Problem Set 04

1. Animals (3p)

By default, WEKA takes the class from the last column in the arff file and uses all other columns as attributes. Take the zoo dataset and only use 'hair', 'airborne', and 'type' as attributes to predict the class 'eggs' with a C4.5 tree (J48). Visualize the tree.

2. Bonsai (4p)

The C4.5 algorithm implemented in WEKA allows us to choose between a pruned and an unpruned tree. Build both trees for the Bank-account dataset (bank.arff), visualize the trees, and compare the results. What is the purpose of using pruning? Which are the two pruning strategies that exist? Explain them briefly with an example.

3. Compare (3p)

Take the five datasets and use the three classifiers Naïve Bayes, Simple Rule, and C4.5. For each dataset-classifier tuple, retrieve the percentage of instances correctly classified. What are your conclusions (without doing any statistical tests)?

Deadline:

October 31, 2016 at 8:00 AM