Machine Learning

Lecture 6 - 7: Data Preprocessing

COURSE CODE: CSE451

2021

Course Teacher

Dr. Mrinal Kanti Baowaly

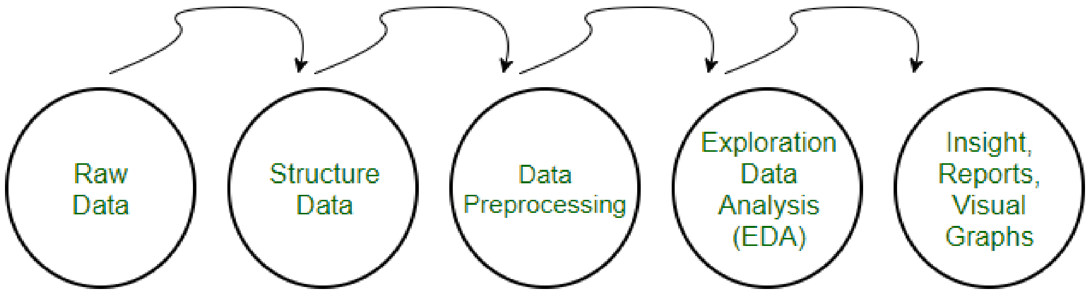
Assistant Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: baowaly@gmail.com

What is Data Preprocessing?

- Data preprocessing is a number of techniques that are used to transform the raw data in a useful and efficient format before feeding it to the algorithm
- Data Preprocessing is the most important step in machine learning to ensure the quality of data
- It directly affects the ability of our model to learn



Source: GeeksforGeeks

Data Preprocessing Techniques

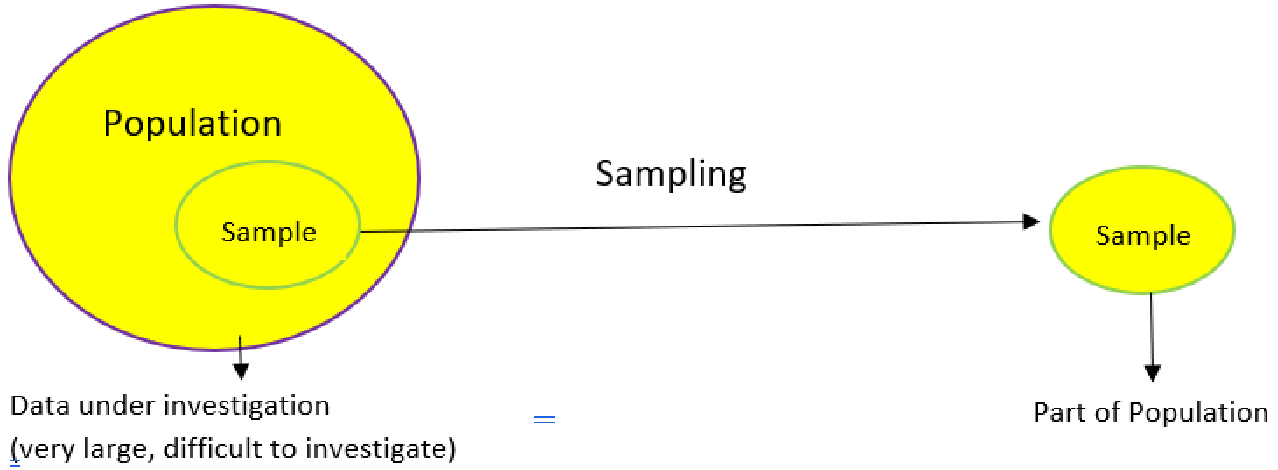
- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
  - Less memory, less processing time
- Disadvantage: the potential loss of interesting details

Sampling

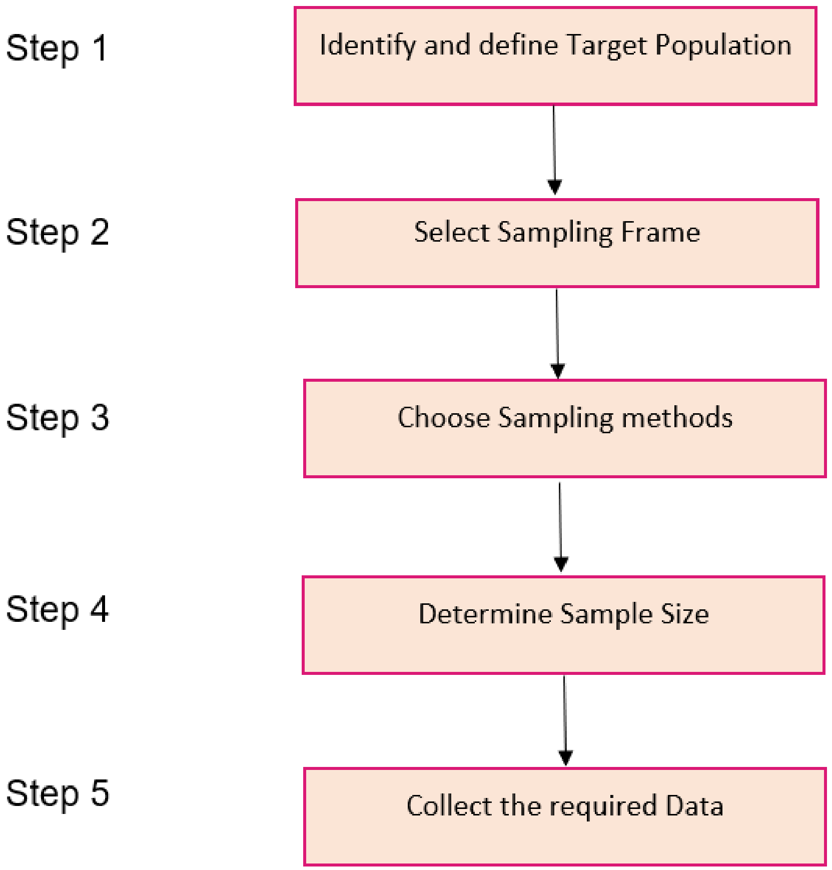
- Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.



What is Representative Sample?

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data sets (or population), if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

Steps Involved in Sampling



Detail: AanalyticsVidhya

Types of Sampling Techniques

- **Simple Random Sampling:** There is an equal probability of selecting any particular item
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample
    - In sampling with replacement, the same object can be picked up more than once

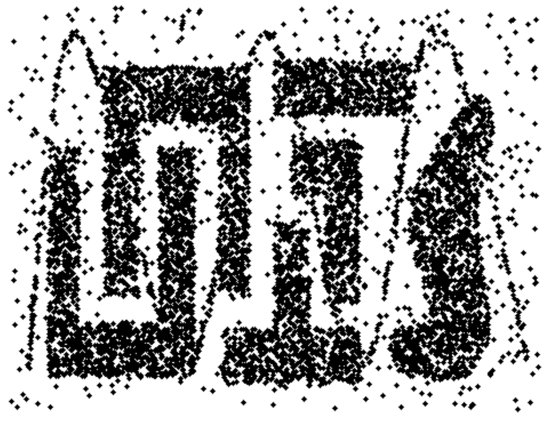
- **Systematic Sampling:** Samples are drawn using a pre-specified pattern, such as at intervals.

Types of Sampling Techniques (Cont.)

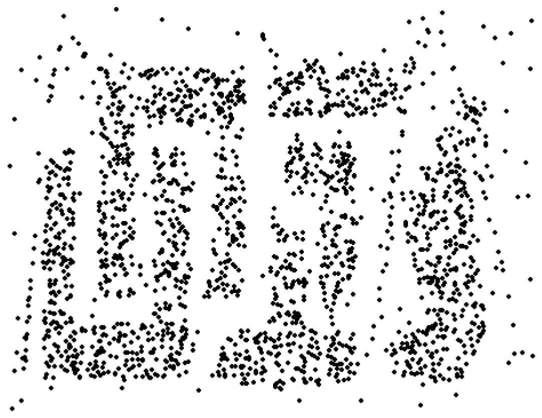
- **Stratified Sampling:** Split the data into several partitions called strata based on different traits like gender, category, etc.; then draw random samples from each partition.
- **Cluster Sampling:** The population is divided into some groups called clusters. Then we select a fixed number of clusters randomly and include all observations from each of the clusters in our sample.
- **Multistage sampling:** It is very much similar to cluster sample but instead of keeping all the observations in each cluster, we collect a random sample within each selected cluster.

Detail: AanalyticsVidhya, Kaggle

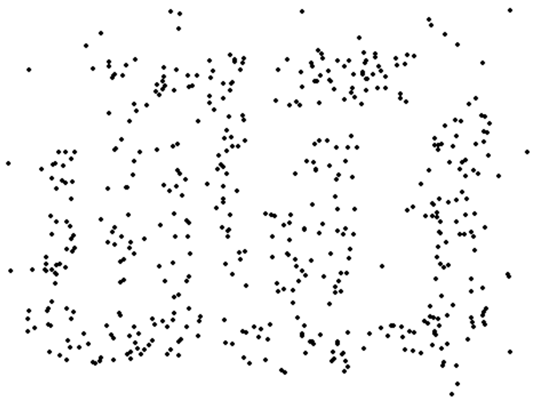
Determine the Proper Sample Size



8000 points



2000 Points



500 Points

Example of the loss of structure with sampling

- **Progressive sampling:** Start with a small sample, and then increase the size until a sufficient sample has been obtained

Curse of Dimensionality

- Many types of data analysis become harder as the dimensionality increases, the data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

Dimensionality Reduction

Purpose:

- Avoid curse of dimensionality
- May help to eliminate irrelevant features or reduce noise
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- Allow model to be more understandable

Techniques:

- Principle Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Others: supervised and non-linear techniques

Dimensionality Reduction vs Feature Subset Selection

Dimensionality Reduction

- Techniques that reduce the dimensionality of a data set by creating new attributes that are a combination of the old attributes

Feature (Subset) Selection

- Techniques that reduce the dimensionality of a data set by selecting only a subset of the attributes

Feature Subset Selection

- Alternative way to reduce dimensionality of data.
- It is desirable to reduce the number of **redundant** and **irrelevant** input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Redundant features add no relevant information to your other features, because they are correlated or because they can be obtained by [linear] combination of other features.
  - Example: date of birth of a student and his age, age can be obtained from date of birth

Feature Subset Selection (Cont.)

- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students’ ID is often irrelevant to the task of predicting their GPA

Feature Subset Selection Techniques

- Brute-force approach:
  - Try all possible feature subsets as input to data mining algorithm
- Embedded approaches:
  - Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches:
  - Features are selected before data mining algorithm is run, using some independent approaches (statistical measures), e.g. Pearson’s Correlation, LDA, ANOVA, Chi-Square etc.
- Wrapper approaches:
  - Use a data mining algorithm as a black box to find best subset of attributes, typically without enumerating all possible subsets

Feature Subset Selection Techniques(Cont.)

- Feature weighting
  - More important weights are assigned a higher weight, while less important features are given a lower weight
  - Some machine learning algorithms (e.g. SVM, GBM) do it automatically during data mining
  - Features with larger weights can be selected

Detail: AnalyticsVidhya, MachineLearningMastery

Feature Creation

Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

Three general methodologies:

- Feature extraction
  - Example: extracting edges from images
- Feature construction
  - Example: dividing mass by volume to get density
- Mapping data to new space
  - Example: Fourier and wavelet analysis

Discretization

Discretization is the process of converting a continuous attribute into an ordinal attribute

- A potentially infinite number of values are mapped into a small number of categories
- Discretization is commonly used in classification
- Many classification algorithms work best if both the independent and dependent variables have only a few values

How can we tell what the best discretization is?

- Unsupervised discretization: find breaks in the data values without using the class label information
    - Common approaches: Equal width, Equal frequency, K-means clustering

- Supervised discretization: Use class label information to find breaks i.e. supervised discretization filter uses the number of classes as the discretization parameter

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}
  - **Common approaches:** Assigning unique integer values [0, m-1] then convert to binary, One-hot encoding

Binarization Example

Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Variable/Attribute Transformation

An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

- Simple functions: x^k, log(x), e^x, |x|, sqrt, sin x, 1/x etc.
- **Purpose**: sqrt, log and 1/x are often used to transform data to Gaussian (normal) distribution, minimizing the huge range of values

Math: Normal distribution, Standard Deviation

Normalization

Normalization scales all numeric variables in the range [0,1]

- Refers to various techniques to adjust the differences among attributes in terms of frequency of occurrence, mean, variance, range
- Before normalization, it is recommended to handle the outliers

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Normalization Example

```
# Normalize the data attributes for the Iris dataset.
from sklearn.datasets import load_iris
from sklearn import preprocessing
# load the iris dataset
iris = load_iris()
print(iris.data.shape)
# separate the data from the target attributes
X = iris.data
y = iris.target
# normalize the data attributes
normalized_X = preprocessing.normalize(X)
```

Standardization

Data standardization is the process of rescaling one or more variables so that they have a mean value of 0 and a standard deviation of 1

- Refers to subtracting off the means and dividing by the standard deviation
- Useful when min and max are unknown or when there are outliers

$$x_{new} = \frac{x - \mu}{\sigma}$$

Standardization Example

```
# Standardize the data attributes for the Iris dataset.
from sklearn.datasets import load_iris
from sklearn import preprocessing
# load the Iris dataset
iris = load_iris()
print(iris.data.shape)
# separate the data and target attributes
X = iris.data
y = iris.target
# standardize the data attributes
standardized_X = preprocessing.scale(X)
```

Lecture 7: Exploratory Data Analysis (EDA)

- An approach to analyze and investigate data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task.
- It can help identify obvious errors, as well as better understand the patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

EDA Lab Works

1. Ultimate guide for Data Exploration in Python using NumPy, Matplotlib and Pandas, by AnalyticsVidhya
2. Introduction to Exploratory Data Analysis (EDA), by AnalyticsVidhya
3. Comprehensive Data Exploration with Python, by Kaggle
4. CheatSheet: Data Exploration using Pandas in Python, by AnalyticsVidhya
5. Python Exploratory Data Analysis Tutorial, by Datacamp
6. Statistical Learning Tutorial for Beginners, by Kaggle