Government of Russian Federation

Federal State Autonomous Educational Institution of High Professional Education

National Research University «Higher School of Economics»

Faculty of Computer Science
School of Data Analysis and Artificial Intelligence

Syllabus for the course
«Data Analysis and Data Mining»

for Bachelor degree specialisation
010400.62 «Applied Mathematics and Informatics»

Author:
Boris G. Mirkin, professor
bmirkin@hse.ru

APPROVED BY

Programme Academic Supervisor of specialisastion 010400 «Applied Mathematics and Information Science»
Anton S. Konushin

_____

«___» __ _____ 2014 г.

Approved on the meeting of the School of Data Analysis and Artificial Intelligence
Head of the School Sergei O. Kuznetsov

_____

«___» __ _____ 2014 г.

Recommended by Academic Council of the Programme «Applied Mathematics and Information Science»

Manager of the School of Data Analysis and Artificial Intelligence Larisa I. Antropova

«___» __ _____ 2014 г.

_____

Moscow, 2014

Data Analysis and Data Mining

A Syllabus

I. Introduction: Subject and background

Instructor and author

Professor Boris Mirkin, PhD (Mathematics), ScD (Engineering), Emeritus Professor of the University of London

Teaching assistant

Ekaterina Chernyak, MSc (Applied Mathematics and Informatics)

Summary

This is an unconventional course in modern Data Analysis. Its contents are heavily influenced by the idea that data analysis should help in enhancing and augmenting knowledge of the domain as represented by the concepts and statements of relation between them. According to this view, two main pathways for data analysis are summarization, for developing and augmenting concepts, and correlation, for enhancing and establishing relations. Visualization, in this context, is a way of presenting results in a cognitively comfortable way. The term summarization is understood quite broadly here to embrace not only simple summaries like totals and means, but also more complex summaries: the principal components of a set of features and cluster structures in a set of entities. Similarly, correlation here covers both bivariate and multivariate relations between input and target features including classification trees and Bayes classifiers.

The material presented in this perspective makes a unique mix of subjects from the fields of statistical data analysis, data mining, and computational intelligence, each of which follows a different system of the subjects.

Another feature of the module is that its main thrust is to give an in-depth understanding of a few basic techniques rather than to cover a broad spectrum of approaches developed so far. Most of the described methods fall under the same least-squares paradigm for mapping an "idealized" structure to the data. This allows me to bring forward a number of relations between methods that are usually overlooked. Although the in-depth study approach involves a great deal of technical details, these are encapsulated in specific fragments termed "formulation" parts. The main, "presentation", part is delivered with no mathematical formulas and explains a method by actually applying it to a small real-world dataset – this part can be read and studied with no concern for the formulation at all. There is one more part, "computation", targeted at studying the computational data processing

issues using the MatLab computing environment. This three-way narrative style targets a typical student of software engineering.

Pre-requisites

Intermediate level spoken English; basics of calculus including the concepts of function, derivative and the first-order optimality condition; basic linear algebra including vectors, inner products, Euclidean distances, matrices, and singular value and eigen-value decompositions; basic probability including conditional probabilities, Bayes theorem, stochastic independence, and Gaussian distribution; and basic set theory notation.

Aims

- To give a student basic knowledge and competence in modern English language and style for technical discussions of data analysis and data mining problems on the international scene
- To provide a unified framework and system for capturing numerous data analysis approaches and methods developed so far
- To teach main methods of data analysis including both bivariate and multivariate approaches including cutting edge techniques such as support vector machine, validation by bootstrapping, and evolutionary optimization techniques
- To give a hands-on experience in real-world data analysis
- To provide an experience in MATLAB coding and computation

Background and outline

The term Data Analysis has been used for quite a while, even before the advent of computer era, as an extension of mathematical statistics, starting from developments in cluster analysis and other multivariate techniques before 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. "A possible definition of data analysis is the process of computing various summaries and derived values from the given collection of data" and, moreover, the process may become more intelligent if attempts are made to automate some of the reasoning of skilled data analysts and/or to utilize approaches developed in the Artificial Intelligence areas (Berthold and Hand 2003, p. 3). 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.

The situation can be looked at as follows. 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 look at those methods that relate to an intermediate level and contribute to the theoretical – rather than any – knowledge of the phenomenon. These would 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 sciences 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.

The view of the data as a subject of computational data analysis that is adhered to here has emerged quite recently. 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. The traveler deals with what comes on their way. Helping the traveler in making sense of data is the task of data analysis. It should be pointed out that this view much differs from the conventional scientific method in which the main goal is to identify a pre-specified model of the world, and data is but a vehicle in achieving this goal. It is that view that 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.

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 is structured somewhat differently by putting 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 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.
- 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 particle 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.

- Social statistics – a discipline for measuring social and economic indexes using observation or sampling techniques.
- 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.

The course describes methods for enhancing knowledge by finding in data either

(a) Correlation among features (Cor) or

(b) Summarization of entities or features (Sum),

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, in our view. 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 of the mainstream.

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 a brief outline of those used in this course:

Mathematical structures:

- linear combination of 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. A fitting 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 indispensible part of the data analysis, which explains why this term is raised into the title.

Currently available computational methods to optimize the criterion encompass three major groups that will be touched upon here:

- 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.

Currently there is no systematic description of all possible combinations of problems, data types, mathematical structures, criteria, and fitting methods available. The course rather focuses on the generic and better explored problems in each of the four data analysis groups that can be safely claimed as being prototypical within the groups:

Summarization	**Quantitative**	**Principal component analysis**
	Categorical	**Cluster analysis**
Correlation	**Quantitative**	**Regression analysis**
	Categorical	**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 as presented in this course. They are unified in the course by the so-called data-driven modeling together with the least-squares criterion. In fact, the criterion is part of a unifying data-recovery perspective that has been developed in mathematical statistics for fitting probabilistic models and then was 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.

There can be distinguished at least three different levels of studying a computational data analysis method. A student can be interested in learning of the approach on the level of concepts only – what a concept is for, why it should be applied at all, etc. A somewhat more practically oriented tackle would be of an information system/tool that can be utilized without any knowledge beyond the structure of its input and output. A more technically oriented way would be studying the method involved and its properties. Comparable advantages (pro) and disadvantages (contra) of these three levels can be stated as follows:

	Pro	**Con**
Concepts	**Awareness**	**Superficial**
Systems	**Usable now**	**Short-term**
	Simple	**Stupid**
Techniques	**Workable**	**Technical**
	Extendable	**Boring**

Many in Computer Sciences rely on the Systems approach assuming that good methods have been developed and put in there already. Although it is largely true for well defined mathematical problems, the situation is by far different in data analysis because there are no well posed problems here – basic formulations are intuitive and rarely supported by sound theoretical results. This is why, in many aspects, intelligence of currently popular "intelligent methods" may be rather superficial potentially leading to wrong results and decisions.

One may compare the usage of an unsound data analysis method with that of getting services of an untrained medical doctor or car driver – the results can be as devastating. This is why it is important to study not only How's but What's and Why's, which are addressed in this course by focusing on Concepts and Techniques rather than Systems. Another, perhaps even more important, reason for studying concepts and techniques is the constant emergence of new data types, such as related to internet networks or medicine, that cannot be tackled by existing systems, yet the concepts and methods are readily extensible to cover them.

This course is oriented towards a student in Computer Sciences or related disciplines and reflects the author's experiences in teaching students of this type. Most of them prefer a hands-on rather than mathematical style of presentation. This is why almost all of the narrative is divided in three streams: presentation, formulation, and computation. The presentation states the problem and

approach taken to tackle it, and it illustrates the solution at some data. The formulation provides a mathematical description of the problem as well as a method or two to solve it. The computation shows how to do that computationally with basic MatLab.

This three-way narrative corresponds to the three typical roles in a successful work team in engineering. One role is of general grasp of things, a visionary. Another role is of a designer who translates the general picture into a technically sound project. Yet one more role is needed to implement the project into a product. The student can choose either role or combine two or all three of them, even if having preferences for a specific type of narrative.

The correlation problems, and their theoretical underpinnings, have been already subjects of a multitude of monographs and texts in statistics, data analysis, machine learning, data mining, and computational intelligence. In contrast, neither clustering nor principal component analysis – the main constituents of summarization efforts – has received a proper theoretical foundation; in the available books both are treated as heuristics, however useful. This text presents these two as based on a model of data, which raises a number of issues that are addressed here, including that of the theoretical structure of a summarization problem. The concept of coder-decoder is borrowed from the data processing area to draw a theoretical framework in which summarization is considered as a pair of coding/decoding activities so that the quality of the coding part is evaluated by the quality of decoding. Luckily, the theory of singular value decomposition of matrices (SVD) can be safely utilized as a framework for explaining the principal component analysis, and extension of the SVD equations to binary scoring vectors provides a base for K-Means clustering and the like. This raises an important question of mathematical proficiency the reader should have as a prerequisite. An assumed background of the student for understanding the formulation parts should include: (a) basics of calculus including the concepts of function, derivative and the first-order optimality condition; (b) basic linear algebra including vectors, inner products, Euclidean distances, matrices, and singular and eigen value decompositions; (c) basic probability; and (d) basic set theory notation. The course involves studying generic MatLab data structures and operations.

This course comes as a result of many years of the author's teaching experience in related subjects: (a) Computational Intelligence and Visualization (MSc Computer Science students in Birkbeck University of London, 2003-2010), (b) Machine Learning (MSc Computer Science students in the Informatics Department, University of Reunion, France, 2003-2004), (c) Data Analysis (BSc Applied Mathematics students in Higher School of Economics, 2008-2014), and (d) Component and Cluster Analyses of Multivariate Data (Postgraduate in Computer Science, School of Data Analysis, Yandex, Moscow, 2009-2010). These experiences have been reflected in the textbook by B. Mirkin, "Core concepts in data analysis: Summarization, Correlation and Visualization" published by Springer-London in 2011. This textbook has been favorably met by the

Computer Science community. Specifically, Computing Reviews of ACM has published a review of the book with these lines: "Core concepts in data analysis is clean and devoid of any fuzziness. The author presents his theses with a refreshing clarity seldom seen in a text of this sophistication. … To single out just one of the text's many successes: I doubt readers will ever encounter again such a detailed and excellent treatment of correlation concepts." (http://www.salereviews.com/review/review_review.cfm? review_id=139186&listname=browseissuearticle visited 27 July 2011). The course is formed of fragments from the first six Chapters of the book, which is thus singled out as the main recommended reading.

Teaching outcomes

After completion of the course, the student will know methods and their theoretical underpinnings for:

- Summarization and visualization of the one-dimensional data
- Summarization and correlation of the bivariate data, both quantitative and categorical as well as mixed
- Multivariate correlation including Linear Regression, Naïve Bayes classifier, and Classification trees
- Principal component analysis, SVD and their main applications
- K-Means clustering, including rules for its initialization
- Computational validation techniques such as bootstrapping

The student will have a computation based experience in analyzing real-world data by using generic MatLab coding.

One more outcome is

- basic knowledge and competence in modern English language and style for technical discussions of data analysis and data mining problems on the international scene.

To achieve this, the fact that, according to the author's experience, up to a 30-35% of students in the Bachelor's program in Applied Mathematics and Informatics cannot follow oral English because of their very limited knowledge, is taken into account. Therefore, each statement and notation is introduced twice, first in Russian, second in English. This may require much more time than usual lessons would which makes me to further reduce the number of topics covered in the course.

II. Schedule

No	Topic	Total hours	In class hours		Self-study
			Lectures	Labs	
Part 1					
1	What is Data Analysis	6	4	0	2
2	1D analysis: Summarization and visualization of a single feature	10	4	2	4
3	2D analysis: Correlation and visualization of two features	28	4	6	18
4	Learning multivariate correlations in data	36	4	8	24
	Part 1, in total	80	16	16	48
Part 2					
5	Principal component analysis	48	6	6	30
6	K-Means and related clustering methods	34	6	6	20
	Part 2, in total	82	16	16	50
	Total	162	32	32	98

III. Reading:

Recommended

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.
2. R.O. Duda, P.E. Hart, D.G. Stork (2001) Pattern Classification, Wiley-Interscience, ISBN 0-471-05669-3
3. H. Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.

Supplementary

4. M. Berthold, D. Hand (2003), Intelligent Data Analysis, Springer-Verlag.

5. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984) Classification and Regression Trees, Belmont, Ca: Wadswarth.

6. S.B. Green, N.J. Salkind (2003) Using SPSS for the Windows and Mackintosh: Analyzing and Understanding Data, Prentice Hall.

7. 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.

8. J. Han, M. Kamber (2010) Data Mining: Concepts and Techniques, 3d Edition, Morgan Kaufmann Publishers.

9. S. S. Haykin (1999), Neural Networks (2nd ed), Prentice Hall, ISBN 0132733501.

10. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156. (There is a Russian translation)

11. L. Lebart, A. Morineau, M. Piron (1995) Statistique Exploratoire Multidimensionelle, Dunod, Paris, ISBN 2-10-002886-3.

12. C.D. Manning, P. Raghavan, H. Schütze (2008) Introduction to Information Retrieval, Cambridge University Press.

13. R. Mazza (2009) Introduction to Information Visualization, Springer, ISBN: 978-1-84800-218-0.

14. B. Mirkin (1985) Methods for Grouping in SocioEconomic Research, Finansy I Statistika Publishers, Moscow (in Russian).

15. T.M. Mitchell (2005) Machine Learning, McGraw Hill.

16. B. Polyak (1987) Introduction to Optimization, Optimization Software, Los Angeles, ISBN: 0911575146 (Russian original, 1979).

17. B. Schölkopf, A.J. Smola (2005) Learning with Kernels, The MIT Press.

18. J. W. Tukey (1977) Exploratory Data Analysis, Addison-Wesley. (There is a Russian translation)

19. V. Vapnik (2006) Estimation of Dependences Based on Empirical Data, Springer Science + Business Media Inc., 2d edition.

20. A. Webb (2002) Statistical Pattern Recognition, Wiley and Son.

IV. Assessment

The assessment includes two main components:

(i) Coursework, a series of home assignments in

a. regression, its validation by bootstrapping and explanation of the data and results,
b. contingency table, Quetelet indexes, Pearson's chi-squared
c. tabular/piece-wise regression and correlation ratio
d. classification and explanation of the data and results,
e. cluster analysis and explanation of the data and results

on an individually chosen and approved by teaching staff dataset from Internet

(ii) Final exam, a written questions/answers in-class work

The total mark is calculated as a weighted mean of the marks for the two components according to formula: T=0.3(i)+0.7(ii).

V. Synopsis

Topic 1. What is data analysis

1.1 Data summarization

1.2 Data correlation

1.3 Data visualization

1.4 Related subjects: statistics, data mining, machine learning, information retrieval, text analysis, computational intelligence, etc.

After studying this, the student will

Know of

- The concept of an "entity-to-feature" data table;
- Main types of data analysis tasks;
- Main types of feature scales;
- Difference between data analysis/data mining and statistics;
- Introductory knowledge of related approaches

Be able to

- Prepare data tables and formulate typical problems in the analysis of them;
- Convert data into quantitative format by enveloping categorical features as sets of 1/0 features;

Have experience in

- Finding datasets, related to a substantive issue, in internet;
- Preliminarily analyzing data tables to see problems in data summarization or correlation ;
- Converting data tables into a quantitative format.

Reading

Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

Supplementary:

4. R.O. Duda, P.E. Hart, D.G. Stork (2001) Pattern Classification, Wiley-Interscience, ISBN 0-471-05669-3

5. H. Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.

Supplementary

4. M. Berthold, D. Hand (2003), Intelligent Data Analysis, Springer-Verlag.

5. 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.

6. J. Han, M. Kamber (2010) Data Mining: Concepts and Techniques, 3d Edition, Morgan Kaufmann Publishers.

7. S. S. Haykin (1999), Neural Networks (2nd ed), Prentice Hall, ISBN 0132733501.

8. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156. (There is a Russian translation)

9. C.D. Manning, P. Raghavan, H. Schütze (2008) Introduction to Information Retrieval, Cambridge University Press.

10. R. Mazza (2009) Introduction to Information Visualization, Springer, ISBN: 978-1-84800-218-0.

11. B. Mirkin (1985) Methods for Grouping in SocioEconomic Research, Finansy I Statistika Publishers, Moscow (in Russian).

12. T.M. Mitchell (2005) Machine Learning, McGraw Hill.

13. B. Schölkopf, A.J. Smola (2005) Learning with Kernels, The MIT Press.

14. J. W. Tukey (1977) Exploratory Data Analysis, Addison-Wesley. (There is a Russian translation)

15. V. Vapnik (2006) Estimation of Dependences Based on Empirical Data, Springer Science + Business Media Inc., 2d edition.

16. A. Webb (2002) Statistical Pattern Recognition, Wiley and Son.

Topic 2. 1D analysis: Summarization and Visualization of a Single Feature

2.1 Quantitative feature: Distribution and histogram

2.2 Further summarization: centers and spreads. Data analysis and probabilistic statistics perspectives. Minkowski metric center.

2.3. Case of binary and categorical features

2.4. Computational validation of the mean by bootstrapping

After studying this material, the student will

Know of

- histogram
- centrality values: mean, median, midrange, mode;

- spread values: variance, standard deviation, and mean absolute deviation;
- approximation meaning of central and spread values;
- the concept of density function and mechanisms leading to Gaussian law and power law density functions;
- the concept of bootstrap as a computational way to validate the sample based values;

Be able to

- computationally visualize histograms in different formats (bar-charts, pie-charts, etc.);
- compute of central and spread values;
- use bootstrap to compute the confidence intervals for the mean using both pivotal and non-pivotal methods;

Have experience in

- using Matlab or another computational platform to visualize histograms in different formats;
- using Matlab or another computational platform to compute central and spread values;
- using Matlab or another computational platform to compute the confidence intervals for the mean using both pivotal and non-pivotal bootstrap methods;

Reading
Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

Supplementary:

2. B. Efron and R. Tibshirani (1993) An Introduction to Bootstrap, Chapman & Hall.

3. A.P. Engelbrecht (2002) Computational Intelligence, John Wiley & Sons.

4. H. Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo.

5. B. Polyak (1987) Introduction to Optimization, Optimization Software, Los Angeles, **ISBN:** 0911575146.

6. J. Carpenter, J. Bithell (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Statistics in Medicine, 19, 1141-1164.

Topic 3. 2D analysis: Correlation and Visualization of Two Features

3.1. Two quantitative features case: Scatter-plot, linear regression, correlation coefficient Validity of the regression; bootstrapping. Non-linear and linearized regression: a nature-inspired approach

3.2 Two nominal features case: contingency tables, deriving conceptual relations, capturing relationship with Quetélet indexes, chi-squared contingency coefficient

After studying this material, the student will

Know of

- concepts related to the analysis of a pair of quantitative features: scatter plot, linear regression, correlation and determinacy coefficients as well as of their properties;
- approximation and probabilistic interpretations of the correlation coefficient;
- concepts related to the analysis of a pair of nominal features: contingency table, conditional frequency/probability, statistical independence, Quetelet indexes, chi-square coefficient and their interrelation;

Be able to

- Compute and visualize concepts related to the analysis of a pair of quantitative features: scatter plot, linear regression, correlation and determinacy coefficients;
- Compute and visualize of concepts related to the analysis of a pair of nominal scale features: contingency table, conditional frequency/probability table, Quetelet indexes, chi-square coefficient and its visualization using Quetelet indexes;

Have experience in

- using Matlab or another computational platform to compute and visualize concepts related to the analysis of a pair of quantitative features: scatter plot, linear regression, correlation and determinacy coefficients;
- using Matlab or another computational platform to compute and visualize concepts related to the analysis of a pair of nominal scale features: contingency table, conditional frequency/probability table, Quetelet indexes, and decomposition of the chi-square coefficient using Quetelet indexes.

Reading

Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

Supplementary:

2. M. Berthold, D. Hand (1999), Intelligent Data Analysis, Springer-Verlag, ISBN 3540658084.

3. A.C. Davison, D.V. Hinkley (2005) Bootstrap Methods and Their Application, Cambridge University Press (7th printing).

4. R.O. Duda, P.E. Hart, D.G. Stork (2001) Pattern Classification, Wiley-Interscience, ISBN 0-471-05669-3

5. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156.

6. H.Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.

7. B. Mirkin (2005) Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC, ISBN 1-58488-534-3.

8. T. Soukup, I. Davidson (2002) Visual Data Mining, Wiley and Son, ISBN 0-471-14999-3

9. J. Carpenter, J. Bithell (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Statistics in Medicine, 19, 1141-1164.

10. B. Mirkin (2001) Eleven ways to look at the chi-squared coefficient for contingency tables, The American Statistician, 55, no. 2, 111-120.

11. K. Pearson (1900) On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling, Philosophical Magazine, 50, 157-175..

Topic 4. Learning Multivariate Correlations in Data

4.1 General: Decision rules, fitting criteria, learning protocols, Occam's razor, metrics of accuracy

4.2 Bayes approach and Naïve Bayes classifier

4.3 Decision trees

After studying this material, the student will

Know of

- concepts involved in the Bayesian approach to data analysis: Bayes theorem, prior and posterior probabilities;
- the entity classification problem (pattern recognition) and the naïve Bayes approach to it in the case of binary features;
- the "bag of words" concept and its application to estimation of parameters of a naïve Bayesian classifier;
- concepts used for evaluation of the quality of a classifier: four-cell table, accuracy, precision, etc.;
- the concepts and main ingredients for building a classification tree or regression tree;

Be able to

- compute and apply a naïve Bayes classifier to data with binary features;
- evaluate quality values for a classifier;

Have experience in

- using Matlab or another computing environment to compute parameters of a naïve Bayes classifier and apply it to classify entities in the case of binary feature data;
- using Matlab or another computing environment to evaluate quality of a classifier.

Reading

Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

2. R.O. Duda, P.E. Hart, D.G. Stork (2001) Pattern Classification, Wiley-Interscience, ISBN 0-471-05669-3

Supplementary:

3. H. Abdi, D. Valentin, B. Edelman (1999) Neural Networks, Series: Quantitative Applications in the Social Sciences, 124, Sage Publications, London, ISBN 0 -7619-1440-4.

4. M. Berthold, D. Hand (2003), Intelligent Data Analysis, Springer-Verlag.

5. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984) Classification and Regression Trees, Belmont, Ca: Wadswarth.

6. A.C. Davison, D.V. Hinkley (2005) Bootstrap Methods and Their Application, Cambridge University Press (7^{th} printing).

7. S.B. Green, N.J. Salkind (2003) Using SPSS for the Windows and Macintosh: Analyzing and Understanding Data, Prentice Hall.

8. P.D. Grünwald (2007) The Minimum Description Length Principle, MIT Press.

9. 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.

10. J. Han, M. Kamber (2010) Data Mining: Concepts and Techniques, 3^{d} Edition, Morgan Kaufmann Publishers.

11. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156.

12. L. Lebart, A. Morineau, M. Piron (1995) Statistique Exploratoire Multidimensionelle, Dunod, Paris, ISBN 2-10-002886-3.

13. H. Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.

14. B. Mirkin (2012) Clustering: A Data Recovery Approach, Chapman & Hall/CRC, ISBN 978-1-4398-3841-9.

15. T.M. Mitchell (2010) Machine Learning, McGraw Hill.

16. J.R. Quinlan (1993) C4.5: Programs for Machine Learning, San Mateo: Morgan Kaufmann.

17. V. Vapnik (2006) Estimation of Dependences Based on Empirical Data, Springer Science + Business Media Inc., 2d edition.

18. A. Webb (2002) Statistical Pattern Recognition, Wiley and Son, ISBN-0-470-84514-7.

19. J. Carpenter, J. Bithell (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, Statistics in Medicine, 19, 1141-1164.

20. T. Fawcett (2006) An introduction to ROC analysis, Pattern Recognition Letters, 27, 861-874.

21. D. H. Fisher (1987) Knowledge acquisition via incremental conceptual clustering, Machine Learning, 2, 139–172.

22. B. Mirkin (2001) Eleven ways to look at the chi-squared coefficient for contingency tables, The American Statistician, 55, no. 2, 111-120.

Topic 5. Principal Component Analysis

5.1. Structure of a summarization problem with decoder; Data standardization; handling mixed scale data.

5.2. Singular values and triplets, associated square matrices, relations between their eigenvalues and singular values.

5.3 Principal component analysis (PCA): PCA model and method; its usage for scoring a hidden factor. Conventional formulation using covariance matrix. Relation between the conventional approach and the model-based approach.

After studying this material, the student will

Know of

- the general formula for data standardization and most popular data standardization options;
- the concept of singular value and vectors, and its relation to eigenvalues and eigenvectors of the associated square matrices;
- the Principal Component Analysis (PCA) model for scoring a hidden factor;
- SVD based and conventional methods for computing PCA scores and loadings;
- application of the PCA to measuring a hidden factor;

Be able to

- compute and interpret PCA scores and loadings;
- apply PCA to a dataset for computing a hidden factor;

Have experience in

- computation of PCA scores and loadings;
- application of PCA to computing a hidden factor.

Reading

Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

Supplementary:

2. 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.

3. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156.

4. L. Lebart, A. Morineau, M. Piron (1995) Statistique Exploratoire Multidimensionelle, Dunod, Paris, ISBN 2-10-002886-3.

5. C.D. Manning, P. Raghavan, H. Schütze (2008) Introduction to Information Retrieval, Cambridge University Press.

6. B. Mirkin (2012) Clustering: A Data Recovery Approach, Chapman & Hall/CRC, ISBN 978-1-4398-3841-9.

7. R. Cangelosi, A. Goriely (2007) Component retention in principal component analysis with application to cDNA microarray data, Biology Direct, 2:2, http://www.biolgy-direct.com/content/2/1/2.

8. S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman (1990) Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41 (6), 391-407.

Topic 6. K-Means and Related Clustering Methods

6.1 Clustering criterion and its reformulations. K-Means clustering as alternating minimization; Partition around medoids PAM; Choosing the number of clusters; Initialization of K-Means; Anomalous pattern and Intelligent K-Means

6.2. Fuzzy clustering.

6.3 Cluster interpretation aids

After studying this material, the student will

Know of

- k-means clustering method метод*;*
- criterion of k-means clustering method and interpretation of the method as an alternating minimization scheme for the criterion;
- advantages and shortcomings of k-means method;
- ways to interpret the results of k-means method applied to a dataset;
- anomalous cluster method;
- using anomalous clustering to specify the number and initial location of k-means cluster centers; использование метода аномального кластера для выбора числа кластеров и их начальных центров;
- c-means clustering method;

Be able to

- apply k-means method to cluster a data table;
- initialize k-means by using the anomalous cluster method;
- interpret the results of a run of k-means method;

Have experience in

- using Matlab or another computing environment to apply k-means method for data clustering;

Reading

Recommended:

1. B. Mirkin (2011) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London.

2. A.K. Jain and R.C. Dubes (1988) Algorithms for Clustering Data, Prentice Hall.

Supplementary:

3. J. Bezdek, J. Keller, R. Krisnapuram, M. Pal (1999) Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers.

4. S.B. Green, N.J. Salkind (2003) Using SPSS for the Windows and Mackintosh: Analyzing and Understanding Data, Prentice Hall.

5. J.A. Hartigan (1975) Clustering Algorithms, Wiley and Sons.

6. L. Kaufman and P. Rousseeuw (1990) Finding Groups in Data: An Introduction to Cluster Analysis, Wiley and Sons.

7. M.G. Kendall, A. Stewart (1973) Advanced Statistics: Inference and Relationship (3d edition), Griffin: London, ISBN: 0852642156.

8. A. Kryshtanowski (2008) Analysis of Sociology Data with SPSS, Higher School of Economics Publishers, Moscow (in Russian).

9. H.Lohninger (1999) Teach Me Data Analysis, Springer-Verlag, Berlin-New York-Tokyo, 1999. ISBN 3-540-14743-8.

10. B. Mirkin (2012) Clustering: A Data Recovery Approach, Chapman & Hall/CRC, ISBN 978-1-4398-3841-9.

11. S. Bandyopadhyay, U. Maulik (2002) An evolutionary technique based on K-means algorithm for optimal clustering in R^N, Information Sciences, 146, 221-237.

12. R. Cangelosi, A. Goriely (2007) Component retention in principal component analysis with application to cDNA microarray data, Biology Direct, 2:2, http://www.biolgy-direct.com/content/2/1/2.

13. J. Kettenring (2006) The practice of cluster analysis, Journal of Classification, 23, 3-30.

14. Y. Lu, S. Lu, F. Fotouhi, Y. Deng, S.Brown (2004) Incremental genetic algorithm and its application in gene expression data analysis, BMC Bioinformatics, 5,172.

15. M. Ming-Tso Chiang, B. Mirkin (2010) Intelligent choice of the number of clusters in K-Means clustering: an experimental study with different cluster spreads, Journal of Classification, 27(1), 3-40.

VI. Final exam questions

Here is a set of examples:

1. What is a histogram of a feature? How can one build a histogram? What is the relation between a histogram and the feature distribution?
2. What is the range of a feature?
3. How can one validate the sample based mean value using bootstrapping?
4. What can you say of the shape of a one-mode feature distribution if its median coincides with its mean? Or, if the median is much smaller than the mean?

5. Occurrence/co-occurrence table

5.1. Of 200 Easter shoppers, 100 spent £100 each, 20 spent £50 each, and 80 spent £200 each. What are the (i) average, (ii) median and (iii) modal spending? Explain. Tip: How can one take into account in the calculation that there are, effectively, only three different types of customers?

5.2. Among the shoppers, those who spent £50 each are males only and those who spent £200 each are females only, whereas among the rest 100 individuals 40% are men and the rest are women. Build a contingency table for the two features, gender and spending.

5.3. Find the Quetelet coefficient for males who spent £50 each and explain its meaning.

6. K-Means clustering.

Consider a specified data table of 8 entities (i1, i2, …, i8) and 2 features (v1, v2).

6.1. Standardize the data with the feature averages and ranges; perform further actions over the standardized data. Would K-Means result differ for this data if the normalization by the range is not performed (yes or no, and why)?

6.2. Set K=2 and initial seeds of two clusters so that they should be as far from each other as possible. Assign entities to the seeds with the Minimum distance rule.

6.3. Calculate the centroids of the found clusters; compare them with the initial seeds.

6.4. Is there any chance that the found clusters are final in the K-means process?

7. Model for the method of Principal Component Analysis and its relation to the problem of the maximum singular value for the data matrix Y.

8. Eigenvalues of YY^T and Y^TY. Contribution of the principal component to the data scatter.

9. Conventional formulation of the PCA method.

10. Linear regression: Let the correlation coefficient between features x and y be 0.8. Can you tell anything of the proportion of the variance of y taken into account by its linear regression over x?

The syllabus is prepared by Boris Mirkin