



NOVA
IMS
Information
Management
School


Data Mining

S1

NOVA-IMS 2025/2026
Fernando Lucas Bação
bacao@isegi.unl.pt
<http://www.isegi.unl.pt/fbacao>

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1



Agenda

L 1	8 Sep	<p>Introduction to the Data Mining</p> <ul style="list-style-type: none"> Course Syllabus Objectives Course projects Grading Bibliography <p>2nd half:</p> <ul style="list-style-type: none"> Practical Sessions Instructions (Setup environment) Project Description (w/ Farina Conteios) <p>NO Practical sessions (Python)</p>
-----	-------	--

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2

NOVA
IMS
Information Management School

Syllabus

NOVA
IMS
Information Management School

DATA MINING

SYLLABUS
2025 - 2026

INSTRUCTOR INFORMATION	FERNANDO LUCAS BAÇÃO 2 nd floor, room 10 Tel: 21 3870413 (ext. 222) fbacao@novaims.unl.pt http://www.novaims.unl.pt/fbacao GASPAREIRA PEREIRA gppereira@novaims.unl.pt ANA CALEIRO acaleiro@novaims.unl.pt
SCHEDULE	TP1 Theoretical Sessions • Monday 11h30 – 13h00 TP2 Theoretical Sessions • Monday 13h30 – 15h00 Practical Sessions (Ana Caleiro) • P1 – Monday 13h30 – 15h00 • P5 – Monday 10h00 – 11h30 • P6 – Monday 15h00 – 16h30 Practical Sessions (Gaspar Pereira) • P2 – Monday 15h00 – 16h30 • P3 – Wednesday 11h30 – 13h00 • P4 – Monday 16h30 – 18h00 • P7 – Monday 11h30 – 13h00
OFFICE HOURS:	Monday from 15h00 – 16h00 (schedule appointment by email) 2nd Floor, Room 10
CONTACT	All communications with the instructors should be done using the Moodle

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3

NOVA
IMS
Information Management School

Syllabus

NOVA
IMS
Information Management School

Data Mining I

```

graph LR
    DM[Data Mining I] --- 1[1. Introduction to the Data Mining Course]
    DM --- 2[2. Introduction to Data Science]
    DM --- 3[3. The canonical tasks in Data Mining and work process]
    DM --- 4[4. Exploratory Data Analysis]
    DM --- 5[5. Data Preparation and Preprocessing]
    DM --- 6[6. Data Segmentation Strategies]
    DM --- 7[7. Association rules]
    DM --- 8[8. Data Clustering]
    DM --- 9[9. Multidimensional Visualization Methods]
    DM --- 10[10. Semi-supervised classification]
  
```

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

NOVA
IMS
Information
Management
School

Course Success

COURSE SUCCESS

In this course success depends on a number of factors:

- Basic knowledge of statistics;
- Attend classes;
- Work during the semester and not only when exams are about to start;
- Develop the course project during the semester, making the most of the practical classes;
- Read the suggested references and the PowerPoint slides made available by the lecturer.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

5

NOVA
IMS
Information
Management
School

Project

Course Projects

Project consists of a practical clustering application using Python. In this project the students will complete the segmentation of a customer database, following all the usual steps of a real-world project. For this the students will receive a set of specific guidelines that they should follow. The guidelines provide information about the type of tasks the students should do and the general results they should achieve.

The end product of the project should be a report about the database and the different customer segments of the company. With this project the students should develop their analytical skills, but also their proficiency in working with large datasets, extracting, transforming and loading tasks and visualization and reporting.

The project will be evaluated based on three components. All components are mandatory:

- **Deliverable 1:** Exploratory Data Analysis (30%). Students should conduct an in-depth EDA of the dataset.
- **Deliverable 2:** Final Report (60%). Students should perform a customer segmentation using the dataset provided.
- **Project discussion:** (10%). After submitting the project the students will be called to discuss the project with one of the instructors. Each student will receive an individual grade based on their contribution to the discussion.

Project groups. The project can be done individually or in groups (the latter is a better option). The recommended group size is 3 and must not exceed 4 students.

Project Deadlines:

- November 4th (Deliverable 1)
- January 3rd (Deliverable 2)

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

6

NOVA

IMS

Information
Management
School

Grading

Grading

Project: 45%

Exam: 55%

Both components of the evaluation (project and exam) are mandatory. There are two opportunities to do the exam. Any delay in the delivery of the project is subject to a penalty of 10% of the grade for each day of delay. Please note that the project will be developed in groups, but each group cannot have more than 4 elements. To obtain approval in the discipline the student **cannot have less than 8 (40%) in the exam grade.**

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

7

NOVA

IMS

Information
Management
School

References

Pattern Recognition Letters 31 (2010) 631–636

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Data clustering: 50 years beyond K-means[☆]

Anil K. Jain^{*}

Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
Department of Post and Computer Engineering, Korea University, Ansan-si, Seoul, 130-702, Korea

ARTICLE INFO

Article history:
Available online 9 September 2009

Keywords:
Data clustering
User clusters
Hierarchical clustering
Regression on clustering
King-Sun Ho prize

ABSTRACT

Clustering data into sensible groupings is one of the most fundamental needs of understanding and learning. As an example, a common scheme of scientific classification puts organisms into a system of ranked taxa: domains, kingdoms, phyla, etc. Cluster analysis is the formal study of methods and algorithms for grouping or clustering objects according to measured or perceived intrinsic characteristics or attributes. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels. The absence of category information distinguishes data clustering (unsupervised learning) from classification or discriminative analysis (supervised learning). The aims of clustering is to find structure in data and is therefore exploratory in nature. Clustering has a long and rich history in a variety of scientific fields. One of the most popular and simple clustering algorithms, K-means, was first published in 1955. In spite of the fact that K-means was proposed over 50 years ago and thousands of clustering algorithms have been published since then, K-means is still widely used. This speaks to the difficulty in designing a general purpose clustering algorithm and the ill-posed problem of clustering. We provide a brief overview of learning, summarize well-known clustering methods, discuss the major challenges and key issues in designing clustering algorithms, and point out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering.

© 2009 Elsevier B.V. All rights reserved.

Data Clustering: A Review

A.K. Jain
Michigan State University
and
M.N. Murty
Indian Institute of Science
and
P.J. Flynn
The Ohio State University

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. This paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We present a taxonomy of clustering techniques and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

Categories and Subject Descriptors: I.5 (Computing Methodologies): Pattern Recognition—Models; I.5.3 (Pattern Recognition): Clustering; I.5.4 (Pattern Recognition): Applications—Computer Vision; H.3.3 (Information Storage and Retrieval): Information Search and Retrieval—Clustering; I.2.6 (Artificial Intelligence): Learning—Knowledge Acquisition

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

8

4

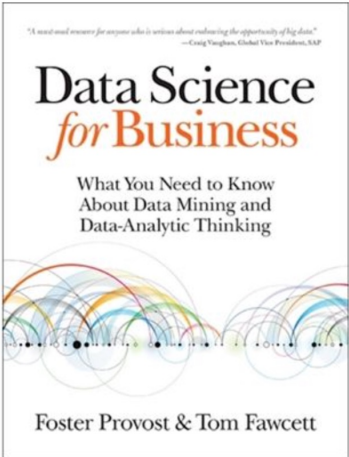
NOVA
IMS
Information Management School

References

"A must-read resource for anyone who is serious about embracing the opportunity of big data."
—Craig Vaughan, Global Vice President, SAP

Data Science for Business

What You Need to Know
About Data Mining and
Data-Analytic Thinking



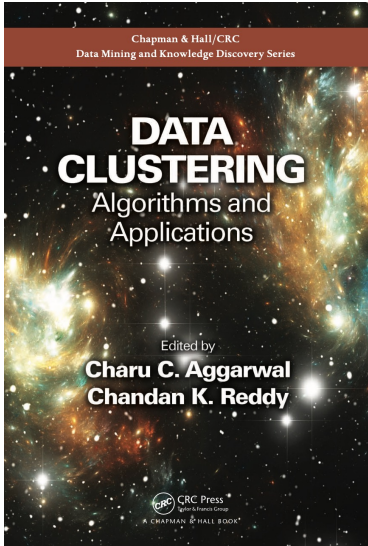
Foster Provost & Tom Fawcett

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

DATA CLUSTERING

Algorithms and
Applications

Edited by
Charu C. Aggarwal
Chandan K. Reddy



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

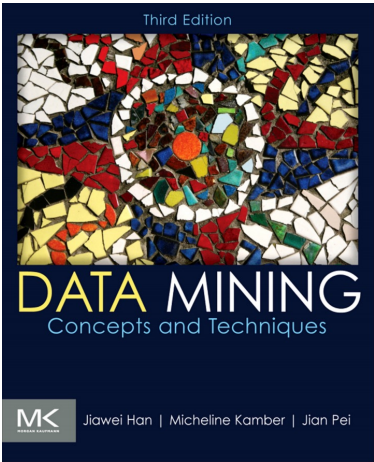
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

9

NOVA
IMS
Information Management School

References

Third Edition



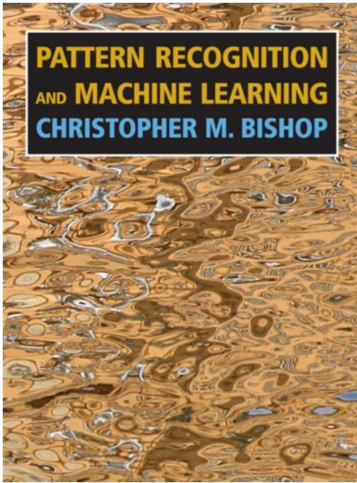
DATA MINING

Concepts and Techniques

MK Jiawei Han | Micheline Kamber | Jian Pei

PATTERN RECOGNITION AND MACHINE LEARNING

CHRISTOPHER M. BISHOP



Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

10

NOVA
IMS
Information
Management
School

Project

- Practical Sessions Instructions (Setup environment)
- Project Description (w/ Farina Pontejos)

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

11

Questions?

12

12