Chapter 1: Data Mining– an Overview

Nischal Regmi

Everest Engineering College

2022

Defining Data Mining

- Data mining is the extraction of knowledge from large amounts of data.
- Some argue that the term ‘knowledge mining from data’ or ‘knowledge mining’ would be better. But the term ‘data mining’ is popular.
- For example, from a large data set that consists values of points $X = (x_1, x_2, \ldots, x_n)$, we may be interested in knowledge such as
    - Which variable is the cause and which is the effect? For example, x_1 could be number of cigarettes consumed per day, x_2 is a variable that measures the level of depression, and x_3 the person’s income. The question then is to identify if depression leads to smoking habit, low income leads to depression and so on. This problem is called *causal mining*.
    - How can the points be grouped according to their similarity? For example, x_1 could be household income, x_2 the highest level of education of the householders, and x_3 the size of household. The question is then to group the household according to their socioeconomic status. This problem is called *clustering*

Data Mining Task Primitives

- Before performing any sort of data mining analysis, the analyst needs to be precise about the following points.

1. The set of data that is relevant to the task in hand.
2. The kind of knowledge to be mined. In other words, the analyst should be clear about what type of information he or she is trying to seek from the data.
3. The background knowledge required for the discovery process. A knowledge of the problem domain is usefult in analyzing data and interpreting the results.
4. The *interestingness measures and threshold* to evaluate the utility of the pattern obtained from data mining.
5. The expected representation for visualizing the discovered patterns. That is, the analyst should be precise about how to present the final results.

Knowledge Discovery from Data

- Knowledge discovery from data (KDD) is the process of extracting knowledge from data.
- KDD is an iterative process with the following steps.
  1. Data cleaning – remove noise and inconsistent data.
  2. Data integration – combine (if needed) data from multiple sources
  3. Data selection – select relevant data for the problem in hand
  4. Data transformation – apply preliminary calculations to change the format of data as per the requirement of data mining algorithm
  5. Data mining – apply appropriate algorithm on the transform data to extract knowledge/pattern.
  6. Pattern evaluation – identify which pattern is useful for the problem in hand
  7. Knowledge presentation – present the final output using appropriate plots.
- Many people take data mining and KDD as synonymous, but from above, it is clear that data mining is only one step in KDD.

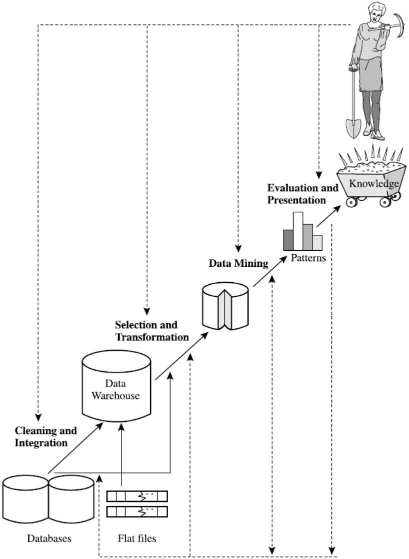


Figure: The KDD process. Data mining is one step in the process. (fig. source – Han and Kamber, 2006)

KDD Step 1: Data Cleaning & Integration – Missing Values

- Tabular data can have entries with missing values. The data analyst needs to be precise about how to handle the missing values.
- Following are some techniques used to handle missing data.
  - Ignore the row containing a missing data: This strategy could be useful only when the number of missing data is very less compared to the number of rows in the table.
  - Fill in the missing value manually: Filling the missing value manually by observing other entries is more reliable. However, it is much time consuming and is not feasible if the data is huge.
  - For numeric data, one insert the mean value for the missing value. But this creates a biased data set.
  - Use the most probable value to fill in the missing value: This method also biases the data.
  - For numeric and ordered data, one can use interpolation to fill the missing data.

KDD Step1: Data Cleaning – Noisy Data

- Noise is a random error or variance in a measured variable.
- For example, we have time-series data consisting of pairs (y, t), where t denote time and the actual relation between x and t should be of the form

  $$y = \sin t$$

  But because of unknown reasons, in our data set, the relation between y and t appears as

  $$y = \sin t + \epsilon$$

  where ϵ is a normally distributed term with mean 0 and very small standard deviation (recall normal distribution from statistics). Then we say the data is noisy.
- This situation is described in the following slide.

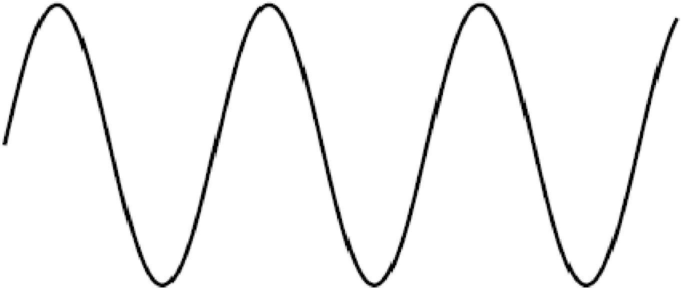


Figure: The actual relation ship between x and t. Vertical axis shows x and horizontal t

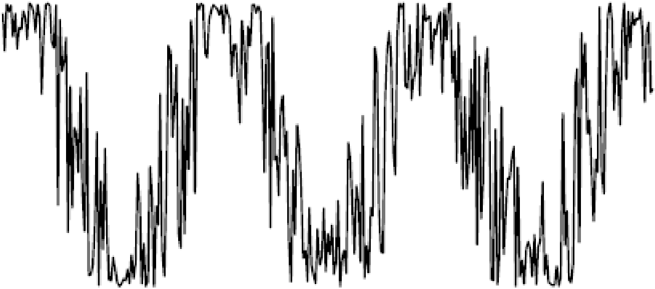
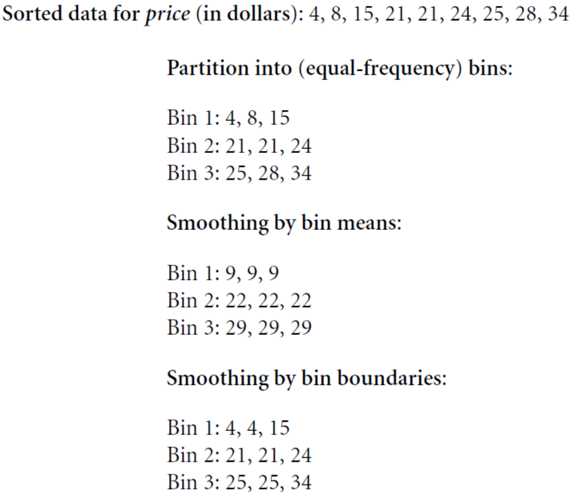


Figure: The figure obtained by plotting x and t in the data set. The plot is jig-jagged because of the presence of noise term ϵ

- Noise can be reduced using techniques that are collectively called as ‘smoothing’.
- There are many techniques for reducing noise. Examples are binning, regression, and clustering.

Noise Reduction using Binning

- Sort data according to their values in ascending order.
- Partition the sorted data into bins of equal size.
- Replace the data in each bin using one of the following smoothing techniques.
  - *Smoothing by bin means*: Replace all values in a bin by the mean value of the bin.
  - *Smoothing by bin medians*: Replace all values in a bin by the median value of the bin.
  - *Smoothing by bin boundaries*: Replace each value by the maximum or minimum value in the bin, whichever is closest.



Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure: Binning methods for data smoothing (fig. source – Han and Kamber, 2006)

Noise Reduction Using Regression

- The process of reducing noise in a multivariate numeric data using is as follows.
  1. Fit a regression model that consider one of the variables as a function of other variables.
  2. Replace the original values by the values predicted by the regression model.
- Consider a noisy data set comprising of points (x_i, y_i) where i denotes a particular data point. Then,
  1. First plot the values (x_i, y_i) and observe the pattern.
  2. The fit the regression model $y = f(x)$, where f depends on the pattern.

     2.1 For example if the pattern looks like a straight line, we take $f(x) = ax + b$, then the regression model will be $y = ax + b$, a linear regression model.

     2.2 If the pattern looks like a wave, we could take $f(x){=}a\sin(bx)$, then the regression model will be $y = a\sin(b)$, a non-linear regression model.
  3. Predict the y-values using actual x-values as $\hat{y}_i = f(x_i)$. Replace all (x_i, y_i) in the data set by $(x_i, \hat{y}_i)$

- Suppose we have following data

i	x_i	y_i
1	5	26.41681
2	10	37.55742
3	15	60.77099
4	20	69.99233
5	25	86.68350

- The plot of (x_i, y_i) suggests the relationship is linear (see next slide).
- After performing linear regression, we get the relationship

$$y = 3x + 10$$

- We then use the actual values of x_i in above equation to get predicted values of y, i.e. $\hat{y}_i$. Replacing each (x_i, y_i) by $(x_i, \hat{y}_i)$, we get the smoothed data set

$$\{(5, 25), (10, 40), (15, 55), (20, 70), (25, 85)\}$$

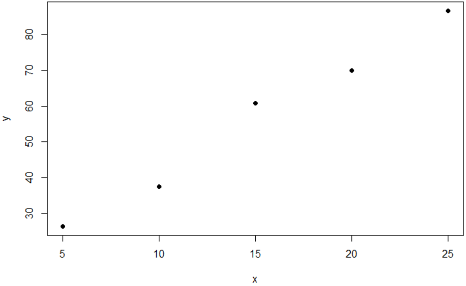


Figure: Plot of (x_i, y_i) suggests a linear relationship.

Noise Reduction Using Clustering

- Data points are clustered using appropriate algorithm.
- The points that do not belong precisely to one of the clusters are identified as 'outliers' and removed from the data set.
- The problem with this method is the results may significantly differ across the choice of clustering algorithm and preciseness measures.
- The method of outlier detection using clustering will be discussed later.

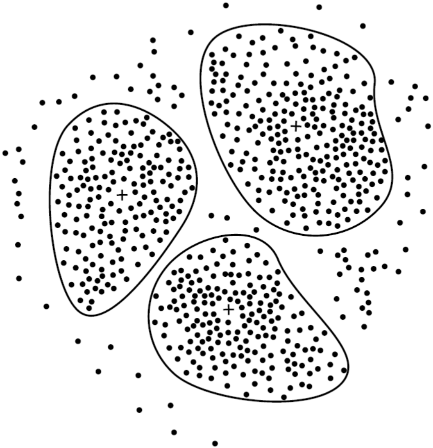


Figure: Outlier detection using clustering. The points that do not fall precisely in one of the clusters are considered outliers and removed. Fig. source – Han and Kamber, 2006

KDD Step 1: Data Integration

- Data integration combines data from multiple source that can be used for further analysis.
- There are some issues in data selection/integration that need to complicates the process
  1. Schema integration and object matching.
     - Schema of data tables may differ in different source
     - A same type of variable, for example, sales, may be stored using different names in different tables.
     - A same attribute name may refer to different variables in different tables.
- Redundancy
  - Sometimes, when we know the value of one variable, then we can predict the value another. This is called redundancy.
  - For example, given the which grade a student currently belongs to, we can guess his/her age.
  - Duplicate data may appear on the combined data set
- Data value conflicts. A same variable for some instance may have different values in different tables. For example, a company XYZ has total sales of Rs. 1 crore in one data table, and Rs. 1.2 crore in another.

KDD Step 2: Data Selection and Transformation

- Data selection is accessing relevant data from the warehouse.
- Data Transformation: Many algorithms require that data should be in a specific format. Data need to be transformed into forms appropriate for mining algorithms. Data transformation can involve
  1. Smoothing – mainly used to reduce noise. Some algorithms may require smoothed data.
  2. Aggregation – summary and aggregation operations applied to data. For instance, daily data can be aggregated to monthly data.
  3. Generalization of data - replace low level data by higher/general level concepts. Example, age can be replaced by youth, middle age, senior, etc.
  4. Normalization - numerical data attributes are scaled to specific range such as $[0, 1]$
  5. Attribute construction – new attributes are constructed and added from the given set of attributes.

- For example, clustering algorithms requires that all numbers are ‘unit-scaled’ (we will study this later)
  - Suppose we have data about individuals. For each individual we know their age and income. We need to group the individuals as their similarity in age and income.
  - Note that maximum realistic age is 100, whereas income takes much larger values, which are in thousands and even in lakhs for some.
  - If we apply clustering algorithm using this raw data, then the variable ‘age’ will have no significant role in defining the clusters.
  - We need to re-scale the variables in same range. For example we can divide all ages by the maximum age and all incomes by the maximum income. This will confine all ages and incomes in the range [0,1].

KDD Step 3: Data Mining

- Data mining is done on the transformed data, and is the third step of the KDD process.
- We will discuss data mining techniques later.

KDD Step 4: Evaluation and Presentation

- A data mining system can generate a very large number of patterns from data available in the warehouse. However, not all are interesting.
- A pattern is interesting if it is
  1. easily understood by data,
  2. valid on new or test data with some degree of certainty,
  3. potentially useful, and
  4. novel
- There are many objective measures of pattern interestingness. By objective, we mean that these measures assign numeric score to each pattern mined from data.
- The data mining system presents the patterns in the order of higher to lower interestingness scores.
- The interesting patterns are/can be presented in different forms like tables and various kinds of charts and plots.

Data Mining Techniques

- Data mining techniques can be classified in two ways, according to (1) the level of user interaction involved, and (2) the methods of data analysis employed.
- According to the level of user interaction, data mining systems can be
  1. *Autonomous Systems* – These systems do not require any user input
  2. *Interactive Exploratory Systems* – These systems allow users to manually define and select data analysis tasks
  3. *Query Driven Systems* – Allow the users to define their data analysis tasks through specific query languages like the Data Mining Query Language (DBML)
- Regarding the method of data analysis, data mining systems could rely on methods from database theory, machine learning, statistics, visualization, pattern recognition, neural networks and so on.
  - A good data mining system relies on more than one of these data mining methods.

Data mining Tools

- There are several data mining tools available, which some are commercial, whereas some are freely available.
- Some important tools:
  1. IBM SPSS Modeler – A commercial data science software by IBM
  2. Oracle Data Mining – a commercial product by the Oracle Corporation.
  3. Weka – A Java based free data analysis software which is highly popular in the academia.
  4. H_2O – An open source Python library for data mining