Data Science Concepts

Lesson01C–Data Pre processing

Objective

After completing this lesson you will be able to:

- Describe the importance of data pre-processing and its impact on the analysis
- Understand the various techniques of data pre-processing

Why Data Preprocessing?

Data in the real world is dirty

- incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“”
- noisy: containing errors or outliers
  - e.g., Salary=“-10”
- inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

No quality data, no quality results!

- Quality decisions must be based on quality data
- e.g., duplicate or missing data may cause incorrect or even misleading statistics.

Data preparation, cleaning, and transformation comprises the majority of the work in a data analytics project (~60%).

Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, or files

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies

- Data transformation
  - Normalization

Data Cleaning

Data cleaning tasks

- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data

Data Cleaning–Missing Data

Data is not always available

- E.g., many tuples have no recorded values for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not registered history or changes of the data

Missing data may need to be inferred.

Data Cleaning–Missing Value Imputation

There are a variety of techniques for missing value imputation; but these should be considered more as scenario-specific than just being a set of pure alternative choices.

There are several missing value imputation techniques:

- Impute Missing Values with ZERO
- Impute Missing Values with MEDIAN
- Impute Missing Values with MEAN
- Impute Missing Values with MODE
- Information based Segmentation
- Impute using Regression on other Non-Missing Predictors
- Logical imputation

Data Cleaning–Noisy Data

Noise: random error or variance in a measured variable

Incorrect attribute values may due to

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which requires data cleaning

- duplicate records
- incomplete data
- inconsistent data

Data Cleaning–Handling Noisy Data

- Combined computer and human inspection
  - detect suspicious values and check by human

- Regression
  - smooth by fitting the data into regression functions

- Clustering
  - detect and remove outliers

Summary

Summary of the topics covered in this lesson:

- Data preparation is a time taking activity and majority of the time in an analytics projects is typically spent in this phase.
- There are several techniques available to improve the quality of the data i.e. data completeness and data consistency.

QUIZ TIME

Quiz Question 1

Quiz 1	What are the typical reasons for missing data?

a. Data not entered due to misunderstanding.

b. Certain data may not be considered important at the time of entry.

c. Inconsistent with other recorded data and thus deleted.

d. All the above.

Quiz Question 1

Quiz 1 What are the typical reasons for missing data?

a. Data not entered due to misunderstanding.

b. Certain data may not be considered important at the time of entry.

c. Inconsistent with other recorded data and thus deleted.

d. All the above.

Correct answer is:

d

There can be many more reasons for missing data but all the above factors into those reasons as well.

End of Lesson01C–Data Pre processing