Data Mining and Knowledge Discovery

Ilaria Battiston

Winter Semester 2019-2020

Contents

1 Data Mining Process

1.1 Data Sources

Examples of data sources are:

- Industrial process data, with many uses such as optimizing processes and predicting, but requires access and management of large quantities of information;
- Business data, mainly generated by shopping transactions, and having applications of market basked analysis and customer segmentation;
- Text (structured data),subject of natural language processing to understand keywords and meaning of documents;
- Image data, important source in terms of amount and especially targeted by deep learning;
- Biomedical data, such as genome and laboratory data.

Data can be projected onto two-dimensional planes, assigning visualization variables (color, height) to characteristics of data.

1.2 Definitions

Data mining is the process concerning the extraction of knowledge from data. Knowledge is defined as interesting patterns, therefore general, non trivial and comprehensive information.

Knowledge discovery consists, in fact, in preprocessing the a priori knowledge and extracting more, followed by evaluation (postprocessing) of the obtained data. The process can be summarised as follows:

1. Preparation: collecting data, planning, generating and selecting features. Generally, this happens before involving data analysts;
2. Preprocessing, all the operations of normalization, cleansing, filtering and correction;
3. Analysis, the whole process of visualization, clustering, correlation etc.;
4. Postprocessing, the interpretation of results using a systematic approach with related documentation.

Patterns aren't always obvious in the beginning: the application of computer systems to the analysis (data analytics) of large datasets can offer great support to decisions.

Getting feedback must involve expert in related areas, such as statistics, pattern recognition, machine learning and operations research.

2 Data and Relations

One of the most famous datasets is Iris, using $n = 150$ vectors of dimension $p = 4$. It contains 50 instances for each of 3 classes, and 4 components representing characteristics of the flower. This collection is widely used for classification, and has all the properties of modern datasets.

Typical questions to ask are whether the data is correct (rounding errors, false assignments) and the correlation between two variables, to have a better identification of each type with its features.