   7.67  4.34  7.96  6.01  9.01  6.01  9.03
  5.64  6.5   2.99  7.8 ]
```

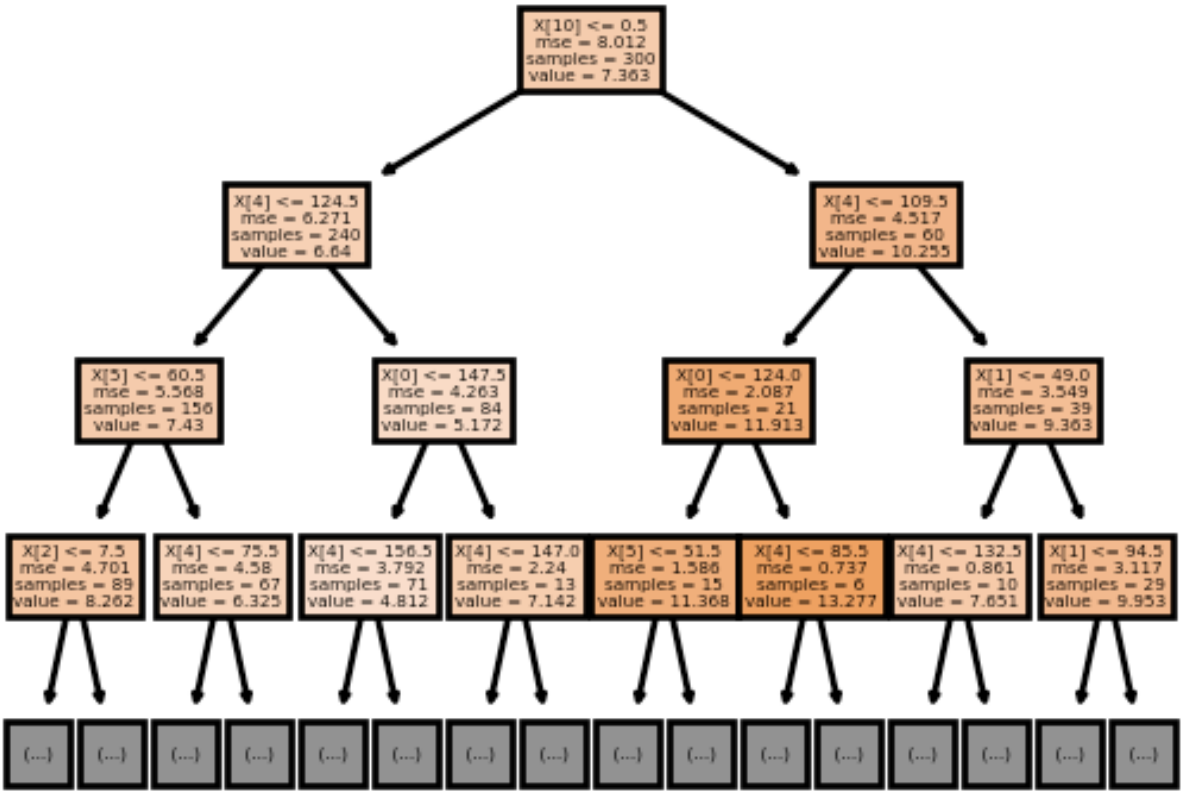
Problem2: losses w.r.t. max depth



Problem2: losses w.r.t. least node size



Tree plot(only show 3 layers because more layer is hard to visualize):

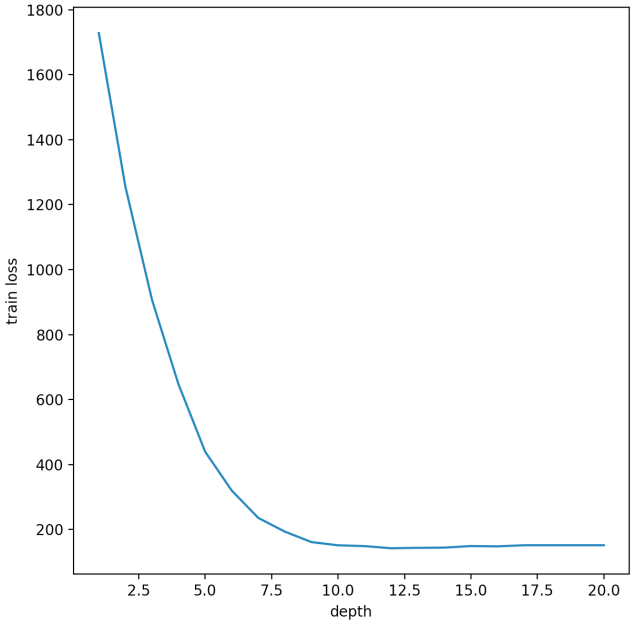


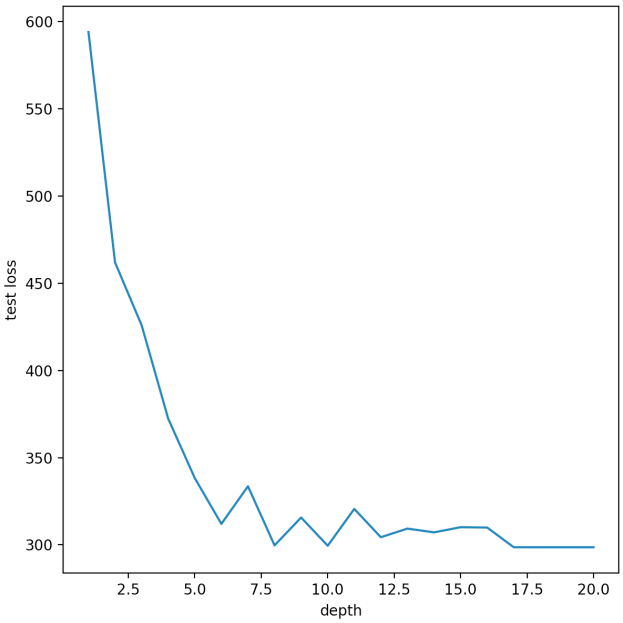
Bagging of trees

Regression of test data(Sales):

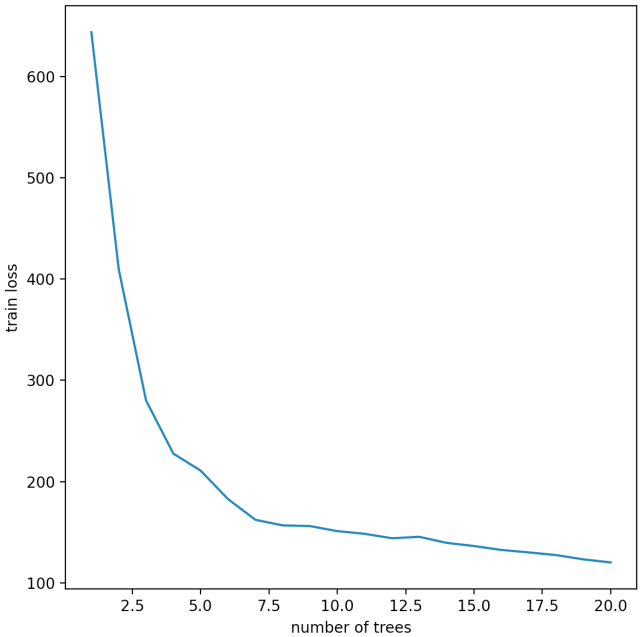
```
test loss of fit with default parameters(bagging):298.61916799999995
Regression of Sales with default parameters(bagging):
[ 6.71   7.649  4.052  9.032  8.798  6.231  6.188  5.907  6.353  9.298
  5.907  7.051  5.05   9.415  9.36   7.345 12.228  8.236  8.455  5.769
  6.933  6.358 10.325  8.694  4.02  10.164  7.06   5.48   3.835 11.909
  5.489  7.324  5.808  6.102  9.733  6.399  5.397  7.29   6.177  9.313
  7.863  7.291  9.006  6.217  9.296  7.568  5.792  9.623 11.028  8.427
  8.359  8.975 10.432  9.631  5.994  8.915  8.097  8.719  6.776  5.741
 10.086  7.677  5.512 10.727  8.345  4.502  5.071 12.367 10.809  9.464
  8.223  6.753  5.896  7.164  8.11   6.036 12.073  5.751  6.739  5.381
  7.614  4.357  6.52   8.545 10.177  5.767  5.465  8.342  9.545  8.54
  6.383  7.14   6.739  6.911  6.268 11.255  7.314  6.056  7.523  9.753]
```

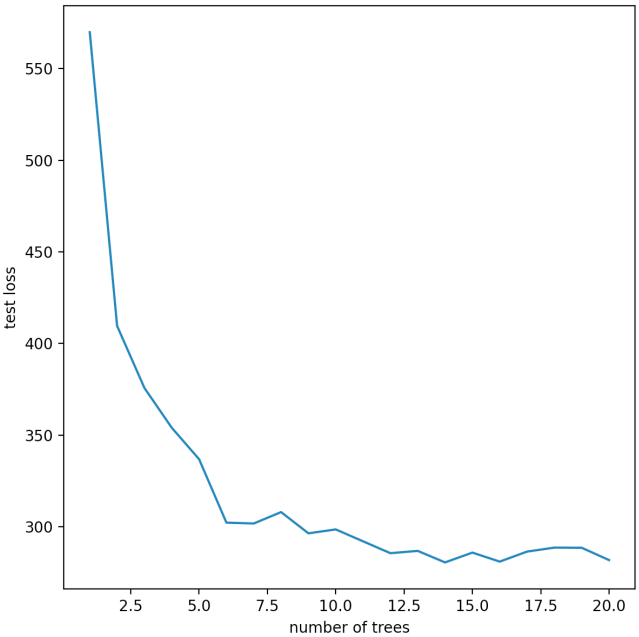
Problem3: losses w.r.t. depth





Problem3: losses w.r.t. number of trees



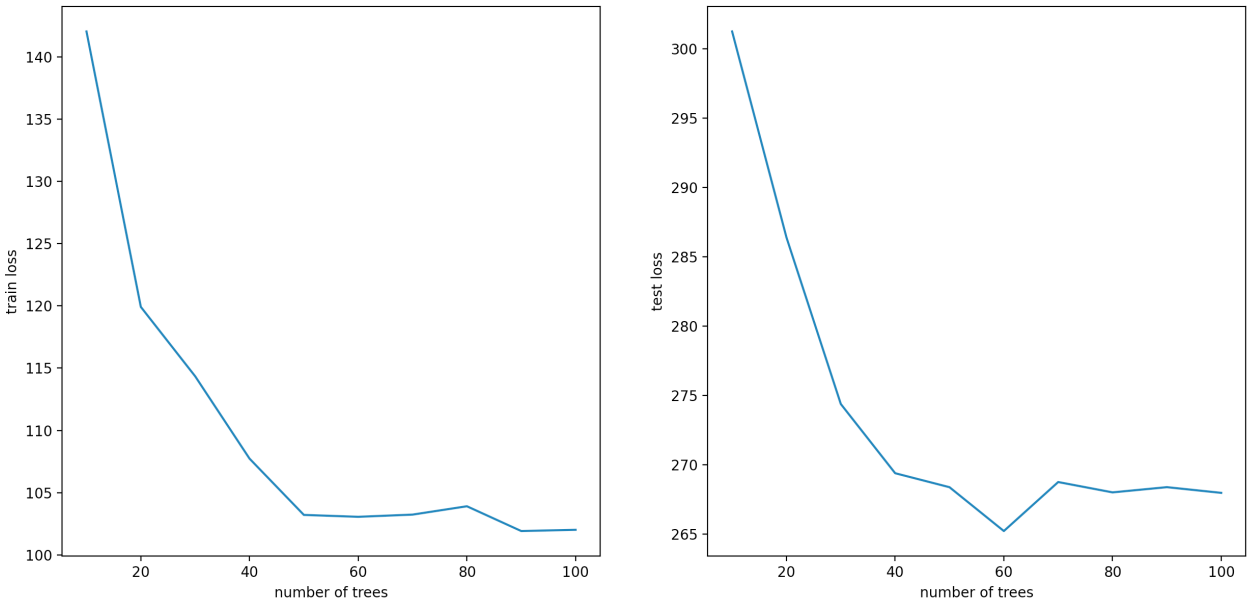


Random forests

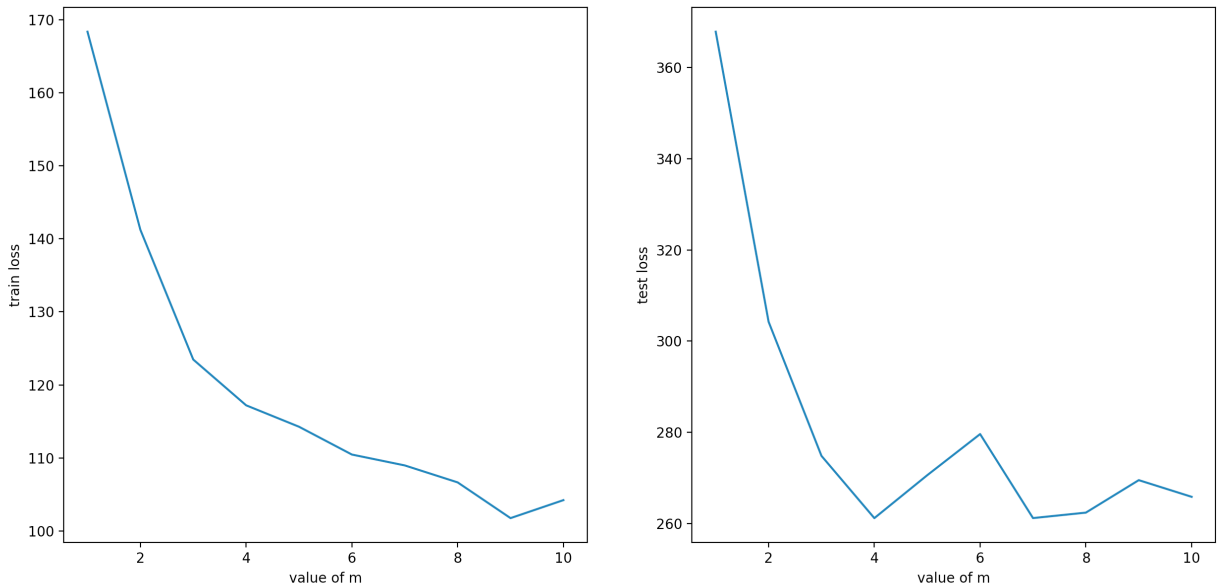
Regression of test data(Sales):

```
test loss of fit with default parameters(random forest):267.96594553
Regression of Sales of test data with default parameters(random forest):
[ 7.5711  8.6974  4.495   9.6889  9.4351  6.3745  6.0455  6.0468  7.1753
  9.339   5.9752  6.3031  4.9216  8.4852  9.1776  7.8333 12.0651  8.9856
  9.2665  6.4058  6.8956  7.1611 10.0672  8.7809  4.0496  9.9974  6.8729
  5.8182  4.4228 11.7324  5.3459  7.2511  6.6785  6.1265  9.3597  7.3153
  4.5405  7.0519  6.6389  9.3896  8.4158  7.1753  8.4976  6.8937  9.837
  8.0733  6.479   9.0777 11.8708  9.4408  8.1333  8.8554 10.0979  9.0443
  5.841   8.6673  8.4082  9.1388  5.6218  5.3983  9.9724  8.4368  5.3372
 10.6765  9.1317  4.6402  5.4568 11.7938 11.3753  9.3265  7.8388  6.8353
  6.1119  6.6744  7.8514  6.2269 12.5795  6.1191  6.4092  5.5898  7.6793
  4.2608  7.1788  9.1725 10.1563  6.2366  5.3066  7.9376  9.1164  8.2575
  6.5236  6.899   6.0005  6.9534  5.9391 10.7589  7.4791  5.8537  6.0866
  9.6148]
```

Problem4: losses w.r.t. number of trees

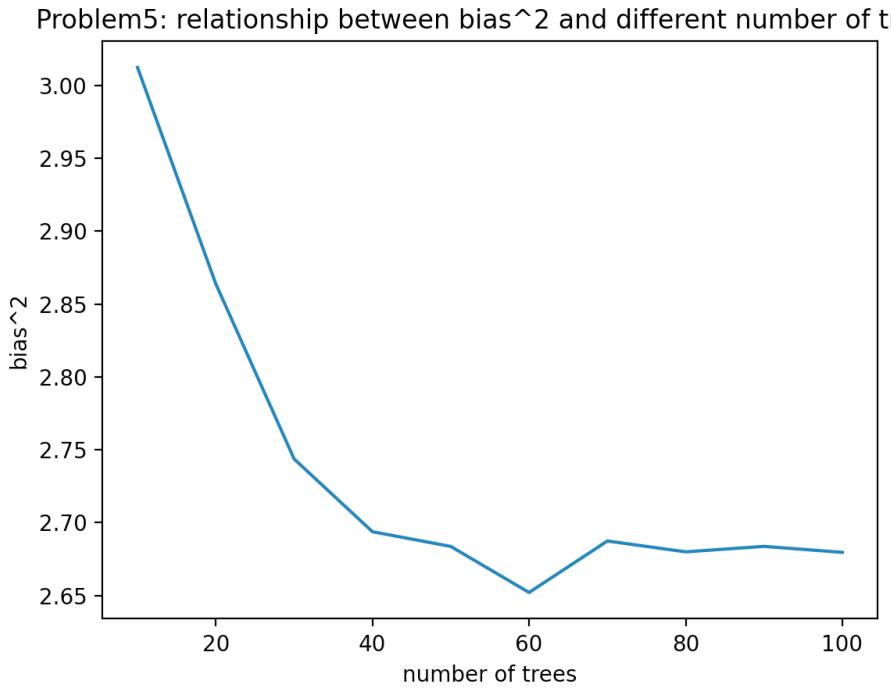


Problem4: losses w.r.t. values of m

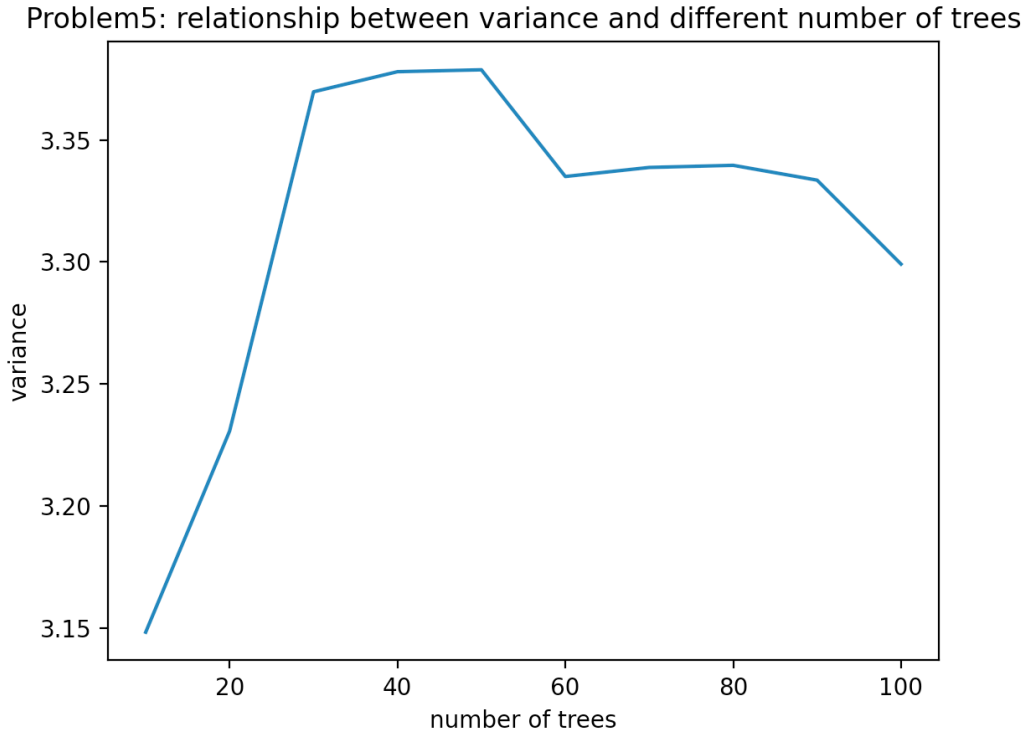


Bias2 and Variance

Since we cannot know the target function of the real data, we assume that the ε is 0, and use y as the target function t(x) to calculate the bias.



Relation: Bias^2 should decrease as the number of trees increase, and the graph is correspond to this conclusion.



Relation: The variance should increase as the number of tree increase, however, it decreases in this graph when the num is larger than 50. This little difference may cause by the dataset.