University of Stavanger, DAT550

Summary of lecture notes

Håvard Godal
Spring 2021

Contents

1 Introduction to Data Mining

1.1 Data Mining

Data mining can be described as:

"Non-trivial extraction of implicit,
previously unknown and potential useful information from data"

"Exploration and analysis, by automatic or semi-automatic means,
of large quantities of data in order to discover meaningful patterns"

1.2 Machine Learning

Machine learning can be described as:

"The field of study that gives computers the ability to learn
without being explicitly programmed"

"A computer program is said to learn from experience E
with respect to some class of tasks T and performance measure P,
if its performance at tasks in T, as measured by P, improves with experience E"

1.3 Data Mining or Machine Learning

Data mining is the process of discovering patterns in large data sets involving methods at the interseciton of machine learning, statistics, and database systems.

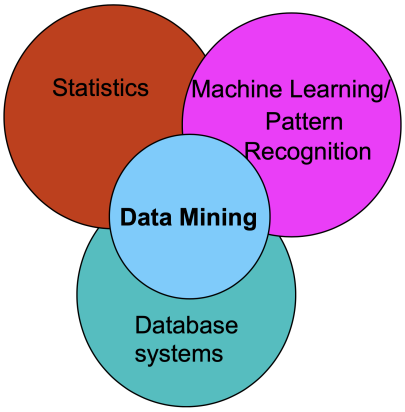

Figure 1.1: Diagram of Data Mining

1.4 Properties of Data Mining

1.4.1 Commercial viewpoint

Lots of data is being collected and warehoused, such as web data, purchases, and transactions. Computation power has become cheap and powerful, and the competitive pressure is strong. It is therefore necessary to provide better, customized servicees for an edge.

1.4.2 Scientific Viewpoint

Data is collected and stored at enormous speeds, up to multiple terrabytes per hour. Such data can be generated from the large hydron collider and from scanning the universe. Such data can not be interpreted by traditional techniques, and requires data mining to classify and segment the data, and form hypothesis formations.

1.4.3 KDD Process

Knowledge Discovery in Databases (KDD) refers to the overall process of discovering useful knowledge from data.

KDD is an integration of multiple technologies for data management such as database management and data warehousing, statistic machine learning, decision support, and others such as visualisation and parallel computing.

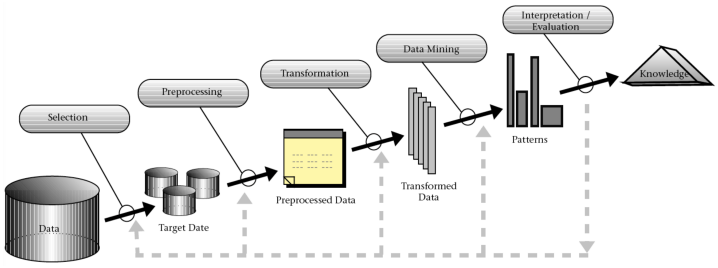

Figure 1.2: The KDD Process

1.5 Data Mining tasks

- Prediction methods
  - Use some variables to predict unknown or future values of other variables.
    - Supervised (classification or regression)
    - Unsupervised (clustering)
    - Semi-supervised
- Descriptive methods
  - Find human-interpretable patterns that describe the data.
    - Rule mining
    - Frequent patterns
    - Anomaly detection

1.6 Regression

1. Given a set of data points/instances along with "correct" answers or labels (training data)
2. Feed it to an algorithm which can learn a model and predict the correct value for an unseen data point (test data)
3. The learned model is an approximate representation of the training data

4. Predict continuous valued output (price in the previous example)
5. The algorithm should produce more right answers (goal)

1.7 Unsupervised Learning

- Unsupervised learning includes unlabeled data
- One way of doing this would be to cluster data into to groups
  - Group data points which are similar together, while separating dissimilar items as much as possible
  - This is a clustering algorithm

1.8 Association Rule Mining

- Given a set of records each of which contain some number of items from a given collection
- Produce dependency rules which will predict occurrence of an item based on occurrences of other items

Such rules can be used for marketing and sales promotion. Rule mining can also be used for inventory management, and much more.

1.9 Challenges in Data Mining

- Scalability
- Dimensionality
- Heterogenous or complex data
- Data ownership and distribution
- Privacy concern
- Data quality
- Evolving/streaming data

2 Data

2.1 Attributes

- An attribute is a property or characteristic of an object
  - Attribute are also known as variables, fields, characteristics, or features
- A collection of attributes describe an object
  - Objects are also known as records, points, cases, samples, entries, or instances

2.1.1 Values

Attribute values are numbers or symbols assigned to an attribute.
The distinction between attributes and attribute values are:

- Same attribute can be mapped to different attribute values
- Different attributes can be mapped to the same set of values

2.1.2 Types

Attribute Type	Operations	Transformations
Nomial	Mode, Entropy, Contingency, Correlation, χ^2-test	Any permutation of values
Ordinal	Median, Rank correlation, Percentiles, Run tests, Sign tests	Order-preserving change of values $new\_value = f(old\_value)$ where f is a monotonic function.
Interval	Mean, Standard deviation, t-tests Pearson's correlation, F-tests	$new\_value = a * old\_value + b$ where a and b are constants
Ratio	Geometric mean, Harmonic mean, Percent variation	$new\_value = a * old\_value$

2.1.3 Discrete and Continuous Attributes

Discrete attributes:

- Finite or countably infinite set of values
- Often integer or binaryvariables

Continuous attributes:

- Real numbers as attribute values
- Typically floating-point variables

2.2 Datasets

Some important characteristics of structured data are:

- Dimensionality
  - Curse of dimensionality
    - As the number of features or dimensions grow, the amount of data needed to generalize accurately grows exponentially.
    - When data moves from one dimensions to i.e. three dimensions, the given data fills less and less of the data space. In order to maintain an accurate representation of the space, the data for analysis grows exponentially.
    - When sorting or classifying data, low dimensional spaces tend to show the data as very similar, but in higher dimensions, the data might be further away from each other.
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale

2.2.1 Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes.

Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.

Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.

Document Data

Each document becomes a *term* vector. Each term is a component (attribute) of the vector.

The value of each component is the number of times the corresponding term occurs in the document.

Transaction Data

A special type of record data, where each record (transaction) involves a set of items.

For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

2.2.2 Graph Data

I.e. chemical structures.

2.2.3 Ordered Data

I.e. sequential data or gene sequences.

2.3 Data Quality

2.3.1 Noise

Noise refers to modification of original values.

2.3.2 Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.

2.3.3 Missing Values

Reasons for missing values may be that some information is not collected or aapplicable to all cases.

Handling missing values can be done by:

- Eliminating data objects
- Estimating the missing values
- Ignoring the missing values during analysis
- Replacing the missing values with all possible values weighted by their probabilities

2.3.4 Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates of one another. This is a major issue when merging data from heterogenous sources.

2.4 Distance/Similarity Functions

- Similarity
  - Numerical measure of how alike two data objects are
  - Higher value means more alike
  - Often falls in the range $[0, 1]$
- Dissimilarity
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

2.4.1 Similarity/Dissimilarity for Simple Attributes

Given that p and q are the attribute values for two data objects:

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & if \quad p = q \\ 1 & if \quad p \neq q \end{cases}$	$s = \begin{cases} 1 & if \quad p = q \\ 0 & if \quad p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval/Ratio	$d = \lvert p - q \rvert$	$s = -d,\ s = \frac{1}{1+d},\ s = 1 - \frac{d-min(d)}{max(d)-min(d)}$

2.4.2 Euclidean distance

Euclidean distance is defined in the following theorem, where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k_th attributes (components) or data objects p and q.

Theorem 2.1: Euclidean Distance

$$distance = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Standarization is necessary if scales differ.

2.4.3 Minkowski Distance

Minkowski Distance is a generalization of Euclidean Distance where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k_th attributes (components) or data objects p and q.

Theorem 2.2: Minkowski Distance

$$distance = \sqrt[r]{\sum_{k=1}^{n}|p_k - q_k|^r}$$

- $r = 1$: Manhattan distance
- $r = 2$: Manhattan distance
- $r \to \infty$: "Supremum" distance
  - This is the maximum difference between any component of the vectors

2.4.4 Properties of Distance Functions

Distances, such as the Euclidean distance, have some well known properties.

- Positive definiteness
  - $d(p, q) \geq 0$ for all p and q
  - $d(p, q) = 0$ only if $p = q$
- Symmetry
  - $d(p, q) = d(q, p)0$ for all p and q
- Triangle Inequality
  - $d(p, r) \leq d(p, q) + d(q, r)$ for all p. q and r

Important 2.3: Metric

A distance that satisfies the properties mentioned above is a metric.

2.4.5 Properties of Similarity Functions

Well known properties of similarities.

- Maximum Similarity
  - $s(p, q) = 1$ (or maximum similarity) only if $p = q$
- Symmetry
  - $s(p, q) = s(q, p)0$ for all p and q

2.5 Similarity/Coefficient

A common situation is that objects, p and q, have only binary attributes.
Let's define:

- M_{01} - Number of attributes where p is 0 and q is 1
- M_{10} - Number of attributes where p is 1 and q is 0
- M_{00} - Number of attributes where p is 0 and q is 0
- M_{11} - Number of attributes where p is 1 and q is 1

2.5.1 Simple Matching Coefficient

$$SMC = \frac{Numberofmatches}{Numberofattributes} \tag{2.1}$$

$$SMC = \frac{M_{00} + M_{11}}{M_{01} + M_{10} + M_{00} + M_{11}} \tag{2.2}$$

2.5.2 Jaccard Similarity/Coefficient

The Jaccard Similarity/Coefficient is used for categorical attributes and sets.

$$J = \frac{Numberof M_{11}}{Numberofnon - both - zeroattributes} \tag{2.3}$$

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \tag{2.4}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cup B|} \tag{2.5}$$

2.5.3 Cosine Similarity

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two documented vectors, then

$$similarity = \cos(\theta) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\|\boldsymbol{A}\|\|\boldsymbol{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.6}$$

2.6 Dot Product

Given two vectors a and b:

$$\overrightarrow{a} \cdot \overrightarrow{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \tag{2.7}$$

2.7 Data Preprocessing

2.7.1 Aggregation

Aggregation is to combine two or more attributes (or objects) into a single attribute (or object).

Purpose:

- Data reduction
- Reduce the number of attributes or objects
- Change of scale
- More "stable" data
- Often reduce variability

2.7.2 Data Sampling

Data sampling is the main technique employed for data selection. It is often used for both the preliminary investigation of the data and the final data analysis. Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming. Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

A good stategy for choosing the sample size is to aim for a sample size that is 10% of the population, as long as the sample size is smaller than 1000. It is also important to have a large enough sample size so that all attributes/groups are represented.

Sampling is effective when:

- The sample is representative
- The sample has approximately the same properties (of interest) as the original set of data

Types of sampling:

- Simple random sampling
  - Equal probability of selecting any item.

- Sampling without replacement
  - As each item is selected, it is removed from the population.
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
- Stratified sampling
  - Split the data into several partitions, then draw random samples from each partition.

Reservoir Sampling

Theorem 2.4: Reservoir Sampling

Keep a reservoir of r samples

1. Keep the first r items in memory
2. When the i^{th} item arrives ($i > r$)
   a) Keep the new item with probability $\frac{r}{i}$,
   or discard the new item with probability $1 - \frac{r}{i}$
   b) Discard one of the items in the reservoir at random
   if the new item was kept.

This means that as i increases, the probability of a new item being kept in the reservoir reduces.

2.8 Dimensionality Reduction

- Purpose
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining
  - Allow data to be more easily visualized
  - May help eliminate irrelevant features or reduce noise
- Techniques
  - Pronciple Component analysis
  - Singular Value Decomposition

3 Exploring Data

3.1 Summary Statistics

- Frequency
- Mode
- Percentiles
- Mean
- Median
- Range
- Variance

Theorem 3.1: Mean

$$mean(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Theorem 3.2: Median

$$median(x) = \begin{cases} x_{(r+1)} & \text{if } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m = 2r \end{cases}$$

Theorem 3.3: Variance

$$variance(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

Theorem 3.4: Average Absolute Deviation

$$AAD(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$$

Theorem 3.5: Mean Absolute Deviation

$$MAD(x) = median\left(\{|x_1 - \bar{x}| \cdots |x_m - \bar{x}|\}\right)$$

3.2 Visualisation

Visualisation is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported. Visualisation allows humans to detect general patterns and trends, as well as detect outliers and unusual patterns.

3.2.1 Representation

The representation is the mapping of information to a visual format. Data objects, their attributes, and the relationships among data objects are translated into graphichal elements such as points, lines, shapes, and colors.

3.2.2 Arrangement

Arrangement is the placement of visual elements within a display. This can make a large difference in how easy it is to understand the data.

3.2.3 Selection

Selection is the elimination or the de-emphasis of certain objects and attributes. Such selections may involve choosing a subset of attributes, or choosing a subset of objects. Only selecting some attributes can be done through dimensionality reduction, while selecting some objects can be done through stratified sampling. Some things to keep in mind when using data selection is to contain the diversity of the objects.

3.2.4 Visualisation Techniques

- Histogram plot
- Box plot
- Scatter plot

- Matrix plot
- Parallel coordinates plot
- Star plot

4 Decision Trees

4.1 Classification Definition