CS771:Machine learning: tools, techniques, applications
Assignment #1: Decision trees

Due on: 18-1-2015, 23.00 12-1-2015
MM:85

1. In this assignment you have to use the Pima Indian Diabetes data set available from the UCI repository:
   https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

   (a) Build a decision tree (DT) classfier using the following three strategies for stopping growth of a tree: i) threshold on the decrease in impurity ii) threshold on the number of data vectors at a node iii) grow the tree as much as possible and then prune.

   (b) Experiment with at least two different impurity functions.

   (c) You will have to handle some errors in the data like absurd values and/or missing data for some attribute values. Carefully, report how you have handled such data errors and what is the reason for the choice you made.

   In all cases report five fold cross validated values for accuracy.

   The problem is a medical diagnosis problem where we expect DTs to work well.

   You can use any ML library that supports DTs. Python libraries, R and Weka have support for DTs to varying levels. Other libraries/packages in other languages that handle DTs also exist.

   [45,20,15=85]