Assignment # 2

The assignment can be done in Python or Matlab.
You need to submit both your report and the source code implementation for all questions.

Question#1 (3)

Suppose that instead of selecting a node using information gain (IG) in a binary decision tree, we select a node randomly from nodes with IG>0:

a) Show that each leaf of the tree contains at least one training data**.**

b) If we have n training data, what is the maximum number of leaf in constructed decision tree? Compare result with the state that we used IG for selecting node.

Datasets Description

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

Datasets are available on http://archive.ics.uci.edu/ml/datasets.html.

For this assignment, you need to download the datasets “Tic-Tac-Toe Endgame” and “Wine” from the above link which are a categorical and a continuous dataset, respectively.

Question#2 (12)

a) Design an ID3 decision tree classifier (Based on Information Gain) to classify each dataset mentioned above. Report the accuracy based on the 10-times-10-fold cross validation approach. You should report mean and variance of the accuracy for each dataset. Draw the confusion matrix for each dataset and analyze your observation.

b) Try the same experiment by using the Gain-Ratio instead of Information Gain in constructing the decision tree. Is the result similar to part (a), why?

Queation#3 (10)

The physical sources of noise in machine learning can be distinguished into two categories: (a) attribute noise; and (b) class noise.

In this question you should check the effect of noise in both categories.

A) Analysis the effect of attribute noise (5)

In this section, for the dataset D, first split it into a training set X, and a test set Y. Train a classifier **C** from X, use **C** to classify instances in Y, and denote the classification accuracy by CvsC (i.e., Clean training set vs. Clean test set). Then corrupt each attribute with a noise (this should be done based on the provided noise percentage and decide whether you should add noise to that attribute or not) and construct a noisy training set X’. Learn classifier C’ from X’, use C’ to classify instances in Y and denote the classification accuracy by DvsC (i.e., Dirty training set vs. Clean test set). In addition, also add the corresponding levels of attribute noise into test set Y to produce a dirty test set Y', and use classifiers C and C' to classify instances in Y'. Denote the classification accuracies by CvsD and DvsD respectively (i.e., Clean training set vs. Dirty test set, Dirty training set vs. Dirty test set). For each dataset, use decision tree (ID3) as a classifier and execute 10-fold cross validation 10 times, and report the average accuracy and its variance as the final result.

- Note: to create these noises, select L% of training data randomly and add a noise (from zero mean normal distribution) to the attributes of the sample in the numerical case. For the nominal case, flip the attribute randomly to another value.

a) Plot one figure for each data set that shows the classification accuracy respect to different attribute noises with the values of (5%, 10%, and 15%). It should be noted that the x-axis and y-axis show noise level and accuracy, respectively. Each figure should contain 4 curves for CvsC, CvsD, DvsC, DvsD results.

b) Analysis the results according to the plots.

B) Analysis the effect of class-label noise (5)

There are two possible sources for class of noises:

c) Contradictory examples. The same examples appear more than once and are labeled with different classifications

d) Misclassifications. Instances are labeled with wrong classes. This type of errors is common in situations that different classes have similar symptoms

To evaluate the impact of class noise, you should execute your experiments on both datasets, while various levels of noise are added.
Then utilize a decision tree learning algorithm to learn from these noisy datasets and evaluate the impact of class noise (both Contradictory examples noise & Misclassifications noise) on them.

- Note: to create these noises, select L% of training data randomly and change them. (Try 10times-10fold cross validation to calculate the accuracy/error for each experiment.)

B) Plot one figure for each data set that shows the classification accuracy in terms of different noise levels (5%, 10%, and 15%). Plot two type of noises overlaid in one figure.

C) How do you explain the effect of noise on ID3 when there is not any pruning variable?

D) In comparison with attribute noise and class noise, which is more harmful? Why?