

**Oleksandr Petrov**  
**Oleksandr Shumeyko**  
**Beata Basiura**  
**Anton Petrov**

# **INTRODUCTION** **TO DATA MINING**



WYDAWNICTWA AGH  
KRAKÓW 2019

Published by AGH University of Science and Technology Press

© Wydawnictwa AGH, Kraków 2019

ISBN 978-83-66364-25-7

Editor-in-Chief: *Jan Sas*

Editorial Committee:

*Andrzej Pach* (Chairman)

*Jan Chłopek*

*Barbara Gąciarz*

*Bogdan Sapiński*

*Stanisław Stryczek*

*Tadeusz Telejko*

Reviewers:

*prof. dr hab. inż. Mikołaj Karpiński*

*dr hab. inż. Grzegorz Ginda*

Authors' affiliation:

*Oleksandr Petrov – AGH Akademia Górniczo-Hutnicza w Krakowie*

*Oleksandr Shumeyko – Dniprovsk State Technical University*

*Beata Basiura – AGH Akademia Górniczo-Hutnicza w Krakowie*

*Anton Petrov – Federal State Budgetary Educational Institution of Higher Education „Kuban State Agrarian University named after I.T. Trubilin”*

Technical editor: *Kamila Zimnicka*

Desktop Publishing: *Andre*

Cover Design: *Paweł Sepielak*

---

AGH University of Science and Technology Press (Wydawnictwa AGH)

al. A. Mickiewicza 30, 30-059 Kraków

tel. 12 617 32 28, tel./fax 12 636 40 38

e-mail: [redakcja@wydawnictwoagh.pl](mailto:redakcja@wydawnictwoagh.pl)

<http://www.wydawnictwo.agh.edu.pl>

---

# Contents

<b>Introduction</b> .....	7
<b>Preface</b> .....	9
<b>1. Least-squares method</b> .....	11
1.1. Ordinary least square method .....	11
1.2. Linearization at the least squares method .....	17
1.3. Examples in Python .....	23
<b>2. Principle Component Analysis</b> .....	32
2.1. The main idea of PCA .....	32
2.2. An iteration scheme of PCA calculating .....	40
2.3. Examples in Python .....	44
2.4. Optimum transition from the RGB model to optimum three-component model .....	48
2.5. Fisher linear discriminant analysis .....	51
2.6. Multidimensional discriminant analysis (MDA) .....	57
<b>3. Application of fuzzy logic in Data Mining</b> .....	72
3.1. What is fuzzy thinking? .....	72
3.2. Fuzzy sets .....	73
3.3. Linguistic variables and linguistic gain .....	77
3.4. Operations on fuzzy sets .....	80
3.5. Properties of operations on fuzzy sets .....	83
3.6. Fuzzy inference rules .....	86
3.7. Defuzzification .....	96
3.8. The choice of alternatives using fuzzy inference rules .....	100
3.9. Ranking alternatives based on heuristic approach .....	110
3.10. Fuzzy decision trees .....	117

<b>4. Soft computing in data handling</b> .....	129
4.1. Introduction to soft computing .....	129
4.2. Evolutionary calculations .....	131
4.2.1. General Introduction .....	131
4.2.2. Genetic algorithm .....	133
4.2.3. Simple example of implementation of GA .....	136
4.2.4. Closer to reality, or the space crossover .....	140
4.2.5. Genetic programming .....	149
4.2.6. To be, or not to be... .....	152
4.2.7. Diophantine equation .....	159
4.3. Swarm intelligence .....	162
4.3.1. The use of ant algorithm for the Traveling Salesman Problem .....	167
<b>5. Clustering methods</b> .....	174
5.1. Clustering. General concepts .....	175
5.2. Hierarchical methods .....	178
5.2.1. Hierarchical methods. Agglomerative algorithms .....	178
5.2.2. Hierarchical methods. Divisive algorithms .....	179
5.3. Examples in Python – clustering hierarchical methods .....	180
5.4. Nonhierarchical algorithms .....	216
5.4.1. <i>K</i> -means method .....	217
5.4.2. Fuzzy <i>k</i> -means .....	220
5.4.3. Gyustafsona-Kessel’s clustering .....	221
5.4.4. Method of correlation galaxies .....	223
5.4.5. Spectral clustering method .....	224
5.5. Examples in Python – clustering nonhierarchical methods .....	228
<b>6. Classifiers</b> .....	248
6.1. Definition of the classification problem .....	248
6.2. Main directions of the research of the classification issue .....	249
6.3. Stochastic classifiers .....	251
6.3.1. Use of the theorem of Bayes for decision-making .....	251
6.4. Naive Bayesian classifier .....	255
6.4.1. Example of sale of the Naive Bayes classifier .....	258
6.4.2. EM algorithm .....	263
6.5. Linear discriminant analysis .....	271
6.5.1. Example 5.1 .....	274
6.5.2. Example 5.2 .....	276
6.5.3. Example 5.3 .....	277

<b>7. Use of genetic algorithms for creation of the vector classifiers .....</b>	<b>279</b>
7.1. Use of Veronoi polyhedron	
in the problem of texts classification .....	282
7.2. Check of the existing classification on the correctness .....	284
<b>8. Support vector machines .....</b>	<b>287</b>
8.1. The main idea .....	287
8.2. SVM for linear separable set .....	289
8.3. SVM for nonlinear separable set .....	290
8.4. Example .....	294
<b>9. Visualization of multidimensional data .....</b>	<b>297</b>
9.1. Multidimensional scaling technic .....	297
9.2. Kohonen self-organizing maps (SOM).....	299
9.2.1. Initialization of the map of Kohonen .....	301
9.2.2. The training algorithm .....	302
9.3. Examples .....	305
9.3.1. Showing similarity of objects .....	305
9.3.2. Showing similarity of European countries .....	306
9.3.3. The world map of poverty .....	308
9.3.4. The traveling salesman problem .....	309
<b>10. Recommender systems .....</b>	<b>311</b>
10.1. General structure of recommendation system .....	313
10.1.1. Collaborative filtering .....	314
10.1.2. The content-oriented recommendations .....	321
10.1.3. Profiles of users .....	322
10.1.4. Training a user model .....	323
10.1.5. Hybrid approaches .....	328
10.2. Analysis of client environments .....	330
10.2.1. Examples of client environments .....	330
10.2.2. Retail chain stores .....	331
10.2.3. Mobile operators .....	331
10.2.4. Online stores of books, audio and video of other products .....	331
10.2.5. Search engines .....	332
10.2.6. Parliamentary elections .....	332
10.2.7. Analysis of texts .....	333
10.2.8. Social networks .....	333

<b>Appendix</b>	
<b>Basic information</b> .....	334
A.1. Background information on linear algebra .....	334
A.2. Background information on probability theory .....	336
<b>References</b> .....	347