Правительство Российской Федерации

Государственное образовательное бюджетное учреждение высшего профессионального образования

«Национальный исследовательский университет Высшая школа экономики»

Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

Программа дисциплины

≪Современные методы анализа данных≫

Modern Methods for Data Analysis

(на английском языке)

для направления 01.03.02 – Прикладная математика и информатика
подготовки магистра

Автор программы
Профессор д.т.н. Б.Г. Миркин bmirkin@hse.ru

Одобрена на заседании
департамента Анализа данных и
искусственного интеллекта

Руководитель департамента

_____С.О.Кузнецов
« ____» _____ 2017 г.

Москва

2017

Modern Methods for Data Analysis

A Syllabus

1. Instructor and author

Boris Mirkin, PhD (Mathematics), DSc (Engineering), Professor of the Department of Data Analysis and Artificial Intelligence FCS NRU HSE Moscow, Emeritus Professor of the University of London

2. Reference to regulatory document

This syllabus is prepared according to the Teaching standard "ОБРАЗОВАТЕЛЬНЫЙ СТАНДАРТ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО АВТОНОМНОГО ОБРАЗОВАТЕЛЬНОГО УЧРЕЖДЕНИЯ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО УНИВЕРСИТЕТА «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ» по направлению подготовки 01.03.02 Прикладная математика и информатика (accepted at the Meeting of Learned Council of NRU HSE 03.03.2017 (minutes № 02)).

3. Summary

This is a course in basic methods for 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. This view distinguishes the subject from related courses such as applied statistics, machine learning, data mining, etc. Two main pathways for data analysis are: (1) summarization, for developing and augmenting concepts, and (2) 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.

Another feature of the class 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 mathematically derived 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 an illustrative real-world dataset – this part can be read and studied with no concern for the mathematical formulation at all. There is one more part, “computation”, targeted at studying the computational data processing issues using the MatLab or any other computing environment: the codes here can be considered as pseudo-codes, just a way for presenting algorithms. This three-way narrative style targets a typical student of software engineering and programming.

4. Pre-requisites

- Spoken English (intermediate level);
- 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 eigenvalue decompositions;
- Basic probability including conditional probabilities, Bayes theorem, stochastic independence, and Gaussian distribution; and
- Basic set theory notation.

5. Objectives:

- 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 modern methods of data analysis including cutting edge techniques such as intelligent and spectral clustering, community detection, SVD and principal component analysis, validation by bootstrapping, and evolutionary optimization techniques
- To give a hands-on experience in real-world data analysis
- To provide an experience in using modern computational tools and computation

6. Place of the course in the structure of the educational program

This course is an intermediary in the educational program. It is taught after students have learned some English, as well as basic concepts and methods of mathematics, statistics, informatics, and coding. 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. Here its scope is limited to techniques helping the user to enhance theoretical knowledge by developing new concepts or relations between concepts. The former is covered by such methods as the Principal component analysis (combining quantitative features in a new quantitative feature) and Cluster analysis (producing a new nominal feature). The latter, by Regression analysis and Bayes classifiers. An extensive training in using these techniques After this course students go on to study more advanced methods in machine learning, knowledge discovery, data processing, distributed systems or systems analysis.

7. Guidelines for students and teacher:

To position the course, one can consider this. 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) so far concentrated mostly on the problem of learning categories of objects, thus leaving such important problems as quantitative summarization outside of the mainstream.

A teacher, and student alike, should be aware that 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 indispensable part of the data analysis, which explains why such attention is given to visualization in this course.

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** **Categorical**	**Principal component analysis** **Cluster analysis**
Correlation	**Quantitative** **Categorical**	**Regression analysis** **Classifier**

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.

8. Education technologies:

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. The second edition of the book is in preparation.

The described lecturing system is supported by Homework which comprises several tasks including search and preprocessing of a data table as well as application of studied methods to the data, that are to be done independently by each of the project teams. Usually a team is composed by one or two students. The laboratory hours are devoted to consulting students with regard to the homework tasks over their specific dataset. Typical individual issues are discussed and explained in class. The final output is a Word-editor file of a project report to be graded by the instructor or teaching assistant.

9. Student's competences after the course:

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

- Mixed scales data, quantification, pre-processing, standardization
- K-Means clustering, including rules for its initialization and interpretation
- Comparing cluster means with computational validation techniques such as bootstrapping
- Interpreting clusters in nominal scales, Quetelet indexes and Pearson's Chi-squared
- Hierarchical clustering
- Clustering similarity and network data including community detection, spectral clustering, and consensus clustering
- Principal component analysis (PCA), SVD and data visualization

- Matrices of covariance and correlation indexes; conventional formulation for PCA
- Correlation and determinacy indexes at different parspectives

The student will have a computation based experience in analyzing real-world data by using generic MatLab coding or other computing environment. These relate to items УК-2, УК-3 of the Teaching standard (Ability to explore the essence of issues and solve emerging problems by using mathematics), as well as those of ПК1, ПК3, ПК7, ПК8, ПК9 of that (Ability to formalize issues and apply mathematics concepts and methods to them; ability to collect data and apply methods of machine learning and data analysis; ability to use existing codes or develop a computational code for solving occurring problems).

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.

10. Contents and schedule

No	Topic	Total hours	In class hours		Self-study
			Lectures	Labs	
Part 1					
1	K-Means clustering: method and properties	11	3	2	6
2	Data preprocessing; scale types; quantification	6	2	2	2
3	Cluster interpretation: comparison of means, bootstrap for confidence intervals	13	3	2	8
4	Cluster interpretation at mixed scale features, Correlation ratio, Pearson chi-squared, Quetelet indexes	14	2	4	8
5	Clustering similarity and network data; k-means converted criterion and algorithms	24	4	4	16
6	Consensus clustering; two criteria; reduction to network clustering	10	2	2	6
	Part 1, in total	78	16	16	46

Part 2					
7	Principal component analysis (PCA), Singular value decomposition (SVD), using PCA for data visualization	28	4	10	14
8	PCA: covariance and correlation matrices, meaning and properties of correlation coefficient in three perspectives; conventional formulation of PCA	32	3	9	20
9	Spectral clustering	14	2	4	8
	Part 2, in total	74	9	23	42
	Total	152	25	39	88

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

4. L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984) Classification and Regression Trees, Belmont, Ca: Wadswarth.0

5. S. Das (2014) Computational Business Analytics, CRC Press.

6. S.B. Green, N.J. Salkind (2010) Using SPSS for the Windows and Mackintosh: Analyzing and Understanding Data, Prentice Hall, 6th Edition.

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

8. 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.

9. J. Han, J. Pei, M. Kamber (2011) Data Mining: Concepts and Techniques, 3d Edition, Morgan Kaufmann Publishers.

10. S. S. Haykin (2008), Neural Networks and Learning Machines, Prentice Hall.

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

12. Daniel T. Larose and Chantal D. Larose (2014) Discovering Knowledge in Data: An Introduction to Data Mining, J. Wiley and Sons (2d edition)

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

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

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

16. W. McKinney (2014) Python for Data Analysis, O'Reilly.

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

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

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

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

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

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

23. V. Vapnik (2013) The Nature of Statistical Learning Theory, Springer.

24. A. Webb, K. Copsey (2011) Statistical Pattern Recognition, Wiley and Sons.

25. Ian H. Witten,Eibe Frank,Mark A. Hall,Christopher J. Pal (2017) Data Mining: Practical Machine Learning Tools and Techniques, Elsevier (4th Edition)

26. Zimmermann, H. J. (2011). *Fuzzy set theory—and its applications*. Springer Science & Business Media.

Articles:

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

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

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

30. 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.

12. Learning aids:

- Textbooks by Mirkin (2011, in English; 2014-17, in Russian) and other texts listed above;
- PowerPoint Lecture slides prepared by the instructor to contain all the important formulas, models and methods;
- Consultations by the Instructor over Homework and related issues;
- Set of questions, in the end, from which exam paper questions are taken.

13. Forms of evaluation for the current assessment and attestation

The current assessment is conducted by checking, commenting, and grading the Homework reports of individual projects. Each report comprises a set of 6-7 sections, each describing a task and its solution, which is graded. Specifically, the sections usually include:

a. validation and comparison of means by bootstrapping
b. contingency table, Quetelet indexes, Pearson's chi-squared
c. tabular/piece-wise regression and correlation ratio
d. data visualization by using SVD
e. cluster analysis and interpretation of results

over an individually chosen and approved by teaching staff dataset from Internet.

The average mark over the set is the current assessment mark.

14. Forms of knowledge assessment and grading procedures

Knowledge assessment is conducted as a two-step controlling procedure: 1) Homework and 2) Exam paper. Exam paper comprises 6-7 questions, such as examples in section 15 below, diversified so that each student is to write an individual Exam paper. Each question is assigned with a maximum mark that can be obtained by answering the question, so that the total mark is 100%. At the exam, every student receives a randomly chosen Exam paper. The exam lasts about 100-120 minutes, after which all papers are collected and marked. Each question is marked according to the extent of the answer written on the paper. Then the total is calculated and proportionally converted to the required scale, usually of 10 points. Rounding of the totals such as 53 or 68 may involve some intricacies. Usually students find it fair if, say, mark 53 is converted into 5, 68 into 7, but 64, 65 and 66 are converted into either 6 or 7 depending on the student's diligence.

The **final mark** is computed according to formula **FM=0.4*HW+0.6*EX**, where HW is Homework mark and EX is Exam paper mark.

15. Exam paper 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