Data Mining and Knowledge Discovery

Ilaria Battiston [*]

Winter Semester 2019-2020

[*] All notes are collected with the aid of material provided by T. Runkler.

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.

Numerical data can have very different meanings: according to those and their type (usually nominal values with their occurrence), they can be treated in many ways. Examples of scales are:

- Ratio $(\cdot, /)$ to detect the generalized mean, which allows to add an exponent $\alpha \in \mathbb{R}$ to the mean summation;
- Interval $(+, -)$ to detect the mean and compare it, getting information on outliers;
- Ordinal $(<, >)$ to detect the median by ordering them;
- Nominal $(=, \neq)$, to detect the mode since the only operation to perform is equality counting the number of occurrences in a bar chart.

After understanding the scale, there is a systematic arrangement of data $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ in a matrix. It has to be written in a way in which each row represents one observation, and each column represents a feature.

Object data has the form $O = \{o_1, \ldots, o_n\}$ while relational data is in the form of a matrix R. There are infinitely relationships which can be specified, yet two of the most important ones are similarity and dissimilarity. The feature vectors have to be the same for two objects to determine similarity.

The measure can be computed using the Euclidean distance definition, which has three main properties:

1. $d(x, y) = d(y, x)$;
2. $d(x, y) = 0 \leftrightarrow x = y$;
3. $d(x, z) \leq d(x, y) = d(y, z)$.

The third one is known as triangle inequality and is particularly useful for Euclidean mapping. It is possible to apply other functions.

The norm $d(x, y) = \|x - y\| = \|y - x\| = d(y, x)$ (not to be confused with the hyperbolic norm, a product) is used to measure dissimilarity and has the following properties:

1. $\|x\| + 0 \leftrightarrow x = (0 \ldots, 0)$;
2. $\|a \cdot x\| = \|a\| \cdot \|x\|$;
3. $\|x + y\| \leq \|x\| + \|y\|$.

Matrices have different kinds of norms: Euclidean, diagonal, Mahalanobis, Minkowski and many more. The usual method is Euclidean.

Similarity is defined through formulas such as Dice and Jaccard or proximity measures:

1. $s(x, y) = s(y, x)$;
2. $s(x, y) \leq s(x, x)$;
3. $s(x, y) \geq 0$.

Similarity measures are used to compare internet pages according to the words in them, and give a general idea of the frequency distribution over nominal data. Vectors can be multiplied to obtain similarity as well, yet output needs to be normalized.

Distances for sequences and text are symbol distance, Hamming and edit (Levenshtein). The con with Hamming is the need for two strings to have the same size. Edit distance is a good way to solve this problem, minimizing the number of operations to get the maximum similarity.

3 Signal processing

Continuous signals are sampled starting from an original sinusoidal shape to obtain a set of discrete values which approximately represent the function.

Shannon's sampling theorem states that having band $s(t)$ and Fourier spectrum $|s(j2\pi f)| = 0$ for $|f| > f_{max}$, the sampling time T must be $T_s < \frac{1}{2f_{max}}$ (Nyquist condition) to completely reconstruct the signal.

The theorem assumes infinite number of observations, not really possible in practice: arbitrary values are then chosen for sampling time.

Quantization is the process of constraining an input from continue values to discrete, irreversible because of the quantization error (information gets lost in the process). A good step size needs to be chosen according to scale and placement of samples.

4 Data preprocessing

Data preprocessing is useful to identify individual errors and perform operations such as merging, normalization and modeling.

The first assumption to make is for errors to be random, due to transmission and measurement, so that they can be treated as additive noise.

Outliers can be caused by processing errors (wrong or permuted data), therefore if there are single incorrect values they may be removed - although there is no certain way to ensure a sample has been subject of mistakes.

Detecting outliers can be done using statistical measures (sigma rule) or imposing stronger criteria on range and distribution.

Errors can be replaced with mean, median, minimum or maximum of the valid feature data x^i, or take the value of the nearest neighbor (having the smallest norm). Examples of error handling:

- Linear interpolation for equidistant time series;
- Linear interpolation for non-equidistant time series;
- Nonlinear interpolation (splines);
- Estimation by regression;
- Filtering;
- Outlier removal (of the complete vector);
- Feature removal.