COMP30120 Assignment 2: Feature Selection in Weka

Deadline: Thursday October 29th 2015

Submission: Submit your report as a single PDF file via the COMP30120 CS Moodle page. Include your full name and student ID number in the report.

Overview:
The objective of this assignment is to use the feature selection functionality in Weka to identify useful and redundant features in data and to propose a subset of features that would be effective for building a classifier.

For your assignment, you will need to make use of two datasets described below. When downloading the datasets, please ensure your student number is correct. Submissions using an incorrect dataset will receive a 0 grade.

Dataset 1:
The first data source relates to high/low predictions for the likelihood of using a public bike sharing scheme on a particular day.

You should download your personal dataset from the URL:
http://mlg.ucd.ie/datasets/comp30120/bikes/<STUDENT_NUMBER>.arff
For example, if your student number is 126023491, your dataset is at the URL:
http://mlg.ucd.ie/datasets/comp30120/bikes/126023491.arff

Dataset 2:
The second data source relates to win/loss predictions for basketball games based on various statistics collected for those games.

You should download your personal dataset from the URL:
http://mlg.ucd.ie/datasets/comp30120/basketball/<STUDENT_NUMBER>.arff
For example, if your student number is 126023491, your dataset is at the URL:
http://mlg.ucd.ie/datasets/comp30120/basketball/126023491.arff

Tasks:
Write a report which addresses the following 4 tasks for both datasets:

1. Apply one filter and one wrapper feature selection technique from those available in Weka and report the feature subsets that they select.
2. Quantify and comment on the overlap between these alternative feature subsets.
3. Discuss the performance of these feature selection techniques when combined with two different classifiers of your choice available in Weka (i.e. there will be four experimental combinations for each dataset).
4. When Information Gain is used as a filter in feature selection some features will score 0, suggesting that they have no predictive power and can be ignored. Yet in

the datasets for this assignment some features that have an Information Gain score of 0 are selected by other feature selection methods. See if you can find examples of this and discuss why this might occur.

The recommended page length for the report is 2-4 pages, although there is no penalty for exceeding this length.