CSE471: Statistical Methods in AI -- Spring 2016
Assignment 6: SVM & DT

DUE: Before 12:00 midnight on 13 Apr 2016 (Wednesday)

INSTRUCTIONS:

1. You may do the assignment in Matlab/Octave, R, Python, C/C++ or Java.
2. You need to upload pdf files in the Course Portal. One file should contain your answers, results and analysis. A separate file should contain code you have written and its sample output.
3. At the top-right of the first page of your submission, include the assignment number, your name and roll number.
4. *IMPORTANT:* Make sure that the assignment that you submit is your own work. *Do not copy any part from any source* including your friends, seniors or the internet. Any breach of this rule could result in serious actions including an **F grade** in the course.
5. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

Preamble:

The aim of this assignment is to experiment with *Support Vector Machine* (SVM) and *Decision Tree* (DT) techniques we learned in the class on real world problems.

A. Support Vector Machine (SVM) Classifier

(A.1) Employ SVM classifier (using any standard library such as LibSVM) for linear and RBF kernel on two datasets in the K-dimensional PCA space (where K=10, 100 for the UCI Machine learning dataset: Arcene Cancer Classification data and any two K values of your choice for another dataset of your choice that has real vectorial data representation). Report the following evaluation results using 5-fold cross validation on your dataset: Precision, Recall/Sensitivity, Specificity and Accuracy. Present an analysis and discussion of your results.

B. Decision Tree (DT) Classifier

(B.1) Write your own code for implementing ID-3 Decision Tree (DT) classifier using information gain measure as attribute splitting criterion. Employ DT classifier on the two datasets from UCI Machine learning repository. Select two datasets that have discrete attributes, otherwise discretization scheme needs to be used for real attributes. Report the following evaluation results using 5-fold cross validation on your dataset: Precision, Recall/Sensitivity, Specificity and Accuracy. Present an analysis and discussion of your results. List the rules discovered by the DT classifier on the best fold (in 5-fold CV experiments) for the data sets. Apply linear SVM on one of the datasets and extract support vectors (SVs). Use the SVs and construct DT and list the resulting rules from the SVM+DT combination. Compare the rules extracted from DT directly versus SVM+DT combination and comment on the results.