EC4213/ET5402/CT5303: Machine learning and deep learning **Fall 2019**

Report for CA1

Instructor: Jonghyun Choi *Jwa Younkyung (20165174)*

REPORT1. why is this computation equivalent to computing classification accuracy ?
It is a good classification when the prediction value using X and the value of Y are identical. Since Y is -1 or 1, the Y value greater than 0 and the prediction value greater than 0 are compared and the average of the compared values is accuracy.

REPORT2. Please dicsuss the reason that the training accuracy tend to decrease. Also, discuss the reason that the test accuracy not monotonically increasing.

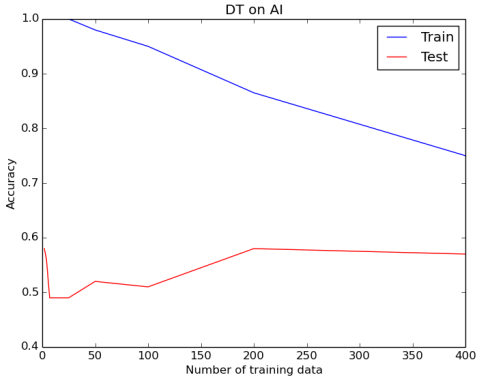


FIGURE 1 – DT on AI.

Large numbers of data drive more bias, which tends to decrease due to bias training accuracy. If the number of data used for training is very small and the test data has similar data to the training data, the accuracy can be larger than others.

REPORT3. If the implementation is correct, you will observe the training accuracy monotonically increasing and test accuracy tumbling. Please discuss the behavior of the training accuracy and test accuracy.

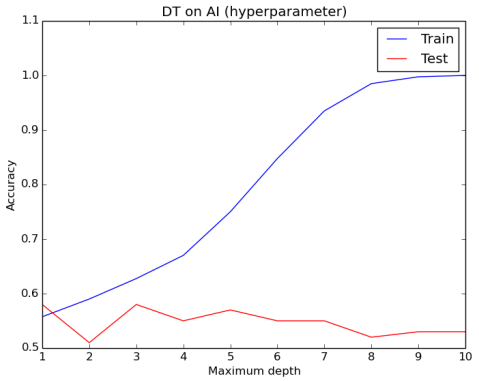


FIGURE 2 – DT on AI(hyperparameter).

Increasing the maximum depth will overfit the model. Therefore, it is very suitable for training sets and improves the accuracy of training sets. As the variance of the model increases, the shape of the decision tree continues to change.

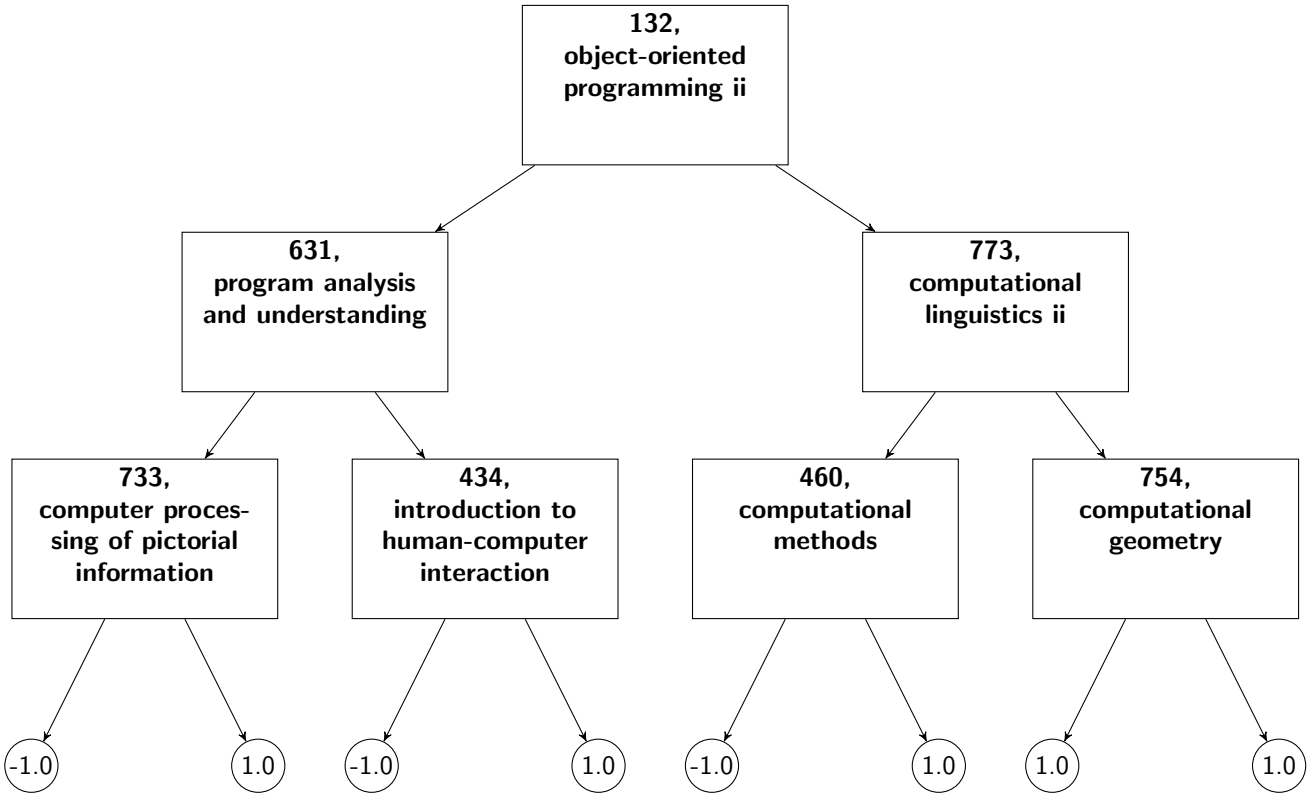
As a result, the test set and model may or may not fit well. So accuracy is tumbling.

REPORT4. Train a decision tree on the CG dataset with a maximum depth of '3'. In the write-up, first, draw out the decision tree for this classifier (but put in the actual course names/ids as the features). Then, discuss about this tree : do these courses are indicative of whether someone might take CG ?

Result of train data is :

Branch 6
- Branch 34
  - Branch 48
    - Leaf -1.0
    - Leaf 1.0
  - Branch 27
    - Leaf -1.0
    - Leaf 1.0
- Branch 54
  - Branch 32
    - Leaf -1.0
    - Leaf 1.0
  - Branch 53
    - Leaf 1.0
    - Leaf 1.0

It may look like this :



This decision tree means that whether someone might take CG depends on whether or not "object-oriented programming ii", "program analysis and understanding", "computer processing of pictorial information", "introduction to human-computer interaction", "computational linguistics ii", "computational methods" and "computational geometry" are taken. Only 7 features are involved in TookCG.

REPORT5. Compare the test accuracy before and after the pruning by plotting a comparative graph. And discuss the effect of pruning in terms of training accuracy, test accuracy and generalization performance. Also, discuss the good hyperparameter (eg., 0.05) for each dataset.

I fixed some code in runClassifier and ran the code as below for make curve

```
>>>curve = runClassifier.learningPrunedCurveSet(dt.DT('maxDepth' : 5), datasets.CFTookAI)
>>>runClassifier.plotCurve('DT on AI with pruning', curve)
>>>curve = runClassifier.hyperparamPCurveSet(dt.DT('maxDepth' : 5), 'maxPvalue', [0.01, 0.02, 0.03, 0.04, 0.05,
0.06, 0.07, 0.08, 0.09, 0.1], datasets.CFTookAI)
>>>runClassifier.plotCurve('DT on AI with pruning (hyperparameter)', curve)
>>>curve = runClassifier.comparativeTestingCurveSet(dt.DT('maxDepth' : 5), datasets.CFTookAI)
>>>runClassifier.plotCurve('Comparative Test', curve)
>>>curve = runClassifier.comparativeTrainingCurveSet(dt.DT('maxDepth' : 5), datasets.CFTookAI)
>>>runClassifier.plotCurve('Comparative Train', curve)
```

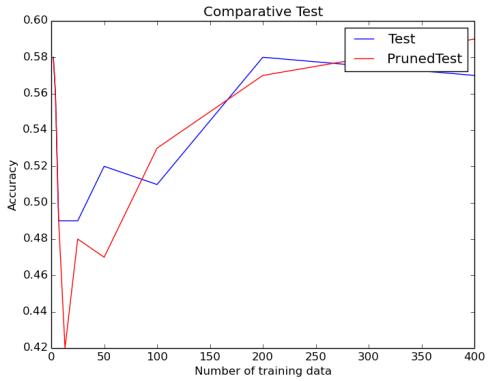


FIGURE 3 – Comparative Test.

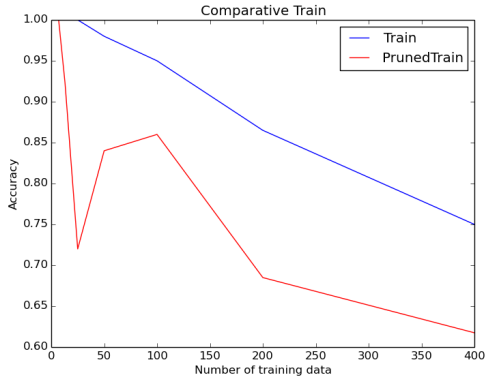


FIGURE 4 – Comparative Train.

Pruning decrease training accuracy and increases testing accuracy little by little. Therefore, pruning can make the model more general.

In Fig.5, accuracy of test set is highest at about 0.08 maxPvalue. So, set maxPvalue = 0.08 is good for pruning.

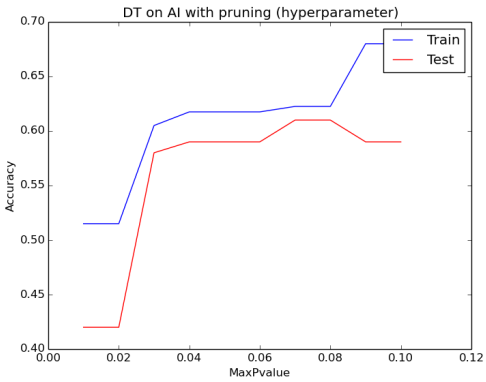


FIGURE 5 – DT on AI with pruning(hyperparameter).