DATA ANALYTICS

Unit 1:Getting to know your data

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Getting to know your data

Mamatha H R

Department of Computer Science and Engineering

Getting to know your data

- Knowledge about your data is useful for data pre-processing ,one of the major tasks of the data analysis process.

Getting to know your data

- You will want to know the following:
- What are the types of *attributes* or fields that make up your data?
- What kind of values does each attribute have?
- Which attributes are discrete, and which are continuous-valued?
- What do the data *look like*?

Getting to know your data

- How are the values distributed?
- Are there ways we can visualize the data to get a better sense of it all?
- Can we spot any outliers?
- Can we measure the similarity of some data objects with respect to others?
- Gaining such insight into the data will help with the subsequent analysis.

Getting to know your data

- “So what can we learn about our data that’s helpful in data pre-processing?”
  - studying the various attribute types.
  - These include nominal attributes, binary attributes, ordinal attributes, and numeric attributes.

Getting to know your data

- Basic statistical descriptions can be used to learn more about each attribute’s values
- Given a temperature attribute, for example, we can determine its mean (average value),median (middle value), and mode (most common value).
- These are measures of central tendency, which give us an idea of the “middle” or center of distribution.
- Knowing such basic statistics regarding each attribute makes it easier to fill in missing values, smooth noisy values, and spot outliers during data preprocessing.
- Knowledge of the attributes and attribute values can also help in fixing inconsistencies incurred during data integration.

Getting to know your data

- Plotting the measures of central tendency shows us if the data are symmetric or skewed.
- Quantile plots, histograms, and scatter plots are other graphic displays of basic statistical descriptions.
- These can all be useful during data preprocessing and can provide insight into areas for analytics.

Getting to know your data

- The field of data visualization provides many additional techniques for viewing data through graphical means.
- These can help identify relations, trends, and biases “hidden” in unstructured data sets.

Getting to know your data

- we may want to examine how similar (or dissimilar) data objects are.
- For example, suppose we have a database where the data objects are patients, described by their symptoms. We may want to find the similarity or dissimilarity between individual patients. Such information can allow us to find clusters of like patients within the data set.
- The similarity/dissimilarity between objects may also be used to detect outliers in the data, or to perform nearest-neighbor classification.

References

Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining by Tan, Steinbach, Kumar, 2nd Edition

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834