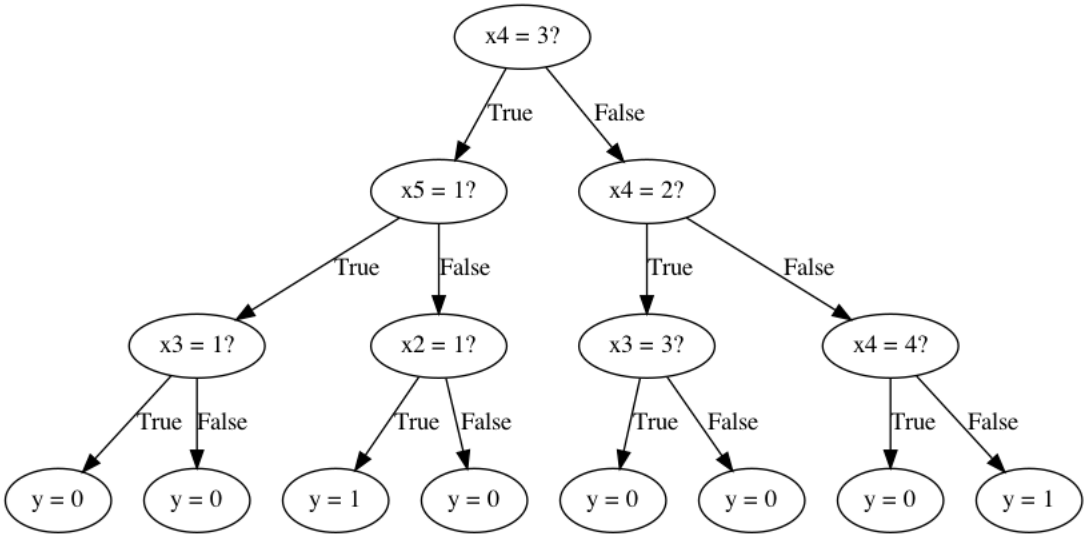
CS6375 Assignment 2 Report

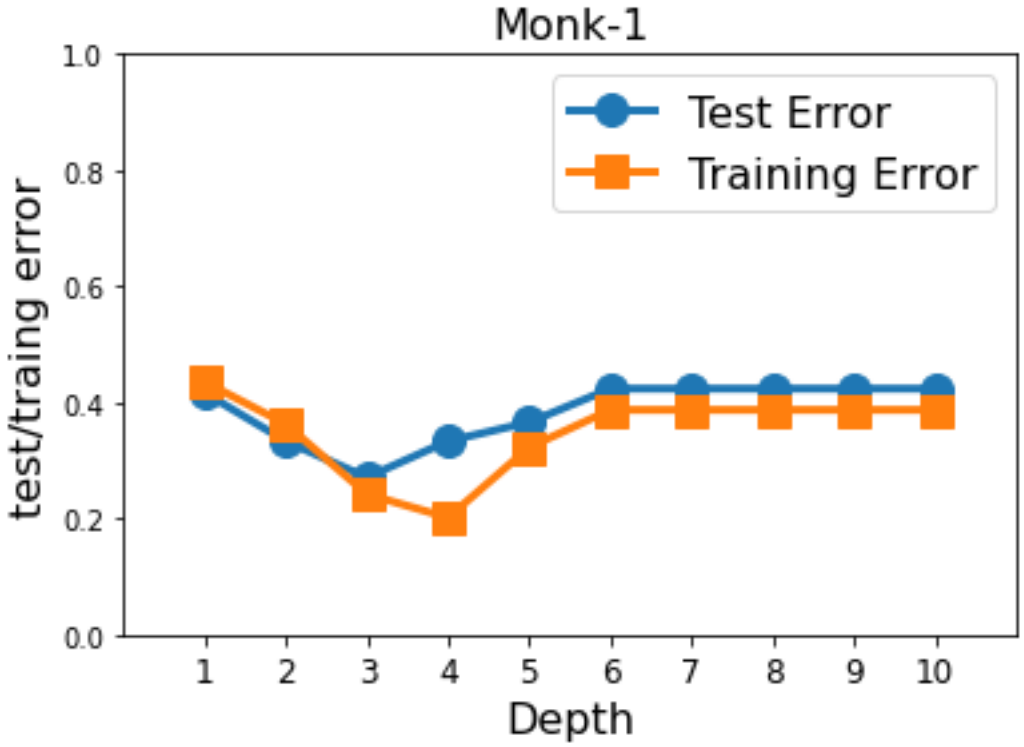
Name: Haoda LE
Netid: HXL180046

a. The python code is in another file decision_tree.py
the running result with default setting (dataset: monks-1.train, monks-1.test; max_depth = 3) is as follows:

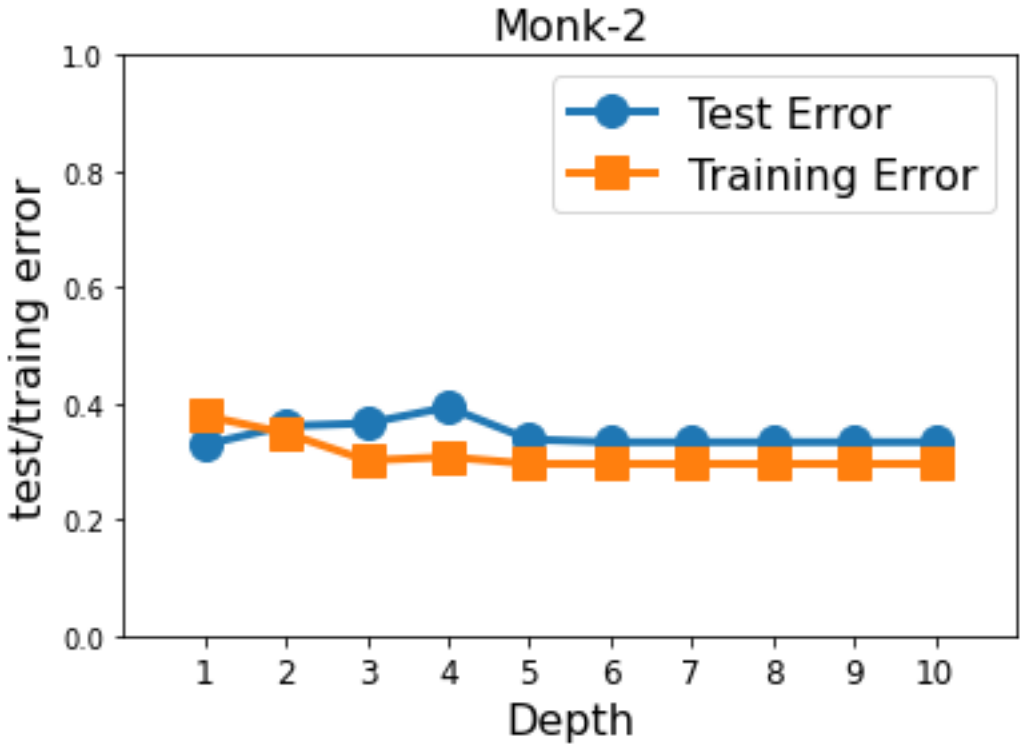


TREE
+-- [SPLIT: x4 = 3 True]
| +-- [SPLIT: x5 = 1 True]
| | +-- [SPLIT: x3 = 1 True]
| | | +-- [LABEL = 0]
| | +-- [SPLIT: x3 = 1 False]
| | | +-- [LABEL = 0]
| +-- [SPLIT: x5 = 1 False]
| | +-- [SPLIT: x2 = 1 True]
| | | +-- [LABEL = 1]
| | +-- [SPLIT: x2 = 1 False]
| | | +-- [LABEL = 0]
+-- [SPLIT: x4 = 3 False]
| +-- [SPLIT: x4 = 2 True]
| | +-- [SPLIT: x3 = 3 True]
| | | +-- [LABEL = 0]
| | +-- [SPLIT: x3 = 3 False]
| | | +-- [LABEL = 0]
| +-- [SPLIT: x4 = 2 False]
| | +-- [SPLIT: x4 = 4 True]
| | | +-- [LABEL = 0]
| | +-- [SPLIT: x4 = 4 False]
| | | +-- [LABEL = 1]
Test Error = 27.08%.
Train Error = 24.19%.

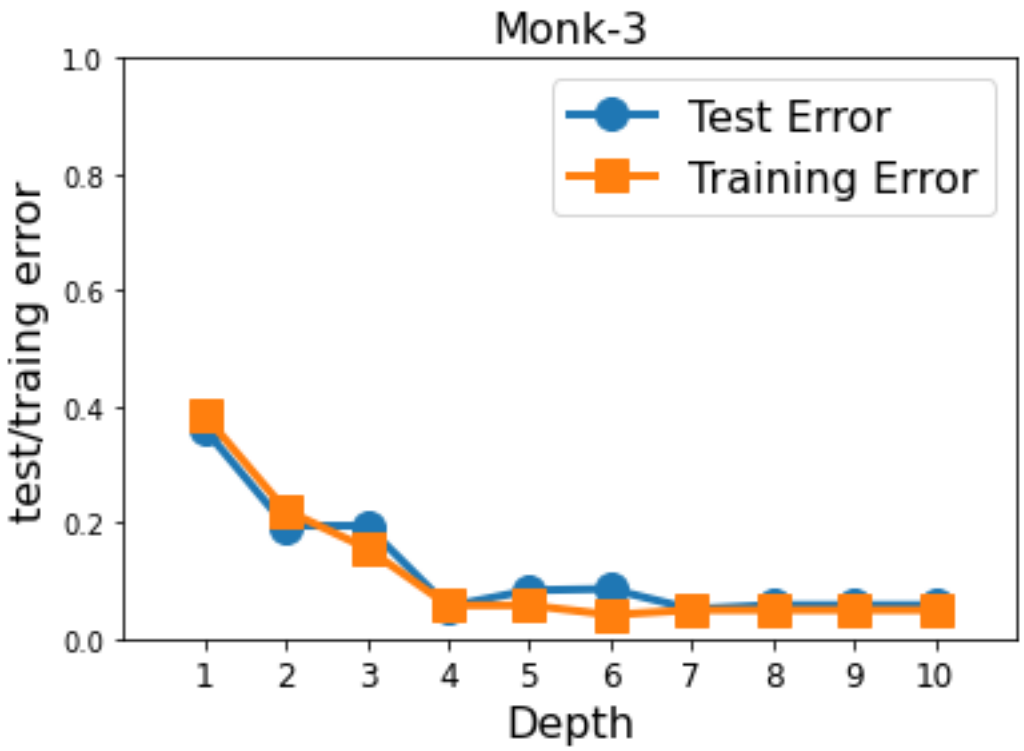
b.



for Monk-1 problem, training/test error shows above. Both training and test error rate declines to depth 3, then test error increase, which may have over fitting.

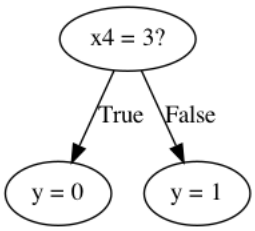


for Monk-2 problem, training/test error shows above. After depth 5, both training and test error remains low.



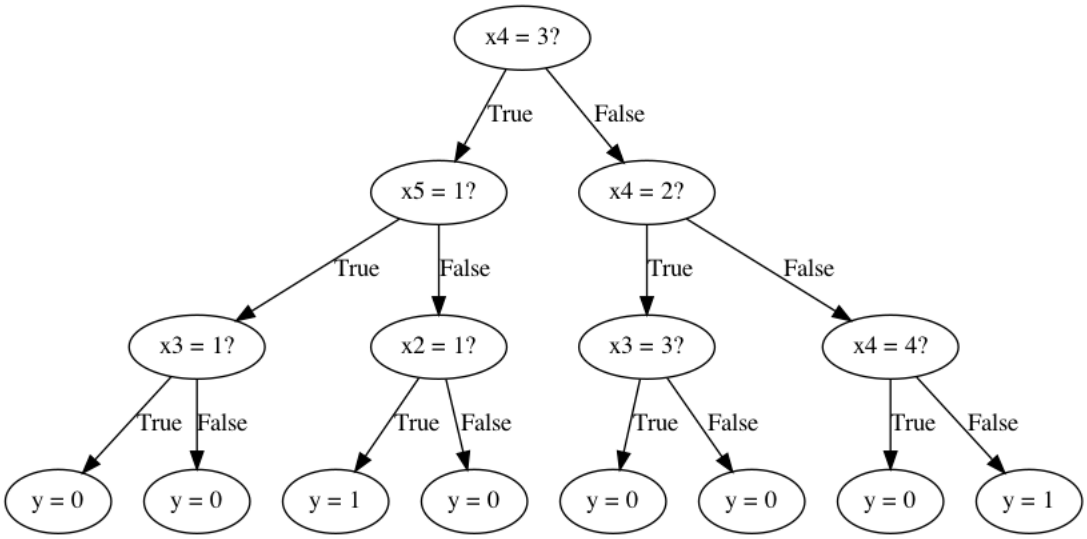
for Monk-3 problem, training/test error shows above. After depth 7, both training and test error remains low.

c.
for monks-1 problem, id3 algorithm, max_depth 1:



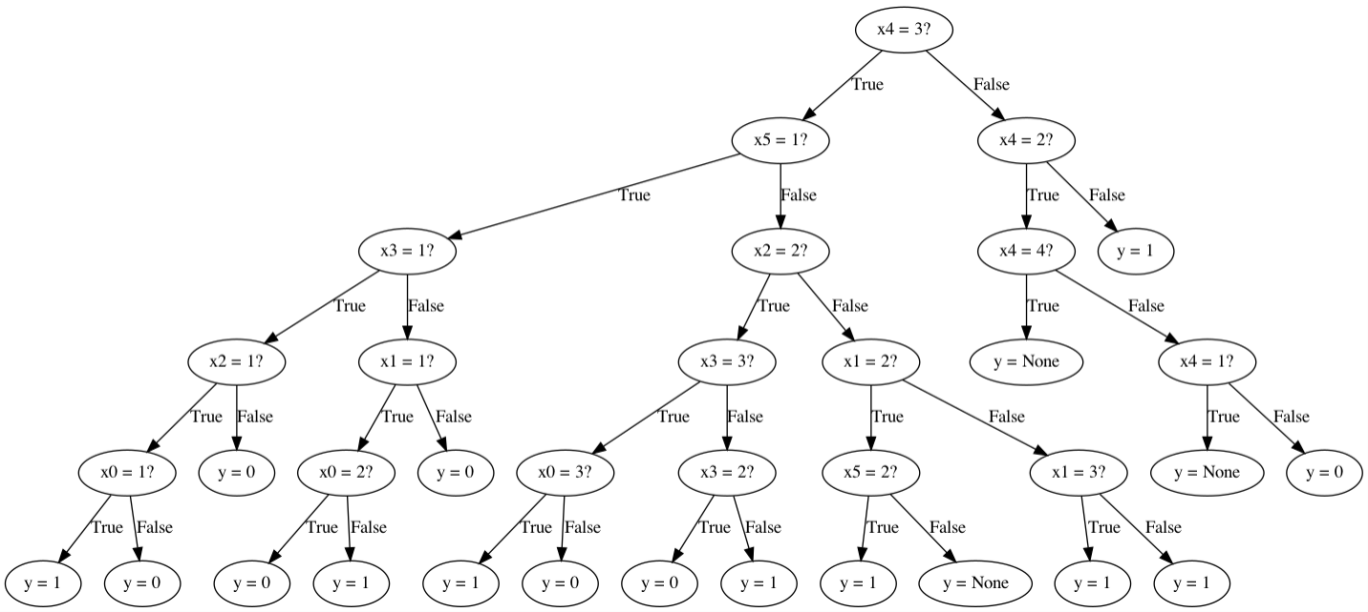
depth:1 confusion matrix
[[72 144]
 [36 180]]

for monks-1 problem, id3 algorithm, max_depth 3:



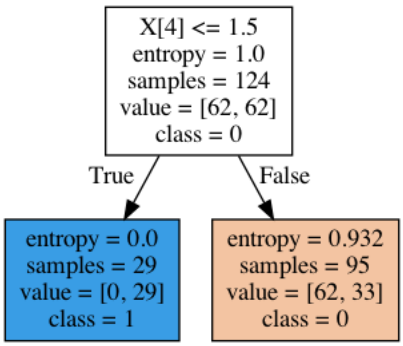
depth:3 confusion matrix
[[198 18]
[99 117]]

for monks-1 problem, id3 algorithm, max_depth 5:



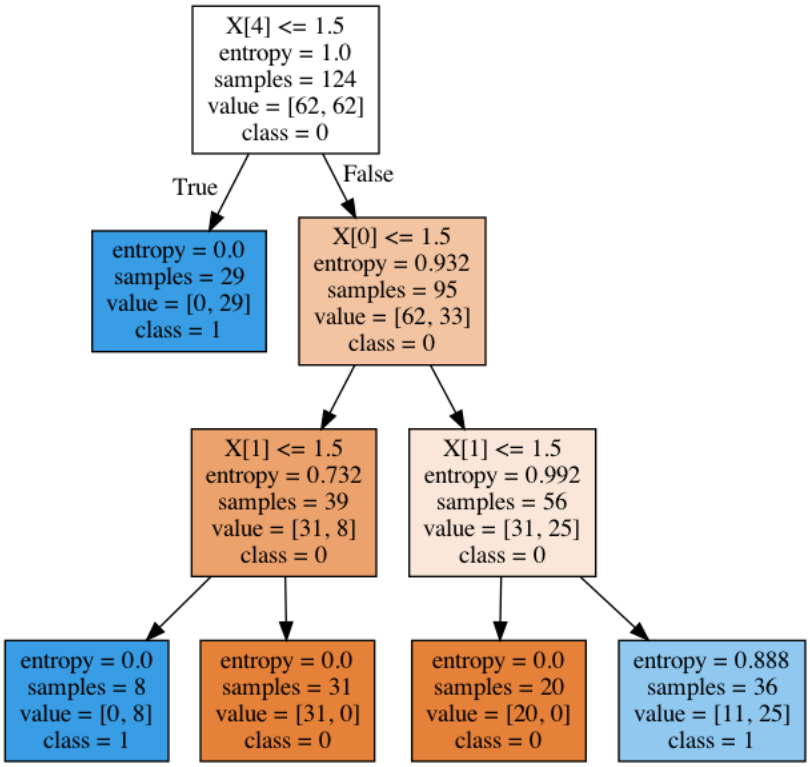
depth:5 confusion matrix
[[112 104]
[54 162]]

d.
for monks-1 problem, scikit-learn's algorithm, max_depth 1:



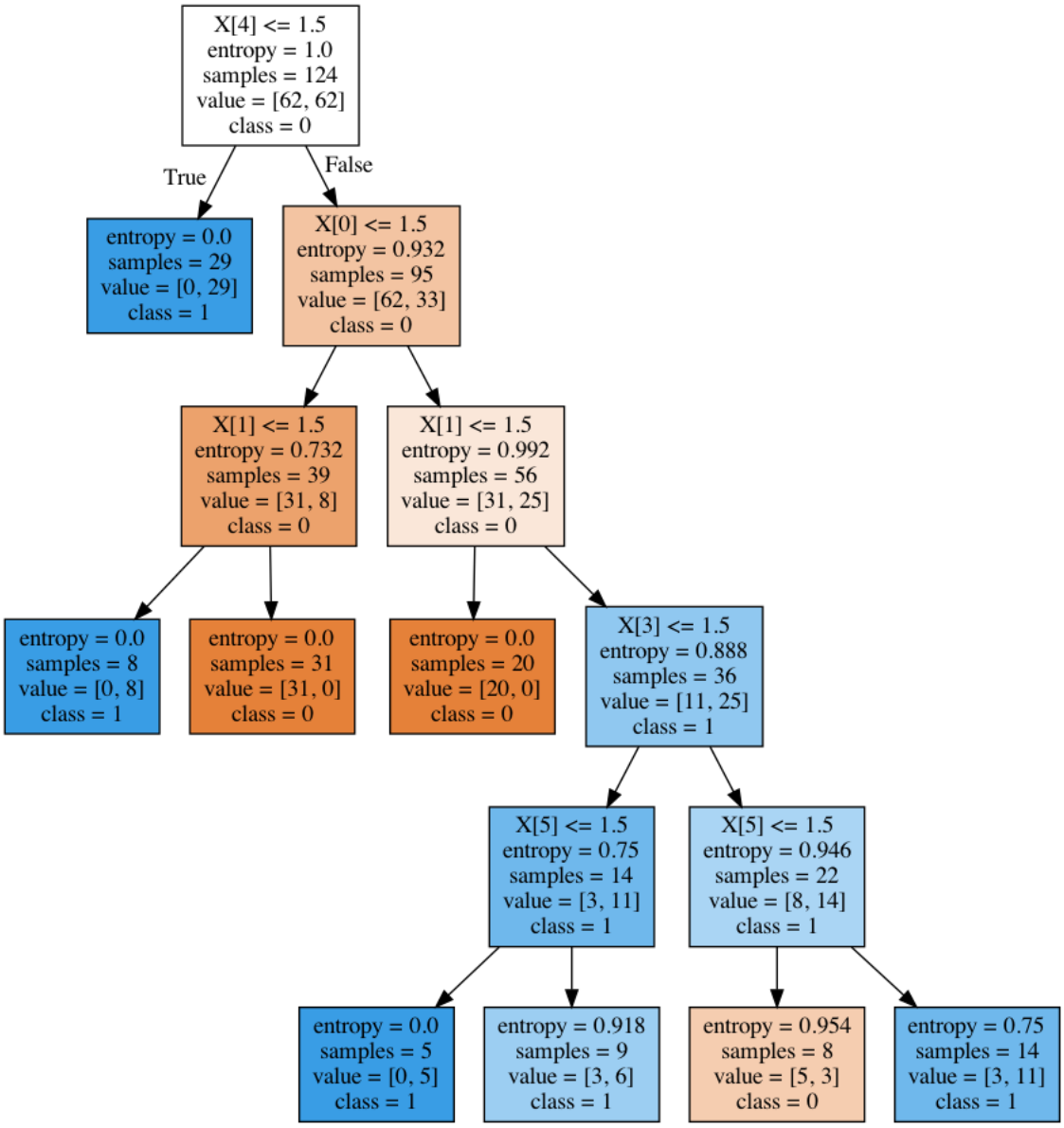
depth:1 confusion matrix
[[216 0]
[108 108]]

for monks-1 problem, scikit-learn's algorithm, max_depth 3:



depth:3 confusion matrix
[[144 72]
[0 216]]

for monks-1 problem, scikit-learn's algorithm, max_depth 5:



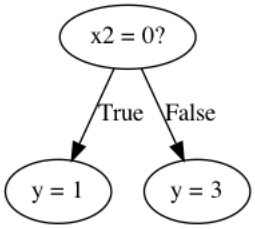
depth:5 confusion matrix
[[168 48]
[24 192]]

e.
Use other data sets in the UCI repository, "Iris Data Set". It has 4 features, totally 150 instances. so 120 instances used for training, and 30 instances for test.

Totally 3 target classes.
class:
#-- 1 : Iris Setosa
#-- 2 : Iris Versicolour
#-- 3 : Iris Virginica

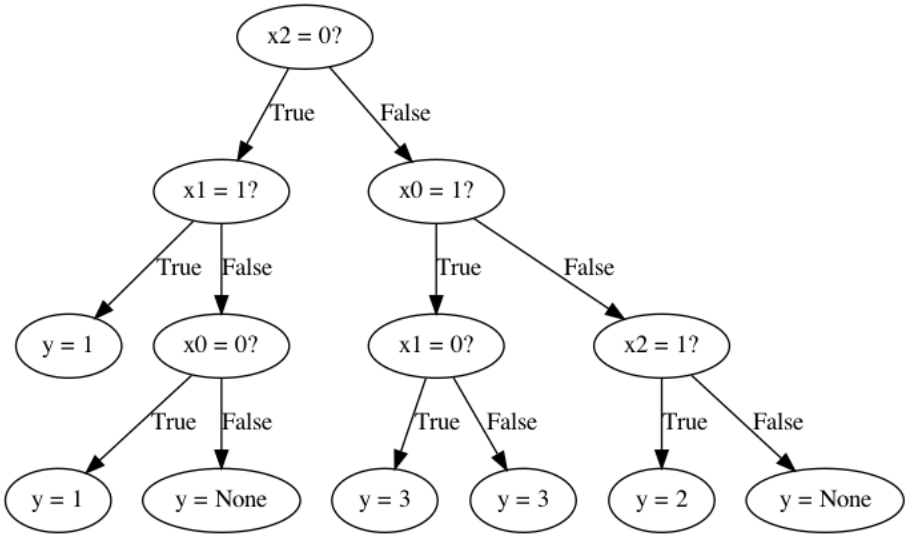
since it is continuous features, use a simple discretization strategy to pre-process them into binary features. x <= mean : 0; x > mean : 1

id3 algorithm, max_depth 1:



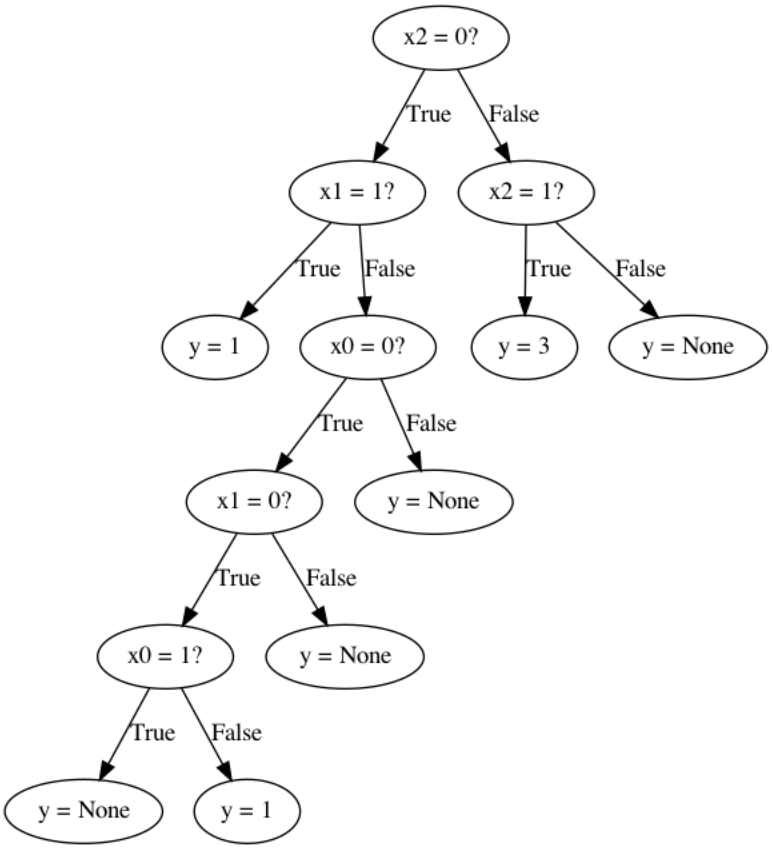
id3, depth:1 confusion matrix
[[10 0 0]
[2 0 8]
[0 0 10]]

id3 algorithm, max_depth 3:



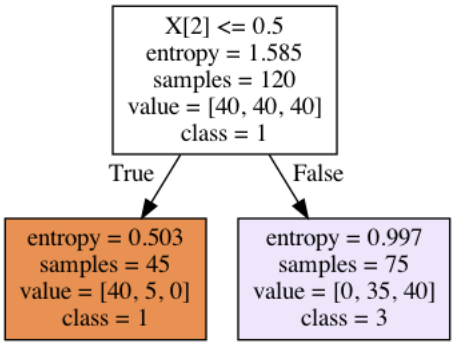
id3, depth:3 confusion matrix
[[10 0 0]
 [2 6 2]
 [0 1 9]]

id3 algorithm, max_depth 5:



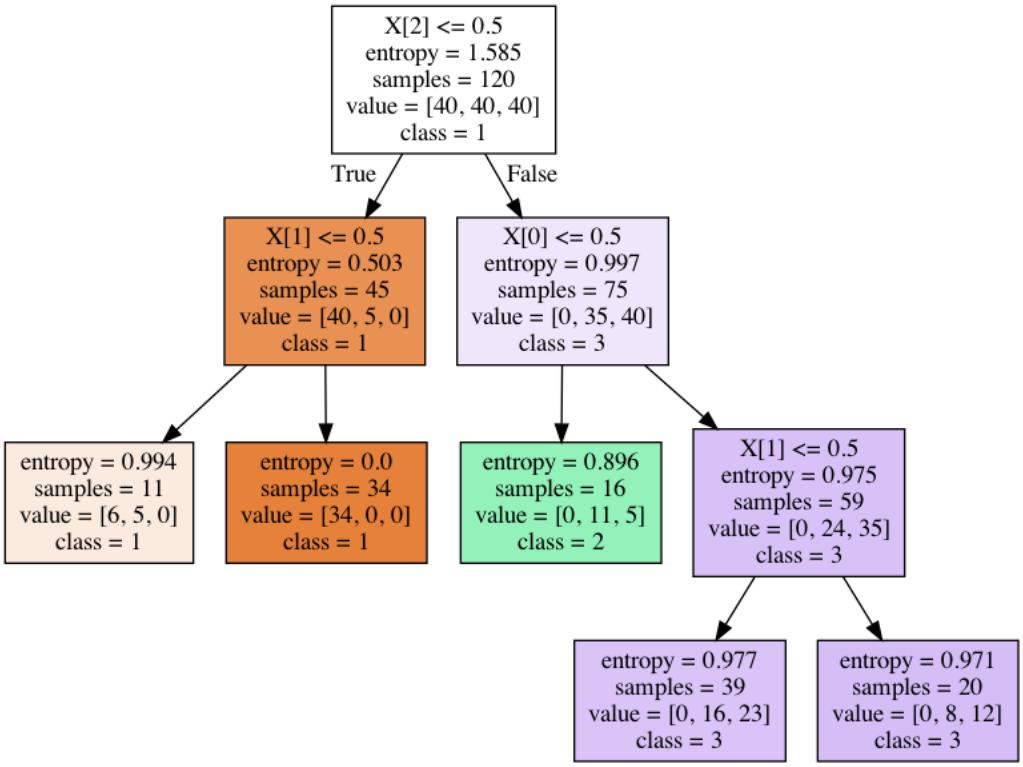
id3, depth:5 confusion matrix
[[10 0 0]
 [2 0 8]
 [0 0 10]]

scikit-learn’s algorithm, max_depth 1:



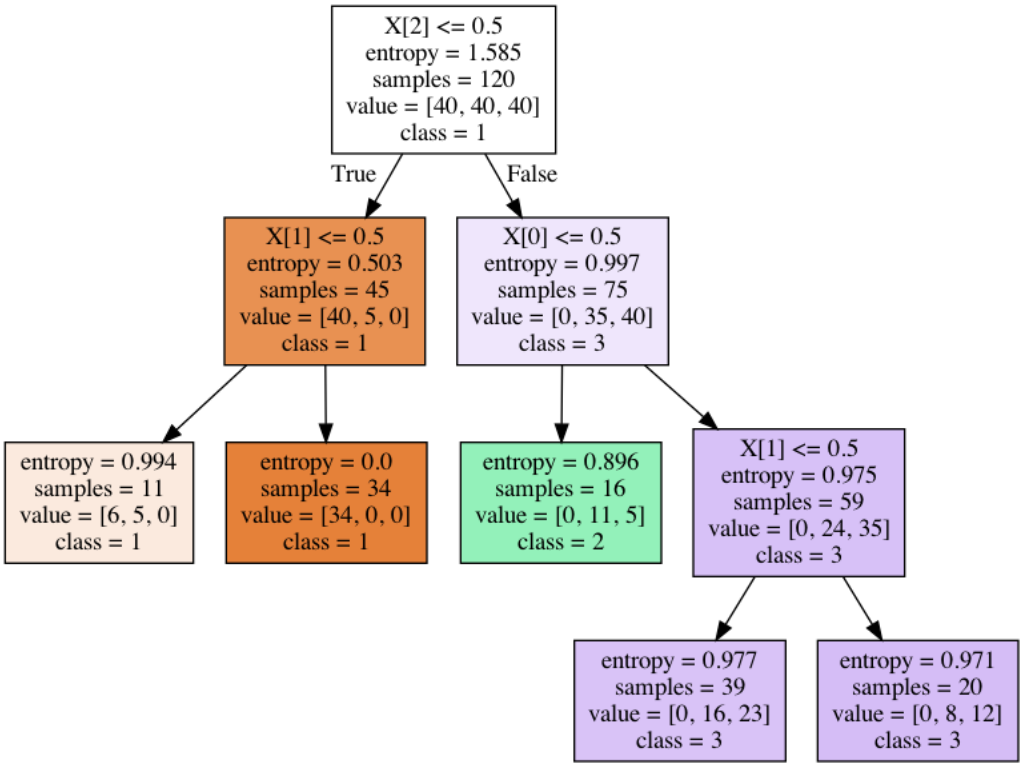
sk, depth:1 confusion matrix
[[10 0 0]
[2 0 8]
[0 0 10]]

scikit-learn’s algorithm, max_depth 3:



sk, depth:3 confusion matrix
[[10 0 0]
[2 6 2]
[0 1 9]]

scikit-learn's algorithm, max_depth 5:



sk, depth:5 confusion matrix
[[10 0 0]
[2 6 2]
[0 1 9]]

Confusion matrix comparison:

Max_depth	id3 confusion matrix	id3 accuracy	Scikit-learn confusion matrix	Sk accuracy
1	[[10 0 0] [2 0 8] [0 0 10]]	0.667	[[10 0 0] [2 0 8] [0 0 10]]	0.667
3	[[10 0 0] [2 6 2] [0 1 9]]	0.833	[[10 0 0] [2 6 2] [0 1 9]]	0.883
5	[[10 0 0] [2 0 8] [0 0 10]]	0.667	[[10 0 0] [2 6 2] [0 1 9]]	0.883

Based on the limited number of training and test data instance, as we can see from the result, on and before max_depth 3, my implemented id3 algorithm has the exact same confusion matrix and accuracy

as scikit-learn algorithm. When max_depth set to 5, my id3 algorithm will make the decision tree deeper, which reduces the accuracy.
While for scikit-learn algorithm, as we can see, the max_depth = 3 has the exact same decision tree as max_depth = 5, because decision tree with depth 3 already has high accuracy, make the tree deeper will reduce the accuracy.
So generally speaking, scikit-learn algorihtm is a little bit better than my implemented id3 algorithm.