Assignment 2 for Applied Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 Part A

1.1 Question 1: Show three decision trees on different datasets

In this question, I trained decision trees seperately on three different datasets. Specifically:
(1) The Figure 1 shows the decision tree trained with "titanic.arff" datasets.

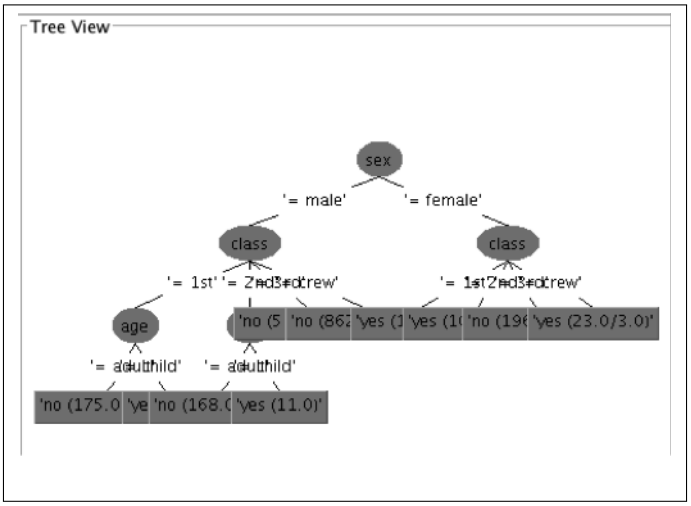

Figure 1: titanic.arff dat trained tree.

(2) The Figure 2 shows the decision tree trained with "titanic-age-numeric.arff" datasets.

(3) The Figure 3 shows the decision tree trained with "titanic_noise.arff" datasets.

1.2 Question 2: Compare results of Figure 1 and Figure 2

The results and the learned decision trees ars quiet different on these two different datasets. According to the figure drew from weka, we can see that in Figure 1, there are more classification rules than in Figure 2 for the "male" subclass. Besides, the Figure 1 shows that there existed a classification hierachy more than Figure 2, the "age".
I think it is almost impossible for the two different datasets to generate the same learning tree, because compared with numeric data, the nominal data lose some "information", like "order" and "distance". What is more, the nominal data would somehow create more classification rules than numeric data.
The J48 settles different different models because those two datasets contains different amout of "information".

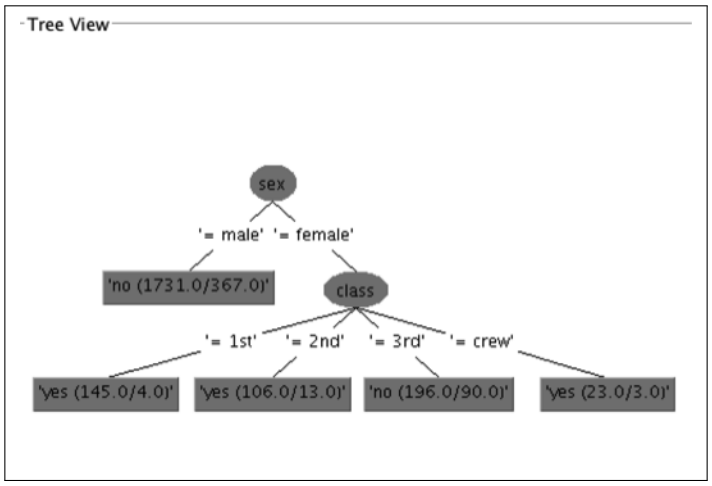

Figure 2: titanic-age-numeric.arff dat trained tree.

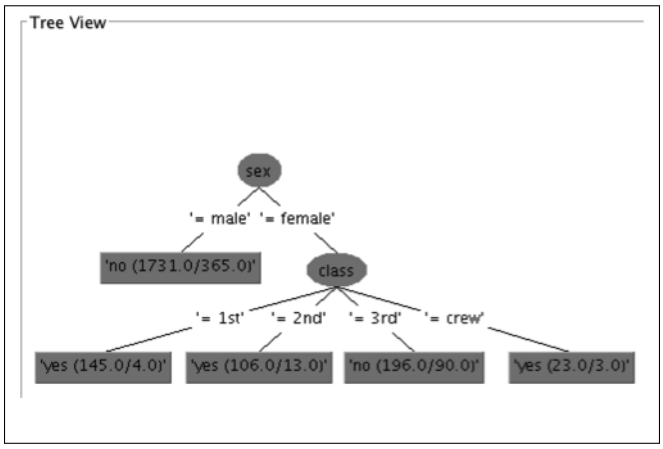

Figure 3: titanic_noise.arff dat trained tree.

1.3 Question 3: Compare results of Figure 2 and Figure 3

As the the two figures showed, the learning results are different. For Figure 2, the "male" classfication value is "no (1731/367)", while for Figure 3, the corresponding value is "no (1731/365)". The noise would slightly influence the classification performance for some instance in the "male" subclass.
I think, the noise-trained tree would make mistakes on instances with value "no (1731/366)". Those instances with the related values falling into the difference gap of the two learned tree would lead the classification tree make mistakes.

2 Part B

Just see the SetupSearchData.py and output.csv.