Dataset

By: Muhammad Qamar Iqbal

Data set

- A **data set** is a collection of numbers or values that relate to a particular subject or task.
  - The test scores of each student in a particular class
  - The transactions at a store

Data Objects

- Data objects can also be referred to as
  - *Samples*
  - *Examples*
  - *Instances*
  - *Data points*
  - *Objects*
  - *Tuples*

Data Objects

- **Data Object**
  - Represents an entity

- Sales database
  - Customers
  - Store items
  - Sales

- Medical database
  - Patients
  - Doctors
  - …

- University database
  - Students
  - Professors
  - Courses

Attribute

- An attribute is a data field, *representing* a **characteristic** or **feature** of a data object

- A.k.a.
  - *Dimension* *(Data Warehouse)*
  - *Feature* *(Machine Learning)*
  - *Variable* *(Statistics)*

Attribute

- Customer (object)
  - *Customer ID*
  - *Name*
  - *Address*
- *Student (object)*
  - *???*
  - *???*
  - …

Types of Attributes

- Qualitative
  - Nominal
  - Ordinal
- Quantitative
  - Numeric
    - Interval
    - Ratio

Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values.
Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value.

Nominal Attributes

- Values of a nominal attribute are
  - **Symbols,** or
  - ***names of things***
  - e.g. category, code, or state

- *Category*
  - Undergraduate, Graduate
- *Code*
  - 065, 266, 105
- *State*
  - Present, Absent

Ordinal Attributes

- An attribute with possible values that have a meaningful **order** or **ranking** among them
- But the magnitude between successive values is not known
- For example, A Pizza ☺
  - *small, medium, large*

Numeric Attributes

- A numeric attribute is **quantitative**
  - i.e. It is a *measurable* quantity
  - Represented in
    - ***integer*** or ***real*** values
- Numeric attributes can be
  - Interval-scaled
  - Ratio-scaled

Standard Public Datasets

- Standard Public Datasets available online
  - Kaggle
  - UCI Machine Learning Repository
  - Open Data Pakistan
  - Google Trends
  - …
- Explore more

Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

Central Tendency of Data

- Mean
- Median
- Mode

Five-number Summary

- Minimum Value
- Q1
- Median
- Q3
- Maximum Value

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Imbalanced datasets

- Oversampling: It involves replicating samples from the minority class to increase its representation. Techniques like random oversampling, synthetic minority oversampling technique (SMOTE), and adaptive synthetic sampling (ADASYN) are commonly used.
- Undersampling: It involves randomly removing samples from the majority class to reduce its representation. Random undersampling and cluster-based undersampling are examples of undersampling techniques.
- Data augmentation: This technique involves creating new synthetic samples by applying various transformations to the existing samples in the minority class. This can help increase the diversity and quantity of the minority class data.

Thank you

Muhamad Qamar Iqbal
https://www.linkedin.com/in/muhammad-qamar-iqbal-3676509a/