

# Data Analysis, Mathematical Statistics, Machine Learning, Data Mining: Similarities and Differences

Boris Mirkin

Division of Applied Mathematics and Informatics, National Research University Higher School of Economics, Moscow RF

Department of Computer Science, Birkbeck University of London UK

## Abstract

An attempt is made to bring some structure in the meanings of the title subjects based on the perspectives at which they view the data. A rigid framework for Data Analysis is proposed pointing out its role in augmenting the knowledge of a phenomenon the data relate to. In particular, the task of summarization as a general problem is discussed. The ingredients of a data analysis problem are described as a base for positioning of various approaches in data analysis and mathematical statistics.

## Introduction

The similarities between the subjects mentioned in the title relate to two fundamental facts: (1) all of them develop methods and procedures to process data, and (2) any data processing algorithm or procedure may belong to any, or better to say, all of them. The differences are in the different perspectives. The difference in perspectives does not affect the procedures but it does affect the choice of them and, even more so, interpretation of concepts and results.

Before I go any further, let me give an example by presenting two concepts that are common at analyzing data: the average and correlation, in the different perspectives.

**Average:** Given observed values  $y_1, y_2, \dots, y_N$  of a variable, compute the average value  $\bar{y} = \sum_i y_i / N$  – same method in all the perspectives.

The meaning:

- An unbiased estimate of the central moment of the probability distribution from which values  $y_i$  have been independently and randomly drawn (Classical mathematical statistics perspective).
- An approximate central tendency  $c$  of the values minimizing the approximation criterion  $L_2 = \sum_i |y_i - c|^2 / N$  (Data analysis perspective).
- An estimate of the most likely value of the variable at a next  $(N+1)$ -th observation to be updated according to formula  $c_{N+1} = (Nc_N + y_N) / (N+1)$  (Machine learning perspective).
- “Norm” to be used for finding “interesting patterns” as those most deviating from the norm (Data mining perspective).

**Correlation coefficient** between two variables: given observed pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$  of two variables,  $x$  and  $y$ , make a z-scoring standardization and compute the average product  $r = \sum_i x_i y_i / N$ .

The meaning:

- An estimate of parameter  $\rho$  of a bivariate Gaussian density function whose standardized covariance matrix is  $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$  (Classical statistics perspective).
- Parameter defining the optimal linear approximation of  $y$  as  $ax+b$  so that  $a=r\sigma_y/\sigma_x$  where  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$ , respectively, from their means; the variance of the residuals being less than that of  $y$ , by  $r^2$  (Data analysis perspective).
- An estimate of the parameter  $\rho$  of a bivariate Gaussian density function whose value can be updated incrementally upon newly observed pairs arriving (Machine learning perspective).
- An association scoring index experimentally proven to be useful in applications such as collaborative filtering (advising the user based on behavior of similar users) (Data mining perspective).

## Data analysis and other paradigms

The term Data Analysis has been conventionally used as an extension of mathematical statistics, starting from the developments in factor analysis, cluster analysis and other multivariate techniques in Psychology before the WWII and eventually bringing forth the concepts of “exploratory” data analysis and “confirmatory” data analysis in statistics (see, for example, Tukey 1977). The former was supposed to cover a set of techniques for finding patterns in data, and the latter to cover more conventional mathematical statistics approaches for hypothesis testing. Overall, the term Data Analysis is usually applied as an umbrella to cover all the various activities mentioned above, with an emphasis on mathematical statistics and its extensions.

Classical statistics takes the view of data as a vehicle to fit and test mathematical models of the phenomena the data refer to. The data mining and knowledge discovery discipline uses data to add new knowledge in any format. It should be sensible then to consider the data analysis as a perspective relating to an intermediate level to contribute to the theoretical – rather than any – knowledge of the phenomenon. This should focus on ways of augmenting or enhancing theoretical knowledge of the specific domain which the data being analyzed refer to. The term “knowledge” encompasses many a diverse layer or form of information, starting from individual facts to those of literary characters to major scientific laws. But when focusing on a particular domain the dataset in question comes from, its theoretical knowledge structure can be considered as comprised of just two types of elements: (i) concepts and (ii) statements relating them. Concepts are terms referring to aggregations of similar entities, such as apples or plums, or similar categories such as fruit comprising both apples and plums, among others. When created over data objects or features, these are referred to, in data analysis, as clusters or factors, respectively. Statements of relation between concepts express regularities relating different categories. Two features are said to correlate when a co-occurrence of specific patterns in their values is observed as, for instance, when a feature’s value tends to be the square of the other feature. The observance of a correlation pattern can lead sometimes to investigation of a broader structure behind the pattern, which may further lead to finding or developing a theoretical framework for the phenomenon in question from which the correlation follows. It is useful to distinguish between quantitative correlations such as functional dependencies between features and categorical ones expressed conceptually, for example, as logical production rules or more complex structures such as decision trees. Correlations may be used for both understanding and prediction. In applications, the latter is by far more important. Moreover, the prediction problem is much easier to make sense of operationally so that the data analysis approaches so far have paid much attention to this.

What is said above suggests that there are two main pathways for augmenting knowledge: (i) developing new concepts by “summarizing” data and (ii) deriving new relations between concepts by analyzing “correlation” between various aspects of the data. The quotation marks are used here to point out that each of the terms, summarization and correlation, much extends its conventional meaning. Indeed, while everybody would agree that the average mark does summarize the marking scores on test papers, it would be more daring to see in the same light derivation of students’ hidden talent scores by approximating their test marks on various subjects or finding a cluster of similarly performing students. Still, the mathematical structures behind each of these three activities – calculating the average, finding a hidden factor, and designing a cluster structure – are analogous, which suggests that classing them all under the “summarization” umbrella may be reasonable. Similarly, term “correlation” which is conventionally utilized in statistics to only express the extent of linear relationship between two or more variables, is understood here in its generic sense, as a supposed affinity between two or more aspects of the same data that can be variously expressed, not necessarily by a linear equation or by a quantitative expression at all.

It would be useful to spell out that view of the data as a subject of computational data analysis that is adhered to here. Typically, in sciences and in statistics, a problem comes first, and then the investigator turns to data that might be useful in advancing towards a solution. In computational data analysis, it may also be the case sometimes. Yet the situation is reversed frequently. Typical questions then would be: Take a look at this data set - what sense can be made out of it? – Is there any structure in the data set? Can these features help in predicting those? This is more reminiscent to a traveler’s view of the world rather than that of a scientist. The scientist sits at his desk, gets reproducible signals from the universe and tries

to accommodate them into the great model of the universe that the science has been developing. In contrast, the traveler deals with what comes on their way. Helping the traveler in making sense of data is the task of data analysis. This view also underlies the development of data mining, though the aspect of data being available as a database, quite important in data mining, is rather tangential to data analysis.

### **Other approaches**

The two-fold goal clearly delineates the place of the data analysis core within the set of approaches involving various data analysis tasks. Here is a list of some popular approaches:

- Classification – this term applies to denote either a meta-scientific area of organizing the knowledge of a phenomenon into a set of separate classes, to structure the phenomenon and relate different aspects of it to each other, or a discipline of supervised classification, that is, developing rules for assigning class labels to a set of entities under consideration. Data analysis can be utilized as a tool for designing the former, whereas the latter can be thought of as a problem in data analysis.
- Cluster analysis – is a discipline for obtaining (sets of) separate subsets of similar entities or features or both from the data, one of the most generic activities in data analysis.
- Computational intelligence – a discipline utilizing fuzzy sets, nature-inspired algorithms, neural nets and the like to computationally imitate human intelligence, which does overlap other areas of data analysis.
- Data mining – a discipline for finding interesting patterns in data stored in databases, which is considered part of the process of knowledge discovery. This has a significant overlap with computational data analysis. Yet data mining puts more emphasis on fast computations in large databases and finding “interesting” associations and patterns.
- Document retrieval – a discipline developing algorithms and criteria for query-based retrieval of as many relevant documents as possible, from a document base, which is similar to establishing a classification rule in data analysis. This area has become most popular with the development of search engines such as Google or Yahoo! over the internet.
- Factor analysis – a discipline emerged in psychology for modeling and finding hidden factors in data, which can be considered part of quantitative summarization in data analysis.
- Genetic algorithms – an approach to globally search through the solution space in complex optimization problems by representing solutions as a population of “genomes” that evolves in iterations by mimicking micro-evolutionary events such as “cross-over” and “mutation”. This can play a role in solving optimization problems in data analysis.
- Knowledge discovery – a set of techniques for deriving quantitative formulas and categorical productions to associate different features and feature sets, which hugely overlaps with the corresponding parts of data analysis.
- Mathematical statistics – a discipline of data analysis based on the assumption of a probabilistic model underlying the data generation and/or decision making so that data or decision results are used for fitting or testing the models. This obviously has a lot to do with data analysis, including the idea that an adequate mathematical model is a finest knowledge format.
- Machine learning – a discipline in data analysis oriented at producing classification rules for predicting unknown class labels at entities usually arriving one by one in a random sequence.
- Neural networks – a technique for modeling relations between (sets of) features utilizing structures of interconnected artificial neurons; the parameters of a neural network are learned from the data with an incremental approach such as error back propagation underlied by the gradient descent algorithm.
- Nature-inspired algorithms – a set of contemporary techniques for optimization of complex functions such as the squared error of a data fitting model, using a population of admissible solutions evolving in iterations mimicking a natural process such as genetic recombination or ant colony or bee swarm search for foods.
- Optimization – a discipline for analyzing and solving problems in finding optima of a function such as the difference between observed values and those produced by a model whose parameters are being fitted (error).
- Pattern recognition – a discipline for deriving classification rules (supervised learning) and clusters (unsupervised learning) from observed data.

- Text analysis – a set of techniques and approaches for the analysis of unstructured text documents such as establishing similarity between texts, text categorization, deriving synopses and abstracts, etc.

### **Main problems of Data Analysis and their structure**

Let us look further at the core methods for enhancing knowledge by finding in data either

- (a) Correlation among features (Correl) or
- (b) Summarization of entities or features (Summary),

in either of two ways, quantitative (Q) or categorical (C). Combining these two bases makes four major groups of methods: CorQ, CorC, SumQ, and SumC that form the core of data analysis. It should be pointed out that currently different categorizations of tasks related to data analysis prevail: the classical mathematical statistics focuses mostly on mathematically treatable models (see, for example, Hair et al. 2010), whereas the system of machine learning and data mining expressed by the popular account by Duda and Hart (2001) concentrates on the problem of learning categories of objects, thus leaving such important problems as quantitative summarization outside.

A correlation or summarization problem typically involves the following five ingredients:

- Stock of mathematical structures sought in data
- Computational model relating the data and the mathematical structure
- Criterion to score the match between the data and structure (fitting criterion)
- Method for optimizing the criterion
- Visualization of the results.

Here is an outline of the most popular ingredients:

Mathematical structures:

- linear combination of features;
- neural network mapping a set of input features into a set of target features;
- decision tree built over a set of features;
- cluster of entities;
- partition of the entity set into a number of non-overlapping clusters.

When the type of mathematical structure to be used has been chosen, its parameters are to be learnt from the data. Each learning method relies on a computational model involving a function scoring the adequacy of the mathematical structure underlying the rule – a criterion, and, usually, visualization aids. The data visualization is a way to represent the found structure to human eye. In this capacity, it is an indispensable part of the data analysis. The criterion measures either the deviation from the target (to be minimized) or goodness of fit to the target (to be maximized).

Currently available computational methods to optimize a criterion encompass three major groups:

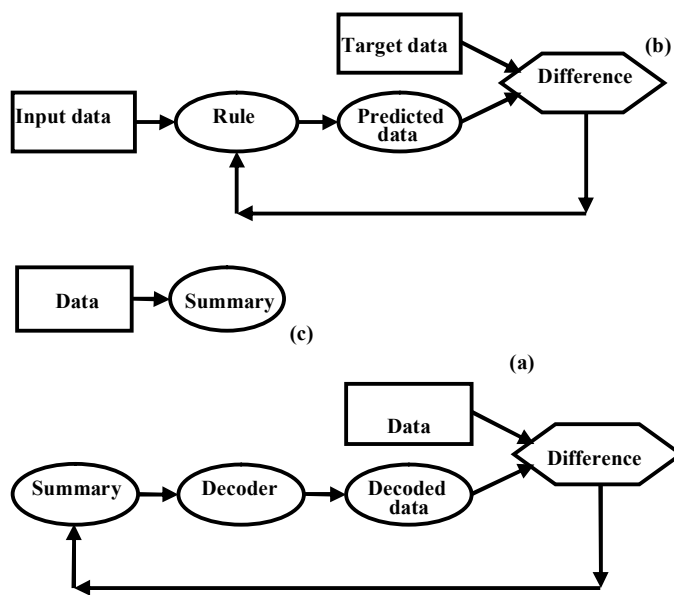
- ❖ global optimization, that is, finding the best possible solution, computationally feasible sometimes for linear quantitative and simple discrete structures;
- ❖ local improvement using such general approaches as:
  - gradient ascent and descent
  - alternating optimization
  - greedy neighborhood search (hill climbing)
- ❖ nature-inspired approaches involving a population of admissible solutions and its iterative evolution, an approach involving relatively recent advancements in computing capabilities, of which the following will be used in some problems:
  - genetic algorithms
  - evolutionary algorithms
  - particle swarm optimization

It should be pointed out that currently there is no systematic description of all possible combinations of problems, data types, mathematical structures, criteria, and fitting methods available. Yet there are some

generic and better explored problems in each of the four data analysis groups that can be safely claimed as being prototypical within the groups:

	Quant	Principal component analysis
Summary	Categ	Cluster analysis
	Quant	Regression analysis
Correl	Categ	Supervised classification

The four approaches on the right have emerged in different frameworks and usually are considered as unrelated. However, they are related in the context of data analysis. Moreover, they can be unified by the so-called data-driven modeling together with the least-squares criterion that I have adopted for them in Mirkin 2011.



**Figure 1.** A diagram for coder/decoder data summarization (a) versus learning input-target correlation (b) or summarization with no decoder (c). Rectangles are for observed data, ovals for computational constructions, hexagons for feedback comparisons.

In a correlation problem, features are divided in two groups, those of “target” and those of “input”. The goal is to establish a relation between the input and target features so that the target features values can be reliably predicted from values of the input features. In a summarization problem, in contrast to a correlation problem, the features are not divided as those belonging to input or output of the phenomenon under consideration. One may think of this as that all features available are target features so that those to be constructed as a summary are in fact “hidden input features”.

The formal structures of the correlation and summarization problems can be captured in the schemes presented on Figure 1. The structure of a correlation problem is conventional (Fig. 1(a)); the conventional structure of a summarization problem is on Fig. 1 (c); the feedback based structure on Fig. 1 (b) has been proposed by the author (Mirkin 2011).

The structure of a summarization problem may be likened to that of a correlation problem if a rule is provided to predict all the original features from the summary. That is, the original data in the summarization problem act as target data in the correlation problem. That implies that there should be two rules involved in a summarization problem: one for building the summary, the other to provide a feedback from the summary to the observed data. Unlike in the correlation problem, though, here the feedback rule

must be pre-specified so that the focus is on building a summarization rule rather than on using the summary for prediction; this is why we refer to the feedback rule as a “decoder” rather than a “predictor”. A proper consideration of the structure of a summarization problem should rely on the existence of a decoder to provide the feedback from a summary back to the data and make the summarization process more or less similar to that of the correlation process (see Figure 1 (a) versus 1 (b)). More exactly, a decoder is a device that translates the summary representation encoded in the chosen summarization rule back into the original data format. This allows us to utilize the same criterion of minimization of the difference between the original data and those output by the decoder: the less the difference, the better.

### **Data recovery approach**

The difference can be measured by a conventional criterion of the summary squared error. The criterion is part of a unifying data-recovery perspective that has been developed in mathematical statistics for fitting probabilistic models and can be easily extended to data analysis. In data analysis, this perspective is useful not only for supplying a nice fitting criterion but also because it involves the decomposition of the data scatter into “explained” and “unexplained” parts in all four methods. The data recovery approach takes in a type of mathematical structure to model the data and it proceeds in three stages:

- (1) fitting a model representing the structure to the data (this can be referred to as “coding”),
- (2) deriving data from the model in the format of the data used to build the model (this can be referred to as “decoding”), and
- (3) looking at the discrepancies between the observed data and those recovered from the model. The smaller are the discrepancies, the better the fit – this is a principle underlying the data-driven modeling approach.

The data recovery approach in data summarization is based on the assumption that there is a regular structure in the phenomenon of which the observed dataset informs. This regular structure  $A$  is the summary to be found. When  $A$  is determined, this can feed back to the observed data  $Y$  in the format of the decoded data  $F(A)$  that should coincide with  $Y$  up to residuals, that are due to possible flaws in any or all of the following three aspects:

- (a) bias in entity sampling,
- (b) selecting and measuring features,
- (c) adequacy of the set of admissible  $A$  structures to the phenomenon in question.

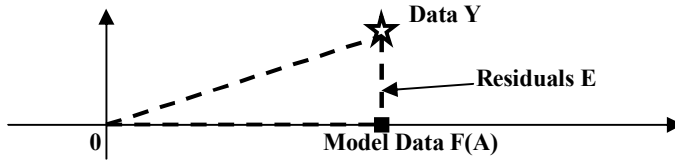
Each of these three can drastically affect results. However, so far only the simplest of the aspects, (a) sampling bias, has been addressed scientifically, in the classical statistics, – as a random bias, due to the probabilistic nature of the data. The other two are subjects of much effort in specific domains but not in the general computational data analysis framework as yet. Rather than focusing on accounting for the causes of errors, let us consider the underlying equation in which the errors are looked at as a whole:

$$\text{Observed\_Data } Y = \text{Model\_Data } F(A) + \text{Residuals } E \quad (1)$$

This equation brings in an inherent data recovery criterion for the assessment of the quality of the model  $A$  in recovering data  $Y$  - according to the level of residuals  $E$ : the smaller the residuals, the better the model. Since a data model typically involves unknown parameters, this naturally leads to the idea of fitting these parameters to the data in such a way that the residuals become as small as possible.

In many cases this principle can be rather easily implemented as the least squares principle because of an extension of the Pythagoras theorem relating the square lengths of the hypotenuse and two other sides in a right-angle triangle connecting “points”  $Y$ ,  $F(A)$  and  $0$  (see Figure 2). The least squares criterion requires fitting the model  $A$  by minimizing the sum of the squared residuals. Geometrically, it often means an orthogonal projection of the data set considered as a multidimensional point onto the space of all possible models represented by the  $x$  axis on Figure 4.2. In such a case the dataset (pentagram), its projection (rectangle) and the origin ( $0$ ) form a right-angle triangle for which a multidimensional extension of the Pythagoras’ theorem holds. The theorem states that the squared length of the hypotenuse is equal to the sum of squares of two other sides. The squared hypotenuse translates into the data scatter, that is, the sum of all the data entries squared, being decomposed in two parts, the part explained by the summary model  $A$ , that is, the contribution of the line between  $0$  and rectangle, and the part left unexplained by  $A$ . The latter part is the contribution of the residuals  $E$  expressed as the sum of squared

residuals, which is exactly the least squares criterion. This very decomposition can be employed in the problems of linear and non-linear regression, classification trees, Principal component analysis and K-Means clustering, as well as additive clustering.



**Figure 2.** Geometric relation between the observed data (pentagram), the fitted model data (black rectangle), and the residuals (connecting line).

When the data can be considered as a random sample from a multivariate Gaussian distribution, the least squares principle can be derived, under some simplifying assumptions, from a major statistical principle, that of maximum likelihood. In the data analysis framework, the data do not necessarily come from a probabilistic population. Still, the least squares framework frequently provides for solutions that are both practically relevant and theoretically sound.

A decoder based summarization problem can be stated as follows. Given  $N$  vectors forming a matrix  $Y = \{(y_i)\}$  with rows  $y_i = (y_{i1}, \dots, y_{iV})$  of  $V$  features observed at entities  $i = 1, 2, \dots, N$  and a set of admissible summary structures  $A$  with decoder  $D: A \Rightarrow R^p$ , build a summary

$$A = F(Y), \quad A \in \mathcal{A}$$

such that the error, which is the difference between the decoded data  $D(A)$  and observed data  $Y$ , is minimal over the class of admissible rules  $F$ . More explicitly, one assumes that

$$Y = D(A) + E \quad (2)$$

where  $E$  is matrix of residual values, or errors: the smaller the errors, the better the summarization  $A$ . According to the least-squares approach, the errors are minimized by minimizing the summary, or average, squared error:

$$E^2 = \langle Y - D(A), Y - D(A) \rangle = \langle Y - D(F(Y)), Y - D(F(Y)) \rangle \quad (3)$$

with respect to all admissible summarization rules  $F$ .

Expression (3) can be further decomposed into

$$E^2 = \langle Y, Y \rangle - 2\langle Y, D(A) \rangle + \langle D(A), D(A) \rangle.$$

In many data summarization methods, such as the Principal component analysis and K-Means clustering, the set of all possible decodings  $D(F(Y))$  forms a linear subspace. In this case, the data matrices  $Y$  and  $D(A)$ , considered as multidimensional points, form a “right-angle triangle” around the origin 0, as presented on Figure 2. In such a case  $\langle Y, D(A) \rangle = \langle D(A), D(A) \rangle$  and the square error (3) becomes part of a multivariate analogue to the Pythagoras equation relating the squares of the “hypotenuse”,  $Y$ , and the “sides”,  $D(A)$  and  $E$ :

$$\langle Y, Y \rangle = \langle D(A), D(A) \rangle + E^2 \quad , \quad (4)$$

or on the level of matrix entries,

$$\sum_{i \in I} \sum_{v \in V} y_{iv}^2 = \sum_{i \in I} \sum_{v \in V} d_{iv}^2 + \sum_{i \in I} \sum_{v \in V} e_{iv}^2 \quad (4')$$

The data is an  $N \times V$  matrix  $Y = (y_{iv})$  that can be considered as either set of rows/entities  $y_i$  ( $i = 1, \dots, N$ ) or set of columns/features  $y_v$  ( $v = 1, \dots, V$ ) or both. The item on the left in (4') is usually referred to as the data scatter and denoted by  $T(Y)$ ,

$$T(Y) = \sum_{i \in I} \sum_{v \in V} y_{iv}^2 \quad (5)$$

## Conclusion

In this text, an attempt is made to distinguish between the four subjects, Classical mathematical statistics, Data analysis, Machine learning and Data mining, not because of the methods each of them encompass, but rather describing the difference in the perspectives they view the data. An attempt is made to conceptualize the contents of Data analysis as related to augmenting the theoretical knowledge of the phenomenon or process in question, thus leading to two principal ways for doing so – correlation and summarization. The summarization, when supplied with a “decoder” becomes an important, feedback based, activity in data analysis, which is yet to be properly developed. In this, such summarization methods as Principal Component Analysis, which are being neglected in the Machine Learning perspective, are raised to the level of main approaches. Five ingredients of a data analysis task are highlighted to give a way for properly systematizing those. As a byproduct a proper positioning to such important approaches as genetic algorithms or neuron nets is achieved.

## References

- M. Berthold, D. Hand (2003), *Intelligent Data Analysis*, Springer-Verlag.
- L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984) *Classification and Regression Trees*, Belmont, Ca: Wadsworth.
- R.O. Duda, P.E. Hart, D.G. Stork (2001) *Pattern Classification*, Wiley-Interscience, ISBN 0-471-05669-3
- S.B. Green, N.J. Salkind (2003) *Using SPSS for the Windows and Mackintosh: Analyzing and Understanding Data*, Prentice Hall.
- J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson (2010) *Multivariate Data Analysis*, 7th Edition, Prentice Hall, ISBN-10: 0-13-813263-1.
- J. Han, M. Kamber (2010) *Data Mining: Concepts and Techniques*, 3<sup>d</sup> Edition, Morgan Kaufmann Publishers.
- S. S. Haykin (1999), *Neural Networks* (2nd ed), Prentice Hall, ISBN 0132733501.
- M.G. Kendall, A. Stewart (1973) *Advanced Statistics: Inference and Relationship* (3d edition), Griffin: London, ISBN: 0852642156.
- L. Lebart, A. Morineau, M. Piron (1995) *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris, ISBN 2-10-002886-3.
- H. Lohninger (1999) *Teach Me Data Analysis*, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.
- C.D. Manning, P. Raghavan, H. Schütze (2008) *Introduction to Information Retrieval*, Cambridge University Press.
- B. Mirkin (2005) *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall/CRC.
- B. Mirkin (2011) *Core Concepts in Data Analysis: Summarization, Correlation, Visualization*, Springer, London.
- T.M. Mitchell (2005) *Machine Learning*, McGraw Hill.
- T. Soukup, I. Davidson (2002) *Visual Data Mining*, Wiley and Son.
- J.W. Tukey (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading MA.
- V. Vapnik (2006) *Estimation of Dependences Based on Empirical Data*, Springer Science + Business Media Inc., 2d edition.
- A. Webb (2002) *Statistical Pattern Recognition*, Wiley and Son.
- S.M. Weiss, N. Indurkha, T. Zhang, F.J. Damerau (2005) *Text Mining: Predictive Methods for Analyzing Unstructured Information*, Springer Science+Business Media. ISBN 0-387-95433-3.