Introduction to Data Mining

Summary

Data Mining is a process of transforming raw data into information. To quote from Wikipedia, data mining "is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems."

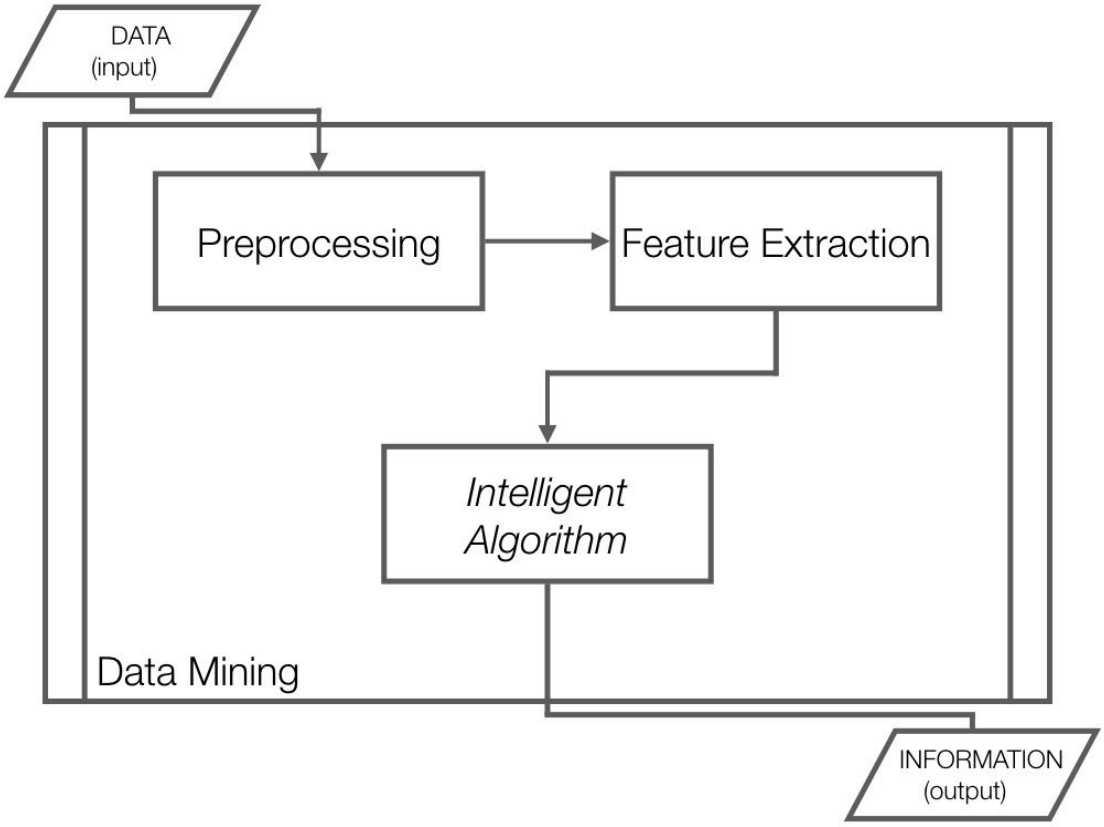
In this module you're going to learn about and experience a number of different aspects of data mining. This first lecture gives an overview of what data mining is and isn't, what types of data we might mine, the typical pipeline of all data mining systems and the two characteristic types of data mining approach.

Key points

- What is data mining?
    - In a single sentence: "the process of analysing and transforming raw data into information"
        - What do we mean by "data"?
            - Any sequence of one or more symbols that can be given meaning by specific acts of interpretation
            - Data IS NOT information
                - You need to interpret or process data in order to extract information
        - What do we mean by "information"?
            - "Actionable Knowledge" - knowledge that allows us to make sensible decisions
            - In terms of prediction, the ability to make predictions with a likelihood better than chance
            - In terms understanding, the ability as humans to *make sense* of the data
    - More detailed definition:
        - Data mining is the process of the discovery of patterns and models that are:
            - Valid: hold on new data with some certainty
            - Useful: should be possible to act on the item
            - Unexpected: non-obvious to the system
            - Understandable: humans should be able to interpret the pattern
        - Fundamentally, there are two types of data mining:
            - Predictive data mining
                - Use some variables to predict unknown or future values of other variables
                    - e.g. Recommender systems, Spam filters
            - Descriptive data mining
                - Find human-interpretable patterns that describe the data
                    - e.g. clustering
- What kinds of data are we interested in mining?
    - All kinds of data: text, multimedia (images, video, audio), "signals", networks, tabular numeric data, relational data, ...

- Lots of different ways in which we might categorise this:
    - Structured (e.g. tabular/relational/...) versus unstructured (images/text/...)
    - Dynamic (data that changes over time) versus static
- Also need to consider data mining where we concentrate on a single modality (Unimodal; e.g. the analysis of a corpus of text documents), versus data mining where we analyse different modalities of data in harmony (multimodal; e.g. the analysis of a dataset of images and textual captions)

- What is the typical data mining pipeline?



- Key components:
    - Preprocessing - cleaning the data, parsing it into usable formats, ...
    - Feature Extraction - making measurements, extracting statistics, creating features, ...
    - *Intelligent Algorithms* - The core of the data mining process; falls into descriptive and predictive techniques

Further Reading

- Wikipedia (https://en.wikipedia.org/wiki/Data_mining) has quite a good page describing data mining as a concept, which also touches on some of the non-technical aspects like legal and ethical issues surrounding the use of data mining.