Homework4

ECE 631A Cyber-Security of Critical Infrastructure

Due Date: October 6, 2019 by 11:55 PM via canvas

This homework is for you to use decision tree based learning as well as PCA based dimensionality reduction followed by any classification algorithm -- to create two models to distinguish between stressed grid condition and nominal grid condition. How the data is generated is described in the paper "Methodology for a Security/Dependability Adaptive Protection Scheme Based on Data Mining" by Bernabeu, Thorp, Centeno.

Check particularly in Section IV A of the paper to know how the data was generated.

The very first column in the data table is marked as '0' if it is a normal condition data and marked '1' if it is a stressed condition data. The data represents positive sequence voltages and angles at various substations, and line voltage/currents for different transmission lines.

However, for this homework, all we want you to do is to divide the data into 50% training set, 25% validation set, and 25% test set.

Find the best machine learning model (among the two methods -- Decision tree and PCA for feature reduction followed by classification).

There will be a competition among the groups as to who gets the best accuracy, and false negative. We will be testing on data that is randomly picked from the given data to test your model and see the accuracy and false negative figures reported by you are in the ball park range.

The final submission should contain a pdf file with report with brief description of each model followed by the table with accuracy, false positive, false negative etc as customary.

You must submit the entire code, the pdf containing report as stated, and any other relevant files to run your model – along with a README file detailing how to run your model.

A group of up to 3 people can submit one homework if they work together. The TAs will interview the entire team to understand the contributions of all the members, and ask the team members to run the model and show results.

The grading scheme will as follows:

Modeling including coding, and efficiency of the training and the actual classification on a few data: 60 points

Accuracy, Quality and details in the report: 20 points

Interview of team members during the demo run: 20 points (this will be different for each team member)

The team getting the best accuracy and false positive combination will get 20 bonus points.

The team with second best accuracy and false positive combination will get 10 bonus points.

The team with the 3rd best accuracy and false positive combination will get 5 bonus points.

The data is available at canvas website in Files → Week4 and Homework4 files.

Note that you have to randomly sort the data – before dividing it into training, validation and test sets – as currently the rows 2-1636 are for grid condition – normal and from 1637 onwards for grid condition stressed. You cannot train the classification models with only normal data. So you have to mix it up first before you get your data that can be divided into 50%, 25% and 25% for training, validation and testing.