Certificate Course in Machine Learning using Python [6 Weeks]

Pre-processing of data for Machine Learning

Attempt: 1

Data Preprocessing

- In any machine learning process, data preprocessing is the step in which data is transformed or encoded so that the machine can process the data easily. The features of data can now be easily interpreted by machine learning algorithms.
- Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.
- It is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Why Data Preprocessing

- For achieving better results from the applied model in Machine Learning projects, the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format.

For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

- Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.
- The data has to be in proper format and any missing values must be processed before applying the Machine Learning algorithms.

Data Preprocessing Process

- Formatting the data to make it suitable for ML (structured format).
- Cleaning the data to remove incomplete variables.
- Sampling the data further to reduce running times for algorithms and memory requirements.
- Selecting data objects and attributes for the analysis.
- Creating/changing the attributes.

Data Preprocessing Steps

- **Step 1:** Importing the libraries
- **Step 2:** Loading the Data set
- **Step 3:** Identify independent and dependent feature
- **Step 4:** Handling of Missing Data
- **Step 5:** Handling of Categorical Data
- **Step 6:** Feature Scaling
- **Step 7:** Splitting the data set into training and testing datasets

Step 1: Import Libraries

First step is usually importing the libraries that will be needed in the program. A library is essentially a collection of modules that can be called and used. Built-in functions are defined in libraries which can be used by the programmer. For example, importing the library pandas and assigning alias as pd.

```
import pandas as pd
```

Step 2: Loading the data set

Load the dataset into pandas data frame using read_csv() function. The read_csv() function reads comma separated values (csv) dataset into pandas dataframe.

```
import pandas as pd

dataset = pd.read_csv('Data_for_preprocessing.csv')
```

Step 3: Identify Independent and Dependent Variables

- The next step of data pre-processing is to identify independent and dependent variables from the data set.
- All the features of any data set are not important for Machine Learning algorithm.
- Classification of dependent and independent feature is very important in Machine Learning.

Independent Variables

- Independent variables (also referred to as Features) are the input for a process that is being analyzes.
- Usually independent features/variables are also known as input features/variables and represented as X.

A	B	C	D
Country	**Age**	**Salary**	**Purchased**
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	NaN	NaN	No
Germany	40	1000	Yes
France	35	58000	Yes
Spain	78	52000	No
France	NaN	79000	Yes
Germany	50	83000	No
France	37	NaN	Yes

Independent Feature

- For example, in Data_for_preprocessing data set, the features such as Country, Age and Salary is known as independent features because they are not dependent to Purchased feature.
- They must be extracted before starting Machine Learning process.
- They can be extracted from the dataset as follows:

```
X=dataset.drop(['Purchased'], axis=1)
```

Dropping the 'Purchased' feature from the data set and initializing the remaining features to X. Here, axis=1 means dropping the column named 'Purchased' from the dataset.

Dependent Variable

- Dependent variables/features are the output of the process.
- Dependent features/variables are also known as output feature/variable and represented as y.

	A	B	C	D
1	**Country**	**Age**	**Salary**	**Purchased**
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	NaN	NaN	No
6	Germany	40	1000	Yes
7	France	35	58000	Yes
8	Spain	78	52000	No
9	France	NaN	79000	Yes
10	Germany	50	83000	No
11	France	37	NaN	Yes

Dependent/Output Variable/Feature

- The result (whether a user purchased or not) is the dependent variable.
- It must be extracted before starting Machine Learning Process.
- It can be extracted as follows:

```
y=dataset['Purchased']
```

Now, the 'Purchased' column of dataset will be assigned to y.

Next

PREVIOUS ACTIVITY

◄ Data File: Admission_Predict.csv: Day 10

Jump to...

NEXT ACTIVITY

Python code: Regression Problem:Apply Different Algorithm: Day 11 ►

Stay in touch

Contact Us

http://nielit.gov.in/gorakhpur/
abhinav@nielit.gov.in or ajay.verma@nielit.gov.in