Data Preprocessing and Data Visualization

CSE-454: Data Warehousing and Data Mining Sessional

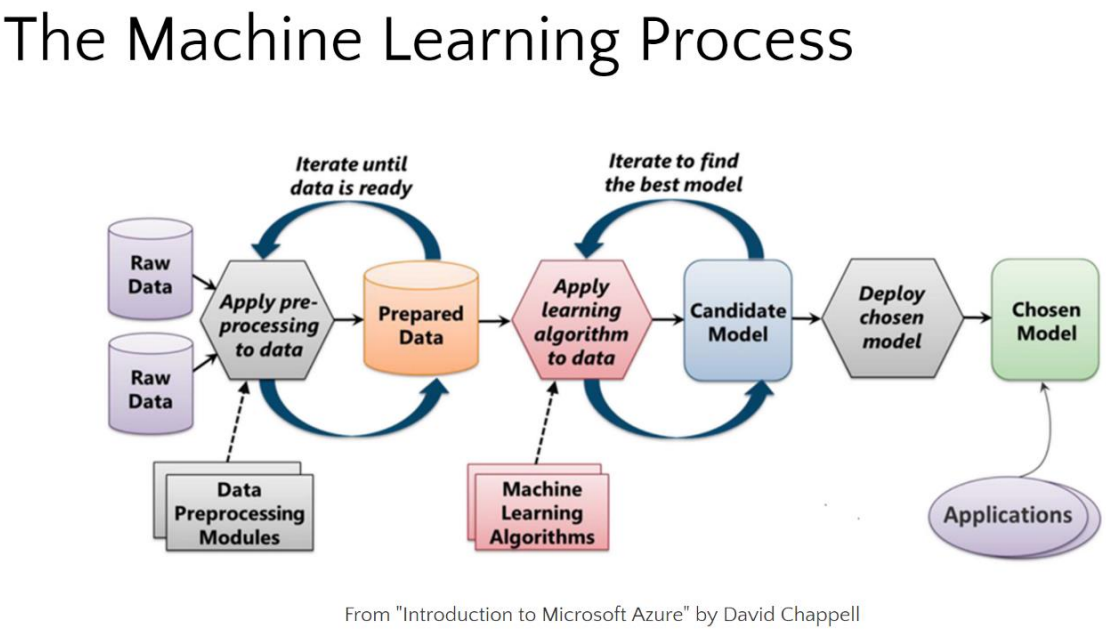
MD. JAKARIA

LECTURER

DEPT. OF CSE, MIST

Data Preprocessing

Definition: Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Data Preprocessing

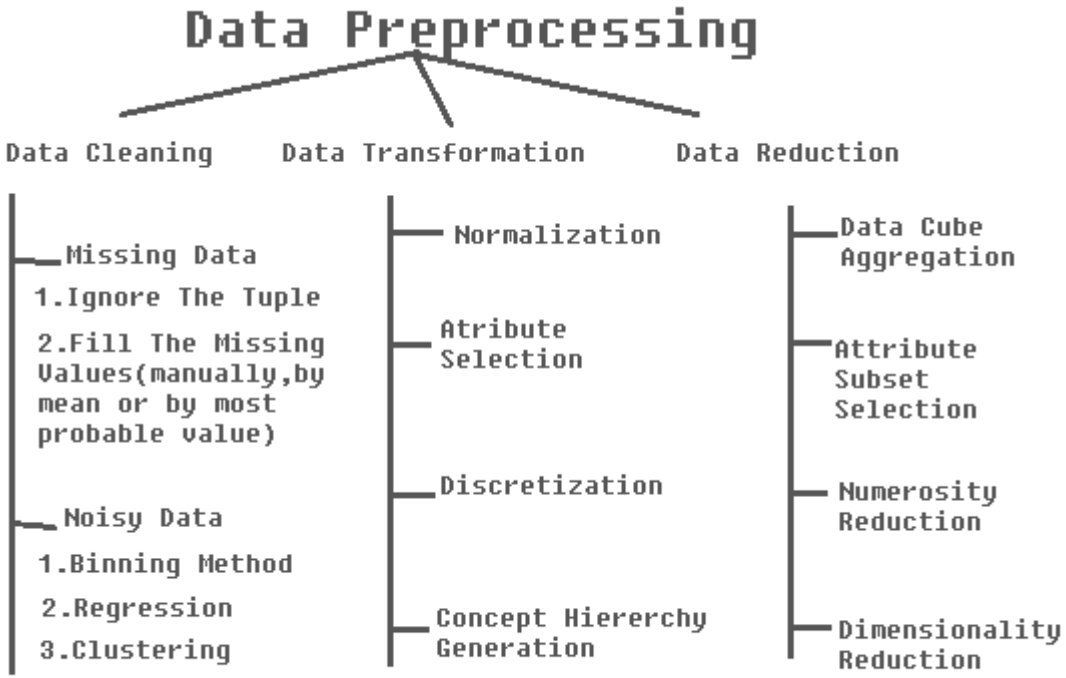
Why we use Data Preprocessing ?

In Real world data are generally-

i. **Incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.
ii. **Noisy:** containing errors or outliers.
iii. **Inconsistent:** containing discrepancies in codes or names.

Data preprocessing is a proven method of resolving such issues.

Data Preprocessing



Steps Involved in Data Preprocessing

Data Preprocessing

Steps Involved in Data Preprocessing

1. Data Cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a) **Missing Data**

This situation arises when some data is missing in the data. It can be handled in various ways.

i. Ignore the tuple

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple

ii. Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

Data Preprocessing

(b) **Noisy Data**

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc.

i. Binning Method

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

ii. Regression

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

iii. Clustering

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Data Preprocessing

2. Data Transformation

This step is taken in order to transform the data in appropriate forms suitable for mining process.

(a) Normalization

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

(b) Attribute Selection

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

(c) Discretization

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

(d) Concept Hierarchy Generation

Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Preprocessing

3. Data Reduction

While working with huge volume of data, analysis became harder. To get rid of this, we uses data reduction technique, which aims to increase the storage efficiency and reduce data storage and analysis costs.

(a) Data Cube Aggregation

Aggregation operation is applied to data for the construction of the data cube.

(b) Attribute Subset Selection

The highly relevant attributes should be used, rest all can be discarded.

(c) Numerosity Reduction

This enable to store the model of data instead of whole data, for example: Regression Models.

(d) Dimensionality Reduction

This reduce the size of data by encoding mechanisms. The two effective methods are: Wavelet transforms and PCA

Data Preprocessing

Features in Machine Learning

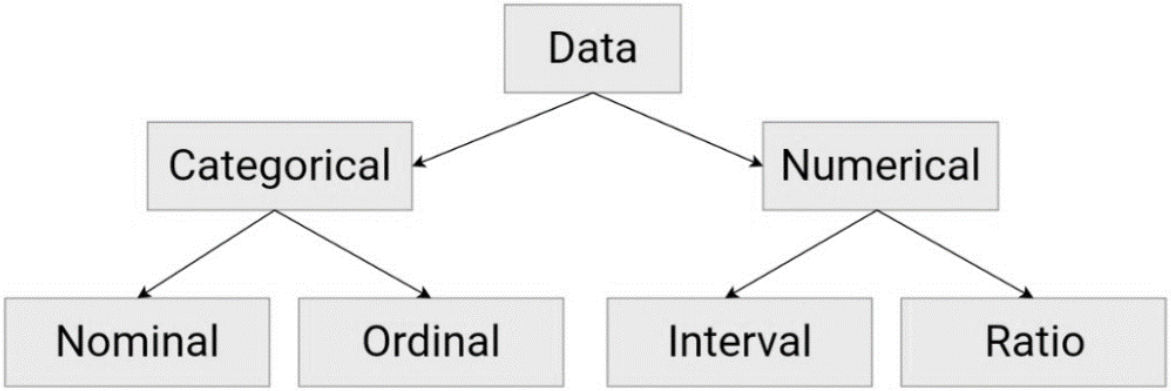
A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, **samples**, observations, or entities.

Data objects are described by a number of features, that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc.. Features are often called as variables, characteristics, fields, **attributes**, or dimensions.

"A feature is an individual measurable property or characteristic of a phenomenon being observed"

Data Preprocessing

Statistical Data Types



- **Categorical:** Features whose values are taken from a defined set of values. Ex., days in a week : {Mon, Tue, Wed, Thu, Fri, Sat, Sun}. Another example could be the Boolean set : {True, False}
- **Numerical:** Features whose values are continuous or integer-valued. They are represented by numbers and possess most of the properties of numbers. For instance, number of steps you walk in a day, or the speed at which you are driving your car at.

Data Preprocessing

Statistical Data Types

Nominal	Ordinal	Interval	Ratio
Categorical variables without any implied order	Categorical variables with a natural implied order but the scale of difference is not defined	Numeric variabes with a defnied unit of measurement, so the differences between values are meaningful	Numeric variables with a defined unit of measurement but both differences and ratios are meaningful
Example : A new car model comes in these colors : Black, Blue, White, Silver	Example : Sizes of clothes has a natural order : Extra Small < Small < Medium < Large < Extra Large - But this does not mean Large - Medium = Medium - Small	Examples : Calender Dates, Temperature in Celsius or Farhenheit	Examples : Temperature in Kelvin, Monetary quantities, Counts, Age, Mass, Length, Electrical Current

Data Preprocessing

Data Splitting

After necessary preprocessing, our dataset is ready for the exciting machine learning algorithms! But before we start deciding the algorithm which should be used, it is always advised to split the dataset into 2 or sometimes 3 parts.

1. **Training data:** This is the part on which your machine learning algorithms are actually trained to build a model. The model tries to learn the dataset and its various characteristics and intricacies, which also raises the issue of **Overfitting v/s Underfitting**.
2. **Validation data:** This is the part of the dataset which is used to validate our various model fits. In simpler words, we use validation data to choose and improve our model hyperparameters. The model does not learn the validation set but uses it to get to a better state of hyperparameters.
3. **Test data:** This part of the dataset is used to test our model hypothesis. It is left untouched and unseen until the model and hyperparameters are decided, and only after that the model is applied on the test data to get an accurate measure of how it would perform when deployed on real-world data.

Data Preprocessing

Data Splitting

After necessary preprocessing, our dataset is ready for the exciting machine learning algorithms! But before we start deciding the algorithm which should be used, it is always advised to split the dataset into 2 or sometimes 3 parts.

Data Split into parts