FIT5047 – Intelligent Systems
Tutorial on Supervised Machine Learning

Question 1: Introduction to weka and categorical attributes

1. Download and install WEKA. Please note that the Explorer and Tutorial files are somewhat out of date. Use also the `PACE_Bootcamp_TS2_WEKA_Intro` file. If you can't find J48, you need to install simpleEducationalLearningSchemes from the WEKA package manager.
2. Get WEKA started by clicking on the Explorer button, or by downloading an `arff` file and clicking on it.
3. Open the `weather.nominal.arff` dataset. Visualize the different variables, and postulate which variables are significant. Why?
4. **J48** (which is algorithm C4.5)
   (a) Run `J48`, with X-validation, and analyze the resulting decision tree. Specifically, trace the computations and changes in class at each node down the different paths in the decision tree.
   (b) Calculate **manually** the Information Gain for `outlook` in level 1 of the tree. Compare this Information Gain with that obtained for `wind` in the lecture.
   (c) Copy `weather.nominal.arff` into your local user space, and try adding the following instances to the ARFF file (you can use `WordPad`):
   ```
   rainy,boiling,high,TRUE,yes
   hot,high,TRUE,yes
   ```
   What happens? How would you solve these problems?
   (d) Now, remove these instances, and try adding the following instances to the ARFF file:
   ```
   overcast,mild,?,FALSE,yes
   rainy,mild,high,TRUE,yes
   rainy,hot,high,TRUE,?
   ```
   What happens? What is the difference in the resultant decision tree?
5. Briefly explain the meaning of the summary measures produced by WEKA, i.e., Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, TP Rate, FP Rate, Precision, Recall, F-measure and ROC.
6. **NaiveBayes**
   (a) Run `NaiveBayes` on the `weather.nominal.arff` dataset. Explain your results.
   (b) Calculate **manually** the probability of playing ball, given that the outlook is overcast, the temperature is hot, the humidity is normal, and the wind is weak.
7. **IBk** (which is k-NN)
   (a) Run `IBk` on the `weather.nominal.arff` dataset. Explain your results.
   (b) Use the Jaccard coefficient to calculate the similarity between the following two data records.
   ```
   sunny, hot, high, FALSE, no
   sunny, mild, normal, FALSE, yes
   ```
8. Compare the performance of `J48` with that of `NaiveBayes` and `IBk`.

Question 2: Continuous attributes

1. Copy `weather.arff` into your local user space, and try running `J48` on it. What happens?
2. Now remove the numeric attributes using the supervised `AttributeSelection` filter, which you will find by clicking `Choose` under the `Filter` heading in the `Preprocess` tab. Analyze your results.
3. Run `J48`, `NaiveBayes` and `IBk` on `weather.arff`, and analyze the resultant summary values. Compare the performance of the three algorithms, and also compare the performance of each algorithm with its performance on the nominal dataset.