CSCI/ARTI 8950 Machine Learning

Assignment Number 1: Due 2/8/2021 **(by eLC)**

For this assignment you will need to use a decision tree learning package. You can use the Weka package downloadable from

`http://www.cs.waikato.ac.nz/~ml/weka/index.html`

and I recommend it. Alternatively, you can use another package of your choice or write your own code.

1. [**50 points**] For this part you will experiment with the *PlayTennis* data from the lecture notes about Decision Trees (Slide 7).

   (a) Use your decision tree learner to learn a decision tree based on all 14 instances. Print the tree if your software easily allows this or hand type it if not. Do you get the same decision tree as the one in Slide 8 of the Decision Tree lecture notes? If not, find out why (it may be because the learner is using a different information gain measure).

   (b) Use the leave-one-out cross-validation method described in class to estimate the error of the decision tree. You can do this manually by removing one example at a time and training on the remaining 13 and then testing on the one you removed, or you can use the package to do this for you automatically by asking for a 14-fold cross-validation if the package supports cross-validation (almost all packages do).

2. [**50 points**] For this part you should use a data set with at least **400** instances. You can choose one of the data sets in the UCI repository at

   `http://archive.ics.uci.edu/ml/`

   . You may use any data set from this web page provided that you specify which one you used. Alternatively, you can use any data set that comes with the Weka package or the package you use. Many of the data sets in the repository have missing attribute values but most packages can handle this automatically.

   (a) Use your full data-set to learn a decision tree. Give the tree, the error on the training set and the time needed for learning (if your package gives the learning time). **Do not use pruning**

   (b) Use 10-fold cross-validation to estimate the error (again without pruning).

   (c) Repeat the learning on the full set and the 10-fold cross-validation with pruning allowed. You may use any pruning method you like but you should describe it. Compare in a **table** between the error on the training set and the 10-fold cross-validation with and without pruning. Comment on the time needed for learning with and without pruning (if your package gives the learning time)