Machine Learning

Programming Assignment 1

Due Date: March 8, 2020

Assignment Guidelines

1. Any kind of plagiarism is not accepted. We will strictly follow institute policies for plagiarism.
2. Recommended programming languages: Python + Sklearn (only for classifiers)
3. *Submission should include:* Working code for each question separately and a report to show the analysis of results and corresponding plots in each of the parts. Make sure to append all the relevant details and plots in your report to support your results. You don't have to upload any of the result in the .zip file.
4. *Submission Guideline:* upload a single .zip file with name 'YourRoll_PA1.zip' and each solution should be a .py file with name 'YourRoll_Q1.py' (consider for first question) and a report file 'YourRoll_report.pdf'. No other file should be added in the folder.
5. Your code should run as $python3 <YourRoll_Q2.py> <path to the dataset containing the train and test set downloaded from the link.>
6. Any error encountered in running the code for TA will fetch you a straight zero for that particular question.

Total: 80 marks

Question 1. Evaluation metrics and evaluation metrics.

1. Implement ID3 Algorithm for Decision Tree discussed in class from scratch, the data can be downloaded from this **link**. You have to use entropy based Information Gain calculation method (as discussed in class) to evaluate different splits. Report Accuracy, Confusion Matrix and F1 Score in the report. What do you think is a good measure - Accuracy or F1 score, support your answer with proper claims.
2. Implement Decision Tree using sklearn library on the same data. (using gini index to calculate Information Gain).
3. Compare results of A and B, report the analysis. **25 + 5 + 5 = 30 marks**

Question 2. Download the dataset from this **link**. Perform multi-class classification using Decision Trees and Random Decision Forest (RDF). You can use Sklearn library. You can perform hyperparameter tuning to improve your results and mention them in the report.

1. Report the following results on the training and test sets: **5 * 6 = 30 marks**
   (a) Accuracy
   (b) Confusion matrix
   (c) Precision and Recall
   (d) Sensitivity and Specificity
   (e) ROC curve (one for Decision tree and one for random forest)
   (f) Plot and visualize your decision tree (Only for Decision Tree, not Random Forest).
2. Compare and report accuracy achieved using Decision Trees and Random Forest on test set provided in question 2 above and suggest which classifier is better and why. Plot a single ROC curve corresponding to both the classifiers and compare their Area Under the Curve (AUC) scores and include them in your report.

5*2 + 5 + 5 = 20 marks