Ghulam Ishaq Khan Institute of Engineering Sciences and Technology

CS351 Lab Semester Project Spring 2021

Deadline: May 11,2021 **Total Marks: 40**

Note: Lab policy about plagiarism is as follows:

- **Students must not share actual program code with other students.**
- **Students must be prepared to explain any program code they submit. (Viva's schedule to be announced later)**
- **Students must indicate with their submission any assistance received.**
- **All submissions are subject to plagiarism detection. For report and code maximum 15% plagiarism is acceptable.**
- **Students cannot copy code from the Internet.**
- **Students are strongly advised that any act of plagiarism will be reported to the Disciplinary Committee.**

---

1 Project Description:

1.1 Objectives:

The objective of the project is to apply the different classification and clustering algorithms to the problem of classifying cyber-attacks in network traffic. This will help your retention of the material and significantly enhance the depth of your understanding. This will also develop your skills in reporting the performance. The project is expected to consume roughly two weeks of moderately concentrated effort. We encourage you to work in a group of (maximum) two students.

1.2 Dataset Description:

The dataset is available in the file section of the teams. It comprises of two text files, the "Dataset.txt" file contains the complete dataset. Each column specifies the attributes of network traffic, out which some or all may be considered as features for classification. The other file "Attack_types.txt" summarizes the possible attack types.

1.3 Scope of Work

For the given dataset, we want to develop classifiers for the prediction of attack type, given their attributes.

The scope of the project includes

- formulation of the problem under consideration.
- cleaning and pre-processing the data.
- apply feature engineering (if needed).
- implement the classification and clustering algorithm. (Details below)
- report the performance of the algorithms and presentation of analysis/findings.

2 Expectations:

There are three components of assessment in the project:

- Report (15 marks)
- Code (15 marks)
- Three minutes video presentation summarizing your work (10 marks).
  Note: Video presentation must include the video of the presenter.

Your report is expected to have the following sections and should be around 600 to 700 words.

1) Abstract (executive summary),

2) Introduction,

3) Data-preprocessing (extraction and cleaning),

4) Feature Engineering (e.g., dimensionality reduction),

5) Use of the following classification and clustering algorithms; subsections on ensemble learning, K-Means/any other clustering algorithm, ANN.

6) Comparison and Performance Evaluation (plots, tables etc.) Detailed comparison is expected for task 2 and 3.

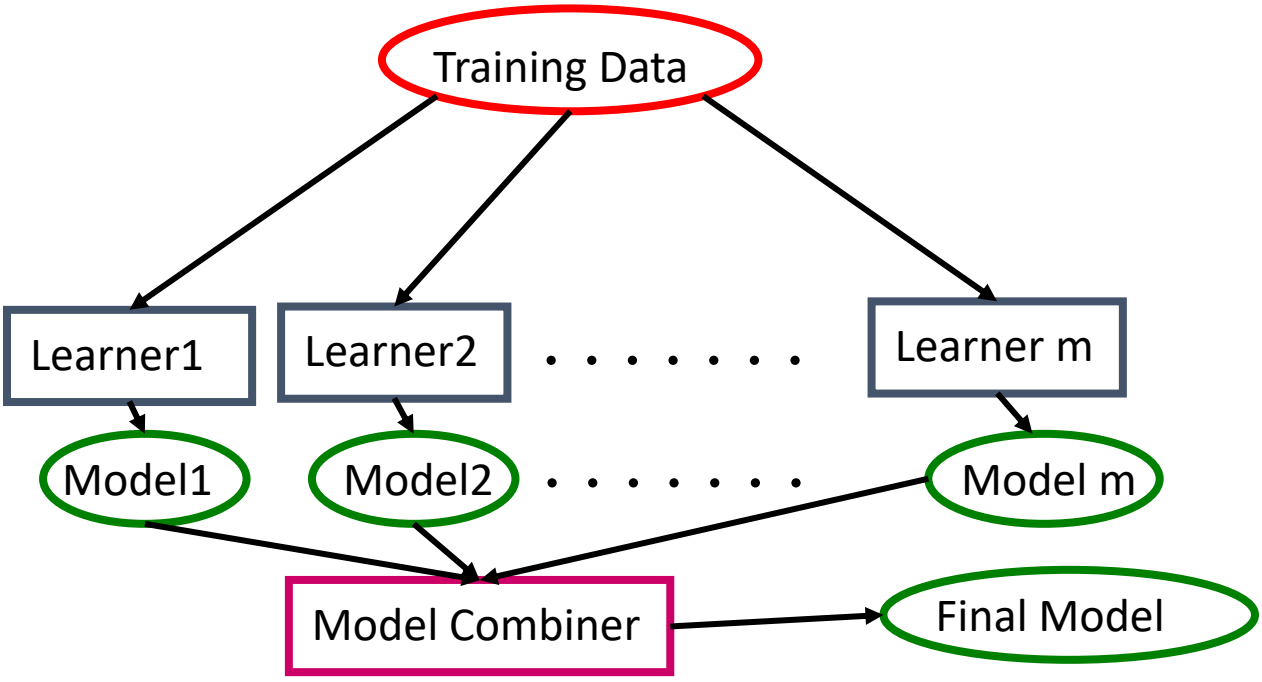
7) Conclusions.

Implementation Details

Note: You may use language of your own choice.

Once the data is preprocessed and ready to be used for learning, you have to perform the following tasks:

1. The dataset provided to you contains 23 different classes (attack types). You need to convert it to 5-classes. (Hint: See the file "Attack_types.txt")
2. The next step is to use the updated dataset for ensemble learning i.e., you will be training different variants of ANN and at the end you will make the decision based on predictions from all ANN variants as show in Figure. The variants of the ANN will be based on their parameters. You must train at least 3 variants.
   Bonus: Implement any other classification algorithm i.e., SVM, kNN, etc within this model.



3. Your last task includes classification followed by clustering. In this task, you will drop the column containing labels from the dataset and label the dataset using any clustering algorithm e.g., k-Means. Clustering will return label for each record. Now again perform the ensemble learning on the newly labelled dataset.