In the name of God
Machine Learning (Spring 2020)
Assignment #1
Decision Tree

Due date: 2th of Ordibehesht

In this assignment, you are about to implement HDDT, which is a decision tree based on Hellinger distance, and it is suitable for unbalanced data. Read the attached documents for more information about the HDDT algorithm and its implementation.

You are required to evaluate the performance of HDDT on the Coronavirus data set. Since the data set is unbalanced, you should use Precision, Recall, F-measure, and AUC measures to evaluate the performance of your tree. It is worth mentioning that these measures are one-class measures; in other words, you should compute these metrics just for the minority class.

Split the data set to train and test parts. Use 70% of the data for training phase and the remaining 30% for testing phase. Run your codes for 10 individual runs and report the average of 10 runs for each performance metric.

In each part of the assignment (step 1 through 3), compare the performance of the trees with the following classifiers: Naïve Bayes, One-nearest-neighbor (OneNN), linear SVM, and kernel SVM (use RBF kernel). Feel free to use built in classifiers. Note that the original version of the data set (Con_Covid-19.csv) has continuous features, which should be used to train these four classifiers. Use the discretized version of the data (Dis_Covid-19.csv) to train the HDDTs.

Step1:

Since the data set has three classes and HDDT is a two-class algorithm you are required to convert the data to a two-class data set by keeping the smallest class as minority and the rest as majority. Implement the two-class HDDT and evaluate its performance in terms of all mentioned metrics.

Step2:

After you implement the two-class HDDT, you need to extend it to handle the multiclass data with OVO (One Versus One) and OVA (One Versus All) approaches. Use the original version of the data sets without converting it to multiclass (Dis_Covid-19.csv). Then evaluate the performance of both approaches with the mentioned metrics and report the results.

Step3:

Repeat both of the experiments in previous parts with the pruned HDDT trees. Simply put, consider the longest height of the tree and prune it with different heights of $\frac{MaxHeight}{2}, \frac{MaxHeight}{3}, \frac{MaxHeight}{4}, \frac{MaxHeight}{5}$ where MaxHeight is the height of the HDDT tree. Compare your results with the unpruned versions of the trees and report the results.

Questions:

What are the properties of the HDDT algorithm? Why is it suitable for unbalanced data?

What are the differences between Hellinger distances, Gini index, and information gain?

Is pruning lead to better results? Why?

Important Notes:

- Pay extra attention to the due date. It is fixed and **will not be extended**.
- Be advised that no submissions after the deadline would **be graded**.
- Be sure to comment your code. Also, include instructions on how to run your code (if necessary).
- Prepare a complete report for you assignment and answer all the questions.
- The name of the uploading file should be your **Lastname_Firstname.**

Good Luck