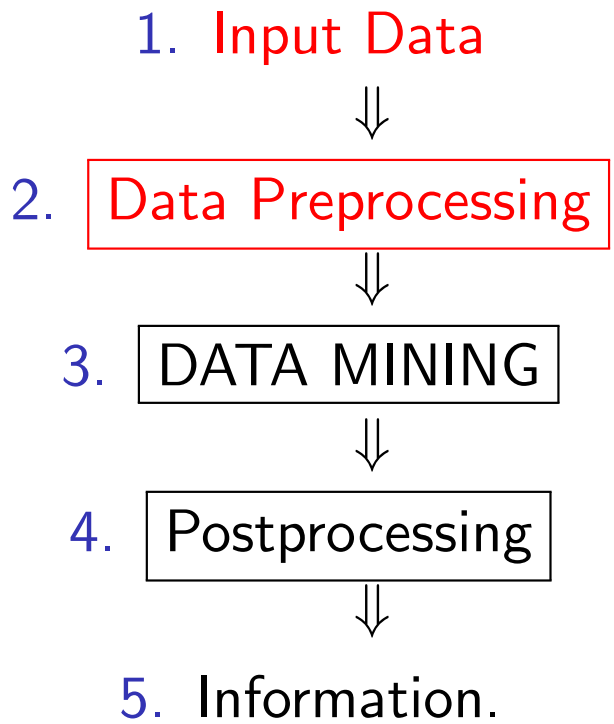
Data

3rd Aug, 2016

Recap

- Looked at a few applications of Data Mining.
- Motivation for Data *Mining*.
- Introduced the concept and importance of Data.

Data Mining and Knowledge Discovery



Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

[1] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

- KNOW YOUR DATA.
- Procuring Data.
- Quality of Data.
- Preprocessing Steps.
- Similarity/Dissimilarity measures.

[1] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Data

Question: Take out a few examples from the discussion till now and tell me how you represent the data.

- KNOW YOUR DATA.
- Procuring Data.
- Quality of Data.
- Preprocessing Steps.
- Similarity/Dissimilarity measures.
- See Example 2.1[1]

[1] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

What makes data

- A data set is a collection of data objects.
  - Record
  - Point
  - Sample
  - Vector
- Data objects are described by a number of *attributes*.
- An attribute is a property of an object.
  - To define and then measure a characteristic of the data object.

Attributes

- A property or characteristic of an object. You have to decide what makes an attribute.
- Needs a measurement scale.
- Character of attribute: Age and ID are both integers, but with different characteristics.
- Operations on numbers can be used for values of attributes
  - $=$ and $\neq$
  - $<, \leq, >, \geq$
  - $+, -, \times, /$
- Attribute types
  - Nominal: ZIP codes, ID. ($=, \neq$)
  - Ordinal: $\{Good, better, best\}$. ($=, \neq$, $<, \leq, >, \geq$)
  - Interval: calender dates. ($=, \neq, <, \leq, >, \geq$, $+, -$)
  - Ratio: monetary quantities, counts, mass, length. ($=, \neq, <, \leq, >, \geq, +, -$, $\times, /$)

Attributes

- Categorical vs Numeric.
- Discrete vs Continuous.
- Special consideration:
  - Asymmetric attributes (zero or nonzero).
  - Binary.

Datasets

Characteristics

- Dimensionality.
  - No. of attributes.
  - Curse of Dimensionality.
  - Dimensionality reduction.
- Sparsity
  - If most of the attribute values are zeros.
  - Advantage or Disadvantage?
- Resolution.
  - Number of students in an institute, in a dept, in a course, yearwise, batch,...
  - Girls and boys distribution.

Datasets

Characteristics

- Dimensionality.
  - No. of attributes.
  - Curse of Dimensionality.
  - Dimensionality reduction.
- Sparsity
  - If most of the attribute values are zeros.
  - Advantage or Disadvantage?
- Resolution.
  - Number of students in an institute, in a dept, in a course, yearwise, batch,...
  - Girls and boys distribution.

Datasets

Types and representation

- Record data.
  - Relational Database.
- Transaction or Market Basket Data.
  - TID and list of items
- Data Matrix.
  - Think of a real matrix.
- Document-term matrix (sparse matrix).
  - Each doc represented as a vector corresponding the vocabulary.

Other datasets and types

- Graph-based
  - Graph capturing relationships among data objects.
    - WWW
    - Links between page objects.
  - Data objects themselves are represented as graphs.
    - Structure of chemical compounds.
- Ordered data.
  - Sequential data.
  - Sequence data.
  - Time series data.
  - Spatial data
  - Temporal and spatial autocorrelation.

Data Quality

Some Issues

- Many a times, data is collected without an application in mind.
- Applications are developed on available data.
- Dealing with data quality issues.
  1. Data cleaning.
  2. Make algorithms tolerant to poor data quality.

Measurement and data collection issues

- Errors, Errors, and more errors...

Measurement and data collection issues

- Errors, Errors, and more errors...
- Human errors
- Limitations of measuring device.
- Procedural errors.

Measurement and data collection issues

- Errors, Errors, and more errors...
- Human errors
- Limitations of measuring device.
- Procedural errors.
- Types of errors.
  - Changes from the true value.
  - Missing values.
  - Even typos.
- NOISE

Evaluating Measurements.

- How do you evaluate the accuracy of weighing machine?
  - Test it with some standards.
- Precision: variation of repeated measurement.
- Bias: variation of measurement from the (correct) quantity being measured
- And finally the accuracy: closeness to the true value.

Other issues

- Outliers
  - Within object.
  - Within attribute.
- Missing values.
- Handling Missing values.
  - Eliminate them
    - Objects.
    - Attributes.
  - Estimate the missing values.
  - Ignore them.
- Inconsistent Values.

Data Preprocessing

Assuming a clean (or unclean) data, we further do the following for making the DM tasks easier (or rather not complicated)

- Aggregation.
- Sampling.
- Dimensionality reduction (Remember the curse).
- Feature subset selection.
- Feature creation
- Discretization and binarization.
- Variable Transformation.

Preprocessing

Aggregation

Less is more sometimes!

- Our purpose is data reduction.
- Reduce the number of objects or attributes by combining them.
- Represent the quantities as a sum or an average across objects.
- Collect items to form a set of all items sold.

Why aggregation

- Less memory and processing time.
- Implies permissibility to use more expensive algos.
- A high-level view against a low-level view
  - But this could be disadvantage as we loose interesting details.
- More stability and less variability.

Preprocessing

Sampling

Handling 10000000 points vs handling 1000 representative samples.

- Selecting a subset of data objects for analysis, instead of the entire population.
- For statisticians, obtaining the entire dataset is expensive, even though desirable.
- For data miners, processing all data is expensive.

Representative samples

- Sample should work as effectively as using the entire data set.
- Characteristics of data shouldn't be lost.
- Mean, variance, covariance of the sample and the population should ideally be close.

Representative samples

- Sample should work as effectively as using the entire data set.
- Characteristics of data shouldn't be lost.
- Mean, variance, covariance of the sample and the population should ideally be close.
- How do we get such samples?
  - Sampling techniques.
  - Sample size.

Sampling Approaches

- Simple random sampling
  - Sampling without replacement
  - Sampling with replacement (Gives independent and identically distributed (i.i.d))
- Stratified sampling
  - Giving importance to objects in different classes, either equally or proportionally.

Sample Size

- Sampling should not lead to loss of information: See Figure 2.9[2]
- What is the proper sample size?
- Progressive sampling
  - Try increasing sample size, until..

[2] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Sample Size

- Sampling should not lead to loss of information: See Figure 2.9[2]
- What is the proper sample size?
- Progressive sampling
  - Try increasing sample size, until..
  - accuracy levels off (leveling-off point)

[2] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley

Preprocessing

Dimensionality Reduction

- Scenario is when you have lots of attributes, but the values could be sparse.
- Dimensionality Reduction can eliminate irrelevant features.
- More *handy* representation.

Preprocessing

Dimensionality Reduction

- Scenario is when you have lots of attributes, but the values could be sparse.
- Dimensionality Reduction can eliminate irrelevant features.
- More *handy* representation.
- New attributes are created as a combination of old attributes.
- Principal Component Analysis (PCA).
- Singular Value Decomposition (SVD).

Preprocessing

Feature Subset Selection

- Eliminate unwanted features.
  - Redundant features.
  - Irrelevant features.
- The problem is to find the best subset of features.
- Best, based on performance compared to the entire set of features.
- Brute force: Try out 2^n subsets of features (n is the total no. of features).

Finding the best subset of features

- Embedded approaches
  - Feature selection is embedded in the data mining algorithm.
  - Decision Trees (Wait for it!)
- Filter approaches
  - Features selected before the DM algorithm is run.
  - Hence independent of DM algorithm.
  - For instance, select attributes whose pairwise correlation is low.
- Wrapper approaches.
  - DM algorithm is a black box to find the best subset.
  - Certain procedure for selecting subsets without enumerating all of them.
- Feature Weighing.

Preprocessing

Feature Creation

- Create new features, with or without retaining the original set of features.
- Feature extraction.
  - Domain specific.
  - Human face detection: Look for specific lines and edges and shades.
- Map data to a new space
  - Fourier Transformation.
  - Ask Image Processing guys.
- Feature Construction
  - $density = mass/volume$

Winding up

- Binarization
- Descritization
- Variable Transformation
  - using functions, say log, order reversal
- Normalization or Standardization.