Assignment 1 (given by Zoran Obradovic)

Due: February 8 in class.

Problem 1:
Read Chapter 4 in the textbook. Summarize (in your own words) up to a half page what you learned in this chapter.

Problem 2:
Solve problem 1.4 in the textbook (page 18)

Problem 3:
Solve problem 2.5 in the textbook (page 48)

Problem 4:
Solve problem 3.4 in the textbook (page 77)

Problem 5:

(a) Perform pre-processing by removing outliers and discretizing attributes to two values (lower vs. higher than medium). Report the results.

(b) Use a 5 cross-validation process to estimate the precision accuracy of a decision trees trained (using either ID3 or C4.5 algorithm) to discriminate diabetes and non-diabetes cases. Compare generalization results of full trees vs trees obtained after prunning. Extract the rules and check their stability (if they change with the training set and to what extend). Report your findings concisely.

(c) Extra credits: Explore if you can obtain higher accuracy by applying more advanced features pre-processing and more advanced tree pruning methods (hint: various methods are implemented in Weka package WEKA).

Note 1: If you use a decision tree software found somewhere make sure that you understand what it does. One that is often used in the machine learning community is Weka software. It is available at http:www.cs.waikato.ac.nz/ml/weka In CIS lab you can find decision trees at the statistical toolbox of Matlab (version 7). You are also allowed to write your programs.

Note 2: Solutions are expected to be your own work (no group work is allowed). Penalty for late homework submissions will be zero points on the assignment (so, do it on time).