6. Data Mining - preparation

It's an **automatic**, **non-trivial extraction** of implicit,potentially useful and previously unknown information from available data.
It's performed by using proper **algorithm.**
Extracted *info* is represented by **patterns**(abstract models).

Analysis Methods

Descriptive methods: extracts *interpretable* and *readable* models describing data
Predictive methods: used to predict future values.

Analysis Techniques

Association rule extraction: Belongs to the descriptive methods, analyzes correlation between data
Classification: Analyzes a collection of data to build a **classifier**. Can belong to both *descriptive* or *predict* methods
Clustering: Partitioning of data objects based on *similar*(thus we need a **similarity** notion) properties.

Data preprocessing

Data

A collection of **data objects** described by their **attributes**.
A collection of **attributes** describes an **object.**

Attribute values

Same **attribute** can be mapped to different values(e.g.: height can be measured in feet or meter).
Different **attributes** can be mapped to the same set of values.
They can be **discrete** or **continuos**.
Also, **continuos** attributes can be converted to **discrete** through a process of **discretization.**

Attribute types

- **Nominal:** Cathegorical attributes. Holds **distinctness**($\neq$). E.g.: ID number,eye colors
- **Ordinal:** Cathegorical attributes Holds **distinctness** and **order**($\neq$ and $> <$). E.g.:rankings,grades

- **Interval:** Numerical attributes. Holds **distinctness** , **order**,**addition**($\neq$, $>$ $<$,+ and -). E.g.:calendar dates
- **Ratio:** Numerical attributes. **Holds** all properties(4th one is multiplication). E.g. : length,time,counts,kelvin temperature

Data set types

Record data

Tables

A collection of record with a fixed set of attributes.

Document data

It includes textual data that can be either **semi-structured** or **unstructured**.
Each document becomes a *term* vector.
Each *term* is a component (attribute) of the vector. Term's value is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

The document1 row is a term vector

Transaction data

Special type ofrecord data where each record,called **transaction** involves a set of items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread

Graph Data

For example a generic graph, a molecule or even webpages.
In general, all information that can be represented using a graph.

Ordered Data

- **Sequences of transactions:** 
  (AB) (D) and (CE) are different transactions. Thus it's a sequence of elements.
- **Genomic sequence data:** like a sequence of transactions but each element is a **gene**.
- **Spatio-temporal data:** Serie of data in different time instants in different geographical position.

Data quality

Certain problems must be dealt with in order to ensure data quality.
These problems may be:

- **Noise:** modification of the original values. For example distortion of someone's voice over the phone
- **Outliers:** data objects with characteristics that are considerably different than most of other data objects in the data set. They can be either :
  - Noise that **interfer** with our **analysis** and thus should be discard
  - The actual **goal** of our analysis. For example *credit card fraud*.
- **Duplicate data:** self explanatory.
- **Wrong data:** self explanatory.
- **Missing values:** There can be several reasons for a missing value:
  - Information not available
  - Certain attributes may not be applicable to all cases.
    They can be handled by **eliminating**,**ignoring** or **estimating** them.

Data preparation

List of actions to perform do prepare data:

Aggregation

Two or more attributes(or objects) into a single attribute(or object).
This can be performed to achieve:

- **Data reduction:** less data objects(or attributes)
- **Change of scale:** Cities aggregated into regions,states etc
- **Data stability:** Stable data has less variability

Sampling

Data reduction technique that *reduces* the cardinality of the set(number of objects in the collection).
It's the main technique employed for data selection.
It's based on the principle that *using a sample will work almost as well as using the entire data set, if the sample is representative*.

Simple Random Sampling

Equal probability of selecting a particular item:

- **Sampling without replacement:** As each item is selected, it's removed from the population so that it wont be picked again
- **Sampling with Replacement:** Objects are not removed from population, thus they can be picked again

Stratified Sampling

Splits the data into several partitions then draw random samples from each partition

Dimensionality Reduction

Curse Dimensionality: when dimensionality increases, data becomes increasingly sparse in the space that it occupies, and this impacts the clusting and outlier detection.
Dimensionality Reduction avoid **curse of dimensionality** and reduce amount of time and memory required by data mining algorithm.
One such technique is **PCA** which consists in finding a projection that capture the largest amount of variation in data.

Feature Subset Selection

Another way to reduce dimensionality of data

- Techniques
  - Brute-force approach
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches
    - Features are selected before data mining algorithm is run
  - Wrapper approaches
    - Use the data mining algorithm as a black-box to find best subset of attributes

Techniques related to feature subset selection

Feature Creation

Create **new** attributes more efficient, from an informative point of view, than the original attributes.
There are three general methodologies:

- Feature extraction
- Mapping Data to New Space(Fourier and Wavelet transform)
- Feature Construction

Discretization

Process of converting a continuous attribute into an ordinal one.
Basically it's a mapping from a potentially infinite set to a finite set.
Commonly used in classification.
It can be either:

- **supervised:** by using class labels
- **unsupervised:** by finding breaks in the data values
  - i.e.: from value 1.0 to value 10, we label it as **low**, from value 10.1 to value 20 we give it the label **medium** and from 20.1 onwards we label it as **high**

Binarization

Mapping of an attribute into one or **more** binary variables.
To map a **continuos** attribute, first map it to a **categorical one** according to the discretization process described above.
Then map it to a set of binary variables
i.e.: Low,medium,high into 1 0 0, 0 1 0, 0 0 1 respectively.

Attribute transformation

It's a function that maps a whole set of values to a new one.
Normalization and **standardization** are cathegorized as **attribute transformation** techniques.

Normalization

Values are scaled down so as to fall into a smaller range, usually $[-1, +1]$

Techniques

min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

z-score normalization $\quad v' = \frac{v - mean_A}{stand\_dev_A}$

decimal scaling

$$v' = \frac{v}{10^j} \qquad j \text{ is the smallest integer such that } \max(|v'|) < 1$$

Similarity and dissimilarity

They are numerical measure of how alike/different two data objects are.
The **similarity** is **higher** when two objects are more alike while the opposite is true for **dissimilarity.**
There are several techniques to compute these measures:

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

*We can see, in the ordinal and ratio section, how the **similarity** can be computed as the opposite of **dissimilarity***

Others techniques:(**TODO**)

- **Euclidean distance**
- **Minkowski distance**
- **Mahalanobis distance**

Correlation

Measures the linear relationship between two data objects