NTUST, CSIE
Machine Learning (CS5087701), Fall 2018

Homework 2 (6pts)

Due date: Nov. 13

Question 2.1. [6pts] Analyze the following two datasets:

the *Bank Marketing* dataset
(http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) from
UCI (http://archive.ics.uci.edu/ml/), and
the Spooky Author Identification from Kaggle
(https://www.kaggle.com/c/spooky-author-identification).

You are recommended to use *C*4.5 (or *C*5.0 that you can use in our lab), the decision tree from scikit-learn of Python or from Weka. After your analysis, you should write a mini report of around one or two pages with a discussion section. In your report, you should include the following items:

(a) List all the parameters for the models that you used.

(b) The prediction accuracy with cross-validation and possible different data partitions.

(c) Explain the result you obtain, e.g., why you have a particular attribute as the root of the tree, the tree size etc.

(d) Which dataset is the one we obtain a better result from decision trees? Why?

(e) Give the reasons why the result is good (or bad) for different experimental settings (pre-pruning, post-pruning strategies, etc.).

(f) (Bonus) Can you suggest any approach for re-building the tree or revising the tree so that the prediction result is better? (hint: manually selecting some particular attributes, transforming the attributes from categorical ones to numerical ones or the other way around.)