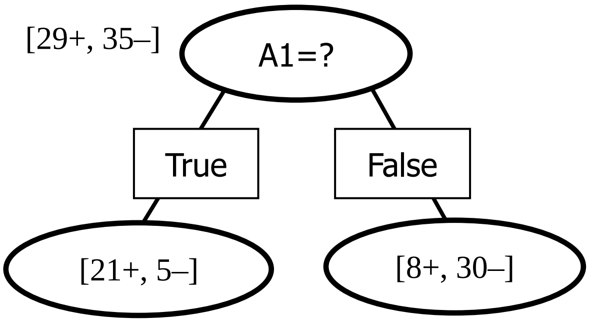
NTUST, CSIE
Machine Learning (CS5087701), Fall 2017

Homework 2 (12pts)

Due date: Nov. 14 (Q2.1 and Q2.2 for the first result) & Nov. 28 (the rest)

Question 2.1. [2pts] Consider the following tree splitting and two questions.



(a) A statement says: "To decide the best attribute in each splitting of decision induction, instead of computing *information gain*, you just need to compute the expected (average) entropy in the lower level". Do you think it is correct or not?

(b) The similar question is asked again, but the criterion "information gain" is substituted by "gain ratio". What is your answer then?

Question 2.2. [10pts] 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 or ANN algorithm from scikit-learn of Python, or the ANN algorithm (called multilayer perceptron) from Weka. After your analysis, you should write a short report and the report should be around three pages with a discussion section. The discussion part should be at least one full page long.

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, how the hidden layer(s) and how many nodes in your ANN can influence the result, etc.

(d) Give the reasons why the result is good (or bad) for different experimental settings (pruning strategies, the number of iterations in ANN, etc.).

(e) (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.)

(f) (Bonus) Can you suggest some rules to decide the network structure for decision tree or ANN in general?