Lab 2 – Data Preprocessing

MACHINE LEARNING

SAUMAY AGRAWAL
16BCE1151

EXPERIMENT

- Explore the Iris dataset as described in the lab.
- Try out:
    - interpolation for missing values
    - cleaning the data
    - Additional plotting/ visualization of dataset desirable.
    - Repeat for a dataset of your choice.

NEED FOR DATA PREPROCESSING

Since the machine learning algorithms learn from the data, it is necessary to make sure that we are feeding the right data in our algorithms. Thus, data preprocessing is done to make sure that all the attributes of the data are fit for further use in the algorithms. This is done in multiple phases.

PHASE 1 – DATA SELECTION

Selection of data is essential. There is no need to start working with all the available data. Therefore, we should select the dataset that best represents our needs. This can be done by combining multiple datasets, and leaving out unwanted attributes from the dataset.

PHASE 2 – DATA FORMATTING

This step involves ensuring that every value in all the attributes field, confirms to the format of respective attributes. For example, there can't be alphanumeric or decimal values in the field of Age.

PHASE 3 – DATA CLEANING

Data cleaning involves the handling of missing values in the dataset. The missing values may be fixed or removed entirely. One of the most popular ways of fixing the missing values is substituting them by the mean values.

PHASE 4 – DATA SAMPLING

Often times the data size will become larger than the optimum size for practical machine learning processes, as a lot of exploration and experimentation is done with different models and algorithms. A large dataset can take up a lot of time in processing, apart from over usage of system resources. Sampling involves taking a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

OBSERVATIONS

IRIS DATASET

- Using the pairplot, we observed that for any pair of petal and sepal attributes, the data is easily classifiable between the classes 'setosa', 'virginica', and 'versicolor'.
- We found some missing values in the petal width attribute, which were substituted with the mean of all the petal width values.

OLYMPICS DATASET

- I started out by combining the Olympic dataset file with the regions.csv file, on the values of NOC (National Olympic Committee) codes. I dropped the 'notes' attribute field from the regions.csv before this step, as it was unnecessary to our use here.
- The resultant dataset didn't need any formatting, as all the values were confirming to their attribute formats.
- However, some fields needed data cleaning as they contained some missing values. The missing values were handled according to the context of the fields containing them.
    - **Regions** – Empty regions contained the athletes from refugee Olympics team, an island named Tuvalu. These regions were named after their 'Team' names. It also had 2 records with most of the values in other attributes missing, hence they were removed entirely from the dataset.
    - **Height, Weight and Age** – The missing values were substituted by the respective mean values of "same region". This was done as different regions have different mean values for the above fields in their population, due to geographical, cultural, lifestyle, and many other differences. The only two female athletes from 'Tuvalu' had their height missing, so substituted them by the mean heights of all the women athletes.
    - **Medals** – Null values represented that the athlete didn't win in event he participated. Simply substituted the missing values by the string 'None'.
- Since the values of the 'Games' field are nothing but the combination of respective 'Year' and 'Season' values, I dropped this field.
- I checked out the statistical distribution of various integer fields like 'Age', 'Height', 'Weight', etc. There wasn't anything logically impossible in these fields.

- Performed the data sampling by dividing the entire processed dataset into three parts:
    - Previous 20 years of Olypmics data (smallest)
    - Previous 40 years of Olympics data (medium-sized)
    - Previous 120 years of Olympics data (whole)

INFERENCE

- Data preprocessing is an essential part of any machine learning task.
- We should be sure about the data we are using for prediction and analysis
- Wrong data can have undesired and misleading consequences.
- We need to check for outliers as they have the potential to affect the results drastically, and we should also check for any illogical values in the dataset
- Data sampling is an essential and beneficial step for large datasets.