Assignment 4 (Knowledge Discovery and Data Mining)

Number of problems/points: Six problems for total of 100 points
Out: February 18, 2016
Due: February 29, 2016 by 12pm in SERC 336 (hard copy).

Problem 1: (10 points)
Solve exercise 1 in section 10. in the textbook (page 343).

Problem 2: (10 points)
Solve exercise 2 in section 10. in the textbook (page 343).

Problem 3. (10 points)
Show that the entropy of a node in a decision tree never increases after splitting it into smaller successor nodes.

Problem 4: (10 points)
Solve exercise 3 in section 10. in the textbook (page 343).

Problem 5: (10 points)
Solve exercise 13 in section 10. in the textbook (page 344).

Problems 6: (50 points)
Find a freeware decision tree classification software (e.g. C4.5) or develop one on your own.
   (i) Test how it decides when to stop growing a decision tree or how it prunes a complete tree. Test how it handles real valued attributes. Report the findings.
   (ii) Apply this classifier on the initial 150 examples of Adults dataset from the homework 1. Use the holdout method (with the initial two thirds of the data used for training and remaining one-third for testing) to estimate accuracy of decision tree classification. Report your findings.
   (iii) Compare this to using ten-fold cross-validation. Report your findings.
   (iv) Repeat experiment (iii) over 100 different folds each time using a random 2/3 fraction of data for training. Report your findings.
   (v) Repeat experiment (iii) using the initial 3000 examples. Report your findings.