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

Report for CA1

Instructor: Jonghyun Choi *Shin Ji Su (20175095)*

REPORT1. Computation used in here :

$$mean((datasets.TennisData.Y > 0) == (h.predictAll(datasets.TennisData.X) > 0))$$

datasets.TennisData.Y is about real data and *h.predictAll(datasets.TennisData.X)* is about predicted data by X given classifier h. So this computation is about how much data is well predicted and it is divided by total data which is the same with computing classification accuracy .

REPORT2. Training accuracy is related to the training data examples.It can go down after increasing number of examples on training because there can be exist more ambiguity which is made by different labels with same feature values. However, test accuracy is predicting from the features and it could increase when you increase the number of examples on training because features that are considered important on training might be vary as the number of examples on training changes. Thus, according to the various examples, testing accuracy can go up or down.

REPORT3. Training accuracy will monotonically increase because it is valued by the given answer examples. However, overfitting problem might happen when you keep trying to increase the training accuracy. This overfitting will cause the test accuracy to be not only not being monotonically increasing but also tumble. Also increasing questions on DT might increase or decrease the testing accuracy so it will tumble.

REPORT4.

```
-introduction to low-level programming concepts [212]
   - program analysis and understanding [631]
     - computer processing of pictorial information [733]
       - Leaf -1.0
       - Leaf 1.0
     - introduction to human-computer interaction [434]
       - Leaf -1.0
       - Leaf 1.0
   - computational linguistics ii [773]
     - computational methods [460]
       - Leaf -1.0
       - Leaf 1.0
     - computational geometry [754]
       - Leaf 1.0
       - Leaf 1.0
```

The course $introductiontolowlevelprogrammingconcepts(212)$ is not considered as a good feature intuitively and I think it should not be at the top of this tree because it is an introductory course that most of the students take.

Also when you see the decision tree, I think it should set the question about the field like AI or graphic which will be helpful to classify the courses because it will be somehow related to the kinds of courses taken.

And $computationalgeometry(754)$ and $computationallinguisticsii(773)$ are considered as a advanced courses. So there are high probability that studetns will take those course after others like AI or graphic. This means that when decision tree is used to expect whether studetns would take AI or graphic, it is highly likely that there will be no considerations about above courses.

REPORT5. After pruning, the test accuracy will increase. Also by the effect of pruning, when we assume that there are lots of examples for training, the training accuracy will decrease, test accuracy will increase and it will be more generalized.