Preface

Every day, enormous amounts of information are generated from all sectors, whether it is business, education, the scientific community, the World Wide Web (WWW), or one of many readily available off-line and online data sources. From all of this, which represents a sizable repository of data and information, it is possible to generate worthwhile and usable knowledge. As a result, the field of Data Mining (DM) and knowledge discovery in databases (KDD) has grown in leaps and bounds and has shown great potential for the future.

Data mining is a multidisciplinary field, drawing work from areas including database technology, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, knowledge based systems, knowledge acquisition, information retrieval, high performance computing, and data visualization. We present the material in this book from an *artificial learning from data perspective*. In a typical scenario, we have an outcome measurement, usually quantitative (like a stock price) or categorical (like heart attack/no heart attack), that we wish to predict based on a set *of features* (like diet and clinical measurements). We have a *training set* of data, in which we observe the outcome and feature measurements for a set of objects (such as people). Using this data we build a prediction model, or *learner*, which will enable us to predict the outcome for new unseen objects. A good learner is one that accurately predicts such an outcome.

The examples above describe what is called the *supervised learning* problem. It is called "supervised" because of the presence of the outcome variable to guide the learning process. In the *unsupervised learning problem* we observe only the features and have no measurements of the outcome. Our task is rather to describe how the data are organized or clustered.

My purpose in writing this lecture notes has been to give a systematic introduction of major concepts and methodologies of artificial unsupervised learning from data and to present a unified framework that makes the subject more accessible to students.

The main audience was the students following the Doctorate courses at the Department of Statistical Sciences, University of Padua and the terminal-year students in informatics at the Faculty of Mathematics and Computer Science, University of Bucharest.

The background material needed to understand these notes is general knowledge of some basic topics in probability and statistics, differential equations and linear algebra, and multivariate calculus.

This course is organized in eight chapters. Chapter 1 introduces the reader to the key concepts of knowledge discovery and unsupervised statistical learning. The general model of learning from observations is studied and the current trends in machine learning and knowledge discovery are presented.

Chapter 2 covers semi-parametric and nonparametric techniques for **probability density modeling**. By these, we mean techniques where we make few or no assumptions about what functional form the probability density takes. We present three main methods of semi-parametric and nonparametric density estimation and their variants: histograms, kernel density estimates, and finite mixtures.

Chapter 3 deals with **cluster analysis**. We discuss some simple and essentially model-free methods for classification and pattern recognition based on *K*-means, *K*-medoids, vector quantization and several other tools.

Unsupervised learning rules for single-unit and single-layer nets are covered in chapter 4. More than seven basic discrete-time learning rules are presented (reinforcement, Hebbian, competitive and feature mapping rules). The presentation of these learning rules is unified in the sense that they may all be viewed as realizing incremental steepest-gradient-descent search on a suitable criterion function.

Chapter 5 is concerned with the **theoretical aspects of unsupervised learning** in artificial neural networks. It investigates mathematically the nature and stability of the asymptotic solutions obtained using the basic Hebbian and reinforcement learning rules, which were introduced in the preceding chapter. Formal analysis is also given for simple competitive learning and self-organizing feature-map learning.

Chapter 6 addresses the issue of **unsupervised learning in multilayer nets** and describes two specific networks (adaptive resonance theory networks and autoassociative clustering network) suitable for adaptive data clustering.

Chapter 7 collects together a number of **data visualization** techniques both linear and non-linear: Probabilistic Principal Component Analysis, a latent variable formulation of Principal Component Analysis that provides a density model, generative topographic Mapping, a non-linear latent variable model and neuroscale, a non-linear topographic (i.e. distance preserving) projection.

I hope that this lectures notes will prove useful to those students who are interested not only in understanding the underlying theory of machine learning from data but also in pursuing researching this area. A list of relevant references is included with the aim of providing guidance and direction for the reader's own search of the research literature.

Acknowledgement. My deep gratitude goes to Professor Fortunato Pesarin of the Department of Statistical Sciences, University of Padua who invited me to teach this course in the autumn term of the university year 2004-05.

I am grateful also to professor Luigi Salmaso for the useful discussions and help.

The Author