In The Name of Allah
Machine Learning (winter 2021)
Instructor: Mahdi Yazdian
TA: Atefeh Mohammadi

Homework#2: Decision Tree
Due Date: 1400.01.27

Part A. Theoretical Example

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.

Suppose that you have n non-overlapping points in [0, 1]*[0, 1] space with + and – labels. Also suppose you can select one feature in different levels:

c) Prove that there exist a tree with at least depth of $\log_{۲} n$ that can classifying data truly.

d) Give an example with n points in defined space that minimal decision tree for them, contains n-1 internal nodes.

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.
Its datasets are available on http://archive.ics.uci.edu/ml/datasets.html.
For this homework assignment, you need to download the datasets “glass” and “Tic-Tac-Toe Endgame” from the above link which are a categorical and a continuous datasets, respectively.

Part B. Analysis the effect of attribute noise

The physical sources of noise in machine learning can be distinguished into two categories: (a) attribute noise; and (b) class noise. In this part you should check the effect of noise in both categories.
In this section, for the dataset D, first split it into a 70% for training and 30% for testing 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 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 (C4.5) as a classifier.

- Note: To add noise to features, generate white Gaussian noise with $N(\cdot,\Sigma)$ and **add** it (additive noise) to the samples, where Σ shows the variance of the noise.

١. Plot one figure for each data set that shows the classification accuracy respect to different feature noises with the variance of (5%, 10%, and 15%) of the feature variances. 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.

٢. Analysis the results according to the plots.

Part C. Analysis the effect of class-label noise

There are two possible sources for class noise:

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

f) 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 executed your experiments on both of datasets, where various levels of class noise are added. Then utilize C4.5 learning algorithms to learn from these noisy datasets and evaluate the impact of class noise (both Contradictory examples noise & Misclassifications noise) on them.

- Note: to create label noises, select L% of training data randomly and change them.

١. Plot one figure for each data set that shows the classification accuracy in terms of different label noise with the level of (5%, 10%, and 15%) of samples. Plot two type of noises over one figure.

٢. How do you explain the effect of noise on C4.5 method?

٣. In comparison with attribute noise and class noise, which is more harmful? Why?

Implementation Note:

- Implement your codes as a functional form.

Report:

- Prepare a report in PDF format including the figures, answer to the questions and discussions mentioned in the homework.
- Make a folder including your report and you codes (Note that your code is needed to be self-comment)
- Submit all things in a zipped folder named as “UrName_UrFamily.rar”

Good Luck