Assignment 2

Machine Learning (CS564)

Date: 29-Oct-2020 **Deadline: 09-Nov-2020**

Instructions:

1. You are supposed to write code from scratch for filter methods but you can use libraries for pre-processing and post-processing.
2. The assignments should be completed and uploaded before the deadline.
3. Markings will be based on the correctness and soundness of the outputs. Marks will be deducted in case of plagiarism.
4. Proper indentation and appropriate comments are mandatory.
5. You should zip all the files and name the zip file as roll_no.zip, eg. 1701cs11.zip.
6. Make necessary assumptions if required. For further clarification, you can write an email: cs5642020@gmail.com

Upload your zip file to the link: https://www.dropbox.com/request/3JcvfpoOpgTDnDlSLvsq

Dataset: https://www.dropbox.com/s/ign7443iz02b4v9/heart.csv?dl=0

Dataset Information: https://www.kaggle.com/ronitf/heart-disease-uci

PROBLEM

In this assignment you are to implement a decision tree classifier and use it to classify whether the person has heart disease or not. There are 14 features in the dataset, with the final column for target with values as either 0 or 1. Use max depth as an early stopping criteria. If no max depth is specified you model should train the whole tree.

Q1. Implement the classifier using information gain and gini index. Compute and compare the accuracy of the model for different options. Set max depth as 3, max split as 5 and train-test split as 70%. Compare the results with the logistic regression and decision tree classifier available in scikit learn library.

Q2. For each one of information gain and gini index,

1) Train the model with [50, 55, 60, 65, 70, 75, 80, 85, 90, 95]% of data as the training data, i.e. increase the percentage by 5% and find the corresponding accuracy. Do you see overfitting? Plot accuracy wrt the training data.
2) Train the model with [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] max depth, and find the corresponding accuracy. Plot accuracy wrt the max depth.