Assignment # 2

SYDE 675 Winter 2019

The assignment can be done in Python or Matlab.
You need to submit both your report and the source code implementation for all questions. The report must be a single pdf and the source code must be a single .py or .m file. Please include brief comments in your code. Be sure to label all figures and include a legend where appropriate.
The due date for this assignment is **March 22th, 2019**. Please also note that late submission will be subject to a penalty of 20% deduction of the assignment mark per day.

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 homework assignment, you need to download the datasets "glass" and "Tic-Tac-Toe Endgame" from the above link. The "glass" dataset is categorical and the "Tic-Tac-Toe" dataset is continuous.

Question 1

Design a C4.5 decision tree classifier to classify each dataset mentioned above. Report the accuracy based on the 10-times-10-fold cross validation approach (20% of training set as the validation set for every experiment). Report the mean accuracy and the variance of the accuracy for each experiment.

Question 2

There are two possible sources for class label noise:

a) Contradictory examples. The same sample appears more than once and is labeled with a different classification.

b) Misclassified examples. A sample is labeled with the wrong class. This type of error is common in situations where different classes of data have similar symptoms.

To evaluate the impact of class label noise, you should execute your experiments on both datasets, while various levels of noise are added. Then utilize the designed C4.5 learning

algorithm from Question 1 to learn from the noisy datasets and evaluate the impact of class label noise (both Contradictory examples & Misclassified examples).

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

a) Plot one figure for each dataset that shows the noise free classification accuracy along with the classification accuracy for the following noise levels: 5%, 10%, and 15%. Plot the two types of noise on one figure.

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

Question 3

Design a feature selection algorithm to find the best features for classifying the Mnist dataset. Implement a bidirectional search algorithm using the provided objective function as the measure for your search algorithm.
Use the first 10000 samples of training set in the Mnist dataset for feature selection and training set for kNN approach. Use **Euclidean distance** to calculate Inter-class and Intra-class distance.

The objective function should be based on this equestion:

$$J = \frac{Inter{-}class}{\sum_{cl} Intra{-}class/(num\_samples\ in\ cl)}$$

If both Inter-class and Intra-class distances are zero use $J = \infty$

a) Select the set of {10, 50, 150, 392} features based on the implemented feature selection approach and report the accuracy on the test set of MNIST based on kNN with k = 3. Note: you can take advantage of data structure tricks to speed up the efficiency of kNN algorithm.

b) Visualize the selected features for each set in {10, 50, 150, 392} by a zero 2-D plane where the selected features are pixels set to a value of 1. Compare the 4 different planes.

c) Apply LDA on the dataset and report the accuracy based on kNN with k =3. Compare the achieved accuracy by the reported accuracies in part (a). Note: you need to implement LDA method by yourself.