EE769 – Introduction to Machine Learning 2020-21-2

Assignment – 2: Comparison of Classifiers

Description: The ministry of water of an African country has made a data set of pumps installed in various places in the country to serve the water for their countrymen. It is a tedious and expensive task to maintain these pumps. This data set contains information such as the kind of pump, when it was installed, and how it is managed. Can you predict which pumps require repairs and which are not functional by using the given data set? A smart understanding of water point failure can improve maintenance operations and ensure that clean and safe water is available to these communities. Experiment with a few methods such as tree based methods, bagging, boosting method, support vector machine etc. and comment on the suitability of the method to predict the classes. Before implementing any model, visualize the data and comment on the data statistics.

Data:

Training: https://drive.google.com/file/d/1-E-EfSZcREnhyMe23tluDmzNnS2pWSE_/view?usp=sharing

Testing: https://drive.google.com/file/d/1-498FRSmoGj1AqpuPIHU5P8tIwuHEzN1/view?usp=sharing

Suggested flow and evaluation criteria:

1. Data exploration and visualization with insightful commentary to determine potentially useful variables. [1]
2. Variable transformation, feature engineering, feature selection or elimination. [1]
3. Declaration of ML frameworks (e.g. SVM with Gaussian kernel) to be used with a prior hypothesis of which method is likely to work better (it does not matter if your initial hypothesis turns out to be wrong) due to insights from the previous steps. Declare some resources on the net that you read to find out which ML framework is better suited for which type of data. [1]
4. Diligent hyperparameter tuning for at least three frameworks. [3]
5. Determination of the relative order of importance of variables, and perhaps a second round of feature selection. [1]
6. Visualization of decision boundaries by either taking the two most important variables into account, or by reducing the input dimensions to two using tools such as t-sne. [1]
7. Submit the final labels (you should not touch the test data before this step) as a CSV file with a single column and no header. [1]
8. Make the notebook easy and insightful to read and declare inspiration sources. [1]