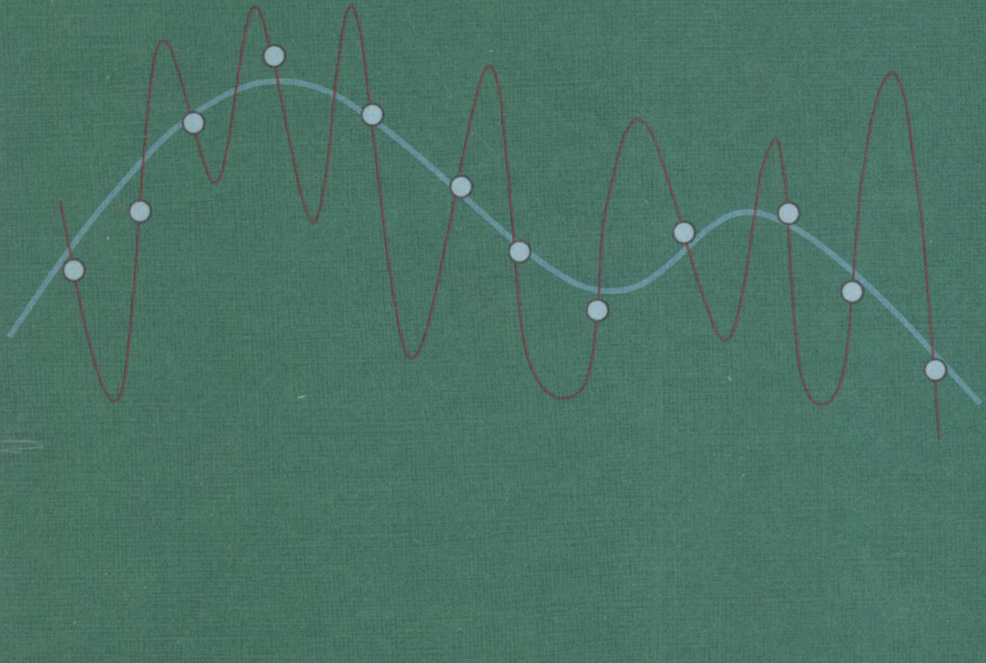
Environmental Sciences Environmental Engineering

MACHINE LEARNING FOR SPATIAL ENVIRONMENTAL DATA

THEORY, APPLICATIONS AND SOFTWARE

Mikhail Kanevski, Alexei Pozdnoukhov and Vadim Timonin



EPFL Press

ENVIRONMENTAL SCIENCES *Environmental Engineering*

MACHINE LEARNING FOR SPATIAL ENVIRONMENTAL DATA

THEORY, APPLICATIONS AND SOFTWARE

Mikhail Kanevski, Alexei Pozdnoukhov and Vadim Timonin

EPFL Press
A Swiss academic publisher distributed by CRC Press

CRC Press
Taylor & Francis Group

Taylor and Francis Group, LLC
6000 Broken Sound Parkway, NW, Suite 300,
Boca Raton, FL 33487

Distribution and Customer Service
orders@crcpress.com

www.crcpress.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

This book is published under the editorial direction of
Professor Christof Holliger (EPFL).

Already published by the same author:
Analysis and Modelling of Spatial Environmental Data,
M. Maignan and M. Kanevski, EPFL Press, 2004.

EPFL Press

is an imprint owned by Presses polytechniques et universitaires romandes, a Swiss academic publishing company whose main purpose is to publish the teaching and research works of the Ecole polytechnique fédérale de Lausanne (EPFL).

Presses polytechniques et universitaires romandes
EPFL – Centre Midi
Post office box 119
CH-1015 Lausanne, Switzerland
E-Mail: ppur@epfl.ch
Phone: 021/693 21 30
Fax: 021/693 40 27

www.epflpress.org

ISBN 978-2-940222-24-7 (EPFL Press)
ISBN 978-0-8493-8237-6 (CRC Press)

Printed in Spain

MACHINE LEARNING FOR SPATIAL ENVIRONMENTAL DATA

This book is dedicated to our families and friends.

PREFACE

The book is devoted to the analysis, modelling and visualisation of spatial environmental data using machine learning algorithms. In a broad sense, machine learning can be considered a subfield of artificial intelligence; the subject is mainly concerned with the development of techniques and algorithms that allow computers to learn from data. In this book, machine learning algorithms are adapted for use with spatial environmental data with the goal of making spatial predictions.

Why machine learning? A brief reply would be that, as modelling tools, most machine learning algorithms are universal, adaptive, nonlinear, robust and efficient. They can find acceptable solutions for the classification, regression, and probability density modelling problems in high-dimensional geo-feature spaces, composed of geographical space and additional relevant spatially referenced features. They are well suited to be implemented as predictive engines for decision-support systems, for the purpose of environmental data mining, including pattern recognition, modelling and predictions, and automatic data mapping. They compete efficiently with geostatistical models in low-dimensional geographical spaces, but they become indispensable in high-dimensional geo-feature spaces.

The book is complementary to a previous work [M. Kanevski and M. Maignan, *Analysis and Modelling of Spatial Environmental Data*, EPFL Press, 288 p., 2004] in which the main topics were related to data analysis using geostatistical predictions and simulations. The present book follows the same presentation: theory, applications, software tools and explicit examples. We hope that this organization will help to better understand the algorithms applied and lead to the adoption of this book for teaching and research in machine learning applications to geo- and environmental sciences. Therefore, an important part of the book is a collection of software tools – the Machine Learning Office – developed over the past ten years. The Machine Learning Office has been used both for teaching and for carrying out fundamental and applied research. We have implemented several machine learning algorithms and models of interest for geo- and environmental sciences into this software: the multilayer perceptron (a workhorse of machine learning); general regression neural networks; probabilistic neural networks; self-organizing maps; support vector machines; Gaussian mixture models; and radial basis-functions networks. Complementary tools useful for exploratory data analysis and visualisation are provided as well. The software has been optimized for user friendliness.

The book consists of 5 chapters. Chapter 1 is an introduction, wherein basic notions, concepts and problems are first presented. The concepts are illustrated by using simulated and real data sets.

Chapter 2 provides an introduction to exploratory spatial data analysis and presents the real-life data sets used in the book. A k-nearest-neighbour (k-NN) model is presented as a benchmark model for spatial pattern recognition and as a powerful tool for exploratory data analysis of both raw data and modelling results.

Chapter 3 is a brief overview of geostatistical predictions and simulations. Basic geostatistical models are explained and illustrated with the real examples. Geostatistics is a well established field of spatial statistics. It has a long and successful story in real data analysis and spatial predictions. Some geostatistical tools, like variograms, are used to efficiently control the quality of data modelling by MLA. This chapter will be of particular interest for those users of machine learning that would like to better understand the geostatistical approach and methodology. More detailed information on the geostatistical approach, models and case studies can be found in recently published books and reviews.

Chapter 4 reviews traditional machine learning algorithms – specifically artificial neural networks (ANN) of different architectures. Nowadays, ANN have found a wide range of applications in numerous scientific fields, in particular for the analysis and modelling of empirical data. They are used in social and natural sciences, both in the context of fundamental research and in applications. Neural network models overlap heavily with statistics, especially nonparametric statistics, since both fields study the analysis of data. Neural networks aim at obtaining the best possible generalisation performance without a restriction of model assumptions on the distributions of data generated by observed phenomena. Of course, being a data-driven approach, the efficiency of data modelling using ANN depends on the quality and quantity of available data. Neural network research is also an important branch of theoretical computer science. Each section of the Chapter 4 explains the theory behind each model and describes the case studies through the use of the Machine Learning Office software tools. Different mapping tasks are considered using simulated and real environmental data. The following ANN models are considered in detail: multilayer perceptron (MLP); radial basis-function (RBF) networks; general regression neural networks (GRNN); probabilistic neural networks (PNN); self-organizing Kohonen maps (SOM); Gaussian mixture models (GMM); and mixture density networks (MDN). These models can be used to solve a variety of regression, classification and density modelling tasks.

Chapter 5 provides an introduction to statistical learning theory. Over the past decades, this approach has proven to be among the most efficient and theoretically well-founded theories for the development of efficient learning algorithms from data. Then, the authors introduce the basic support vector machines (SVM) and support vector regression (SVR) models, along with a number of case studies. Some extensions to the models are then presented, notably in the context of kernel methods, including how these links with Gaussian processes. The chapter includes a variety of important environmental applications: robust multi-scale spatial data mapping and classification; optimisation of monitoring networks; and the analysis and modelling of high-dimensional data related to environmental phenomena.

The authors hope that this book will be of practical interest for graduate-level students, geo- and environmental scientists, engineers, and decision makers in their daily work on the analysis, modelling, predictions and visualisation of geospatial data.

The authors would like to thank our colleagues for numerous discussions on different topics concerning analysis, modelling, and visualisation of environmental data: Prof. M. Maignan, Prof. G. Christakos, Prof. S. Canu, Dr. V. Demyanov, Dr. E. Savelieva, Dr. S. Chernov, Dr. S. Bengio, Dr. M. Tonini, Dr. R. Purves, Dr. A. Vinciarelli, Dr. F. Camastra, Dr. F. Ratle, R. Tapia, D. Tuia, L. Foresti, Ch. Kaiser. We thank many participants of our workshops held in Switzerland, France, Italy, China, and Japan on machine learning for environmental applications, whose questions and challenging real life problems helped improve the book and software tools.

The authors gratefully acknowledge the support of Swiss National Science Foundation. The scientific work and new developments presented in this book were considerably supported by several SNSF projects: 105211-107862, 100012-113506, 200021-113944, 200020-121835 and SNSF Scope project IB7310-110915. This support was extremely important for the research presented in the book and for the software development. The supports of the University of Lausanne (Institute of Geomatics and Analysis of Risk - IGAR, Faculty of Geosciences and Environment) are also gratefully acknowledged.

We acknowledge the institutions and offices that have kindly provided us with the challenging data that stimulated us to better develop the models and software tools: IBRAE Institute of Russian Academy of Sciences (Moscow), Swiss Federal Office for Public Health, MeteoSwiss (Federal Office of Meteorology and Climatology), Swisstopo (Federal Office of Topography), Swiss Federal Statistical Office, Commission for the Protection of Lake Geneva – CIPEL (Nyon, Switzerland).

Finally, we acknowledge the publisher, the EPFL Press, and in particular Dr. F. Fenter for the fruitful collaboration during the preparation of this book.

TABLE OF CONTENTS

CHAPTER 1

LEARNING FROM GEOSPATIAL DATA

Machine learning is a very powerful approach to data analysis, modelling and visualization, and it is developing rapidly for applications in different fields. The key feature of machine learning algorithms is that they learn from empirical data and can be used in cases for which the modelled phenomena are hidden, non-evident, or not very well described. There are many different algorithms in machine learning, adopting many methods from nonparametric statistics, artificial intelligence research and computer science. In the present book, the machine learning algorithms (MLA) considered are among the most widely used algorithms for environmental studies: artificial neural networks (ANN) of different architectures and support vector machines (SVM).

In the context of machine learning, ANN and SVM are important models, not as an approach to develop artificial intelligence, but rather as universal nonlinear and adaptive tools used to solve data-driven classification and regression problems. In this perspective, machine learning is seen as an applied scientific discipline, while the general properties of statistical learning from data and mathematical theory of generalization from experience are more fundamental.

There exist many kinds of ANN that can be applied for different problems and cases: multilayer perceptron (MLP), radial basis function (RBF) networks, general regression neural networks (GRNN), probabilistic neural networks (PNN), mixture density networks (MDN), self-organizing maps (SOM) or Kohonen networks, etc. They were — and still are — efficiently used to solve data analysis and modelling problems, including numerous applications in geo- and environmental sciences.

Historically, artificial neural networks have been considered as black-box models. This is more the way they have been used rather than the essence of the methods themselves that have led to this attitude. A proper understanding of the methods provides many useful insights into what was previously considered as black-box. In recent studies it was also shown that efficiency of ANN models and the interpretability of their results can be radically improved by using ANN in a combination with statistical and geostatistical tools.

Recently a new paradigm emerged for learning from data, called *support vector machines*. They are based on a statistical learning theory (SLT) [Vapnik, 1998] that establishes a solid mathematical background for dependencies estimation and predictive learning from finite data sets. At first, SVM was proposed essentially for classification problems of two classes (dichotomies); later, it was generalized for multi-class classification problems and regression, as well as for estimation of probability densities. In the present book they are considered as important nonlinear, multi-scale, robust environmental data modelling tools in high dimensional spaces.

The present book is an attempt to present traditional (neural networks) and more recent (support vector machines and other kernel methods) developments in the filed of machine learning, with an emphasis on their applications for the analysis, modelling and visualization of spatial geo- and environmental data. These studies were started by the authors around 1992, aiming to model and to map extremely complex data on soil pollution after the Chernobyl accident [Kanevski et al., see references]. The methodology developed for this particular case study was successfully applied for many other interesting problems: pollution of air, water systems and soils, topo-climatic modelling, epidemiological data, crime data, geology and geophysics, natural hazards, etc. Geostatistics, being a well-founded field for modelling spatial data, was also in focus of spatial data modelling. The developments in modern data-driven methods have been carried out in parallel with the latter, the two approaches are complementary [Kanevski and Maignan, 2004].

The book of [Kanevski and Maignan, 2004] is mainly devoted to the approach of geostatistical analysis and modelling to geospatial data. Only a relatively small part was devoted to machine learning algorithms. The present book fills this gap. Of course, in this book attention is mainly paid to the models that have been widely used by the authors in many environmental applications and for which they have experience, including software development. It should be noted that software modules accompanying the book have been under development and tested over the last ten years. Of course, it does not guarantee that they are free of bugs and problems. But we believe that this software is an important, integral part of the book, allowing the readers to reproduce our numerical experiments and to start using it for their own teaching and research purposes. From the methodological point of view, the authors present some theory, applications (with real and simulated case studies) and software modules covering the algorithms presented in the book.

This book is mainly oriented to geo- and environmental community: graduate and Ph.D. students of environmental and earth-sciences departments, environmental engineers, researchers interested in machine learning methods and applications, and others working with geospatial data. Those who actively work in the field of machine learning itself can find some new challenging problems in environmental sciences, as described in this book.

1.1 PROBLEMS AND IMPORTANT CONCEPTS OF MACHINE LEARNING

The goal of machine learning is to develop methods that allow computers to learn. The most important concept of learning is not learning something by heart, but rather the gaining of experience and abilities to generalise the previously seen conditions onto new situations. Learning abilities are essential for human intelligence, and one of the main challenges for artificial intelligence research is to endow the machine with this capacity, either implemented as a set of algorithms or as a stand-alone robot. In the early age of the industrial era, as early as the first radio-controlled mechanism was created, the idea appeared already to make it adaptable, interactive and able to learn from experience [Tesla, 1900]. The engineering origins of machine learning continue to bring many challenging problems to fundamental research. These have brought to life many new scientific branches, such as speech recognition and computer vision, where machine learning is indispensable.

Machine learning has also become a branch of theoretical computer science. Computational learning theory, which studies the properties of learning from empirical data from a statistical perspective, is now an important field. Originating from the biological motivation to model biological neurons and the brain as a system of neurons with learning abilities, it has generalised these views with a solid mathematical theory.

Why is this field important for geo- and environmental sciences? To reply to this question, it is worth considering how scientific research has changed in past decades. With recent technological advances, our abilities to gather data in the world around us have drastically improved. Environmental science is the field which gradually benefits from these advances. Sensors, capable of measuring countless parameters, can be organised in wireless networks to provide great streams of information, often in real-time. While the storage of this data is mainly a technological and engineering problem, it remains an important scientific challenge to make effective use of this, in order to understand the underlying phenomena, to model them and to visualize the obtained results. Exploratory scientific research is becoming data-driven. And the methods for data-driven modelling have to tackle this situation correspondingly, with the field of machine learning giving significant promise.

Below, a rather qualitative introduction to the main machine learning task is presented. The description of the machine learning models, the theoretical concepts of learning from data, and the technical details of implementation of the methods can be found in Chapters 4 and 5. More in-depth descriptions can be found in in the following books [Bishop, 2007; Cherkassky and Mulier, 2007; Hastie et al., 2001; Scholkopf and Smola, 2002; Vapnik, 1995; Vapnik, 1998]. The amount of available literature on the topics of machine learning is gradually increasing, and the references above present only some of the recent editions of the popular textbooks.

1.1.1 Learning from data

The first step for building a system or an algorithm which can learn and generalise from empirical data is the formulation of an appropriate mathematical framework. Let us start with a definition of data. In many cases, an observation can generally be presented as a pair of entities, one describing the conditions where the observed event has happened (input space) and the other characterising the observed event or presenting its outcomes (output space). So, empirical knowledge can be formulated as a set of these input-output pairs. Both input and output data can be encoded as multi dimensional vectors, $\boldsymbol{x} = \{x^1, x^2, \ldots x^d\}$, $\boldsymbol{y} = \{y^1, y^2, \ldots y^s\}$. Sometimes the coordinates of $\boldsymbol{x}$ are called the *input features*. As for the outputs, they usually have much simpler structure, being, for example, simply a one-dimensional categorical (classes) or continuous value.

Setting up the learning problem

By making observations and collecting data, one usually supposes some kind of underlying phenomena that links inputs to outputs. Let us, without precising its nature, denote this dependence with F, such that F maps $\boldsymbol{x}$ to $\boldsymbol{y}$. While a deterministic mapping f(**x**): $\boldsymbol{x} \to \boldsymbol{y}$ is a natural way to link the vector spaces, it would be too optimistic to restrict the real-world processes which generate the data to be purely deterministic. With many factors influencing data and measurement processes in the real world, the whole setting becomes stochastic. A probabilistic description of the latter is thus preferable.

Some probability distribution P($\boldsymbol{x}$, $\boldsymbol{y}$), responsible for generating the data, is assumed to provide an acceptable description of the process. Obviously, the explicit form of this distribution is generally unknown, and only a set of empirical data is available. What is important in order to make some kind of inference from the available data set $\{\boldsymbol{x}, \boldsymbol{y}\}$ generated by P($\boldsymbol{x}$, $\boldsymbol{y}$), is to make sure that this set is consistent and representative enough to provide reliable knowledge about P($\boldsymbol{x}$, $\boldsymbol{y}$). This would mean one generally assumes that the $\{\boldsymbol{x}, \boldsymbol{y}\}$ are independent and identically distributed data sampled from the same population.

The knowledge of P($\boldsymbol{x}$, $\boldsymbol{y}$) would provide a full description of the process, and making any kind of inference about the distribution of $\boldsymbol{x}$, $\boldsymbol{y}$ or conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ becomes evident. The general problem of estimating P($\boldsymbol{x}$, $\boldsymbol{y}$) or even simply P($\boldsymbol{x}$) is a very difficult one. What one would actually like to know is not a distribution, but rather a particular property of the dependence between $\boldsymbol{x}$ and $\boldsymbol{y}$. Often, the actual questions of interest are more specific and more simple conceptually, such as "Would it be reasonable to characterize my new observation, $\boldsymbol{x}_{\text{new}}$, as belonging to the same class type $\boldsymbol{y}$ as I have observed before with a set of samples $\{\boldsymbol{x}_{\text{old}}\}$?" or "If I have observed that $y = 2$ at $\boldsymbol{x} = \{1.3, 2\}$ and $y = 3$ at $\boldsymbol{x} = \{1, 3.2\}$, what would be the value of y at $\boldsymbol{x} = \{1.1, 2.5\}$?". These are quite specific problems and one would not need to know P($\boldsymbol{x}$, $\boldsymbol{y}$) explicitly to answer these, though it would definitely be possible having it known. This is an approach to learning from data known as *discriminative*, as opposed to the *generative* approach when one models P($\boldsymbol{x}$, $\boldsymbol{y}$) first [Jebara, 2004].

Machine learning constructs algorithms able to predict the outputs for previously unknown inputs without making restrictive assumptions about P($\boldsymbol{x}$, $\boldsymbol{y}$). Some of its baseline ideas are purely algorithmic and distribution-independent. It is, however, essential to require that an available empirical data set is good enough (representative) and that the underlying process which we observe is the same we try to model and predict (i.e., the new samples come from the very same distribution/population as the training samples).

Main learning tasks

Depending on the type of observed outputs, or the way one chooses to encode them, different learning tasks may be introduced. First of all, the situation with no outputs at all can be considered. This is the case when one observes the environment without any possibility (or prior intention) of characterizing it by assigning an output to every input. For example, with a set of satellite images it may not be possible to assign a label to every image denoting the weather type or particular meteorological situation in the observed region. And yet, several soil probes have been taken in a region, and a data set of the chemical analysis of the latter is available. The goal of this effort is set as to explore the general dependencies in the data and then to relate them to their spatial distribution. The problems of learning in this case would be to make some kind of inference from the set of inputs $\mathrm{X} = \{\boldsymbol{x}_i\}_{i=1,\ldots \mathrm{N}}$. These types of learning problems are called *unsupervised* (see also Chapter 4).

Unsupervised learning. Two main unsupervised problems are usually considered. The first one is a *clustering* problem, formulated as to find some structures or typical groups (*clusters*) of vectors of inputs. The clusters are the regularities in data enforced by the underlying phenomenon, such as a finite number of typical weather types which produce similar satellite images of the region. In this case an important quantitative measures of similarity and dissimilarity has to be introduced. This task is visualized in Figure 1.1.

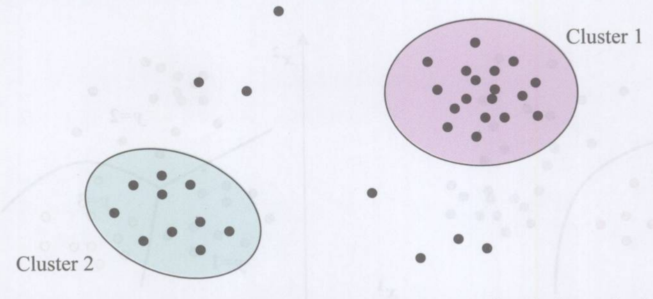


Fig. 1.1 Clustering problem: to find and characterize groups of typical samples in data.

Another type of unsupervised learning problem is *dimensionality reduction*. It is formulated so as to find a simpler, low dimensional representation of the observed high dimensional data which is as descriptive (in some predefined sense) as the original one. For example, one would like to find a low dimensional representation (*embedding*) which preserves some geometrical or topological properties of the original space. If one finds a way to reduce the dimension of the data to one, two or three dimensions, such problems obtain a natural and very important application – *visualization* of high dimensional data. The task of dimensionality reduction is illustrated in Figure 1.2. Dimensionality reduction is an important tool in modelling of noisy high dimensional data and in features extraction/selection analysis [Lee and Verleysen, 2007].

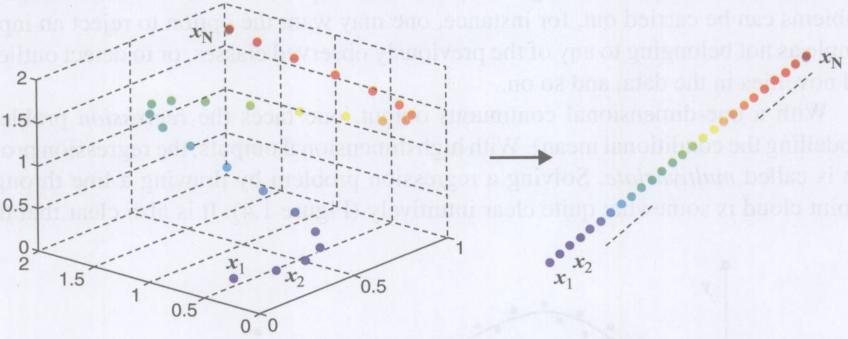


Fig. 1.2 Dimensionality reduction problem: to find a low dimensional representation of data.

The general unsupervised problem is a matter of estimating the probability density of input space $\{x\}$. Having a reliable model for $P(x)$, one would know virtually everything about x, including any kind of regularities and clusters.

Supervised learning. With outputs at hand, the learning problems becomes *supervised*, meaning that the examples which illustrate the input-output dependence are available to supervise our intention to model it. Different types of outputs induce different learning tasks (see also Chapter 4).

Let us first consider one-dimensional categorical outputs, $y = \{1, 2, 3, \ldots M\}$. This is the multi-class *classification* problem, as every observed input x is known to belong to some class y. The task is, using the available examples, to build a classification model, that is, to build a rule that assigns a class label to any previously unseen input vector (Figure 1.3). For instance, the soil type may be known in a finite set of locations in the region. The task at hand would be to construct a classifier which predicts a soil type at every other location in that region. Note that the soil map can be produced by applying this algorithm on the dense coordinate grid. If there are only two classes available, the problem is one of *binary classification*. One may think of other, more specific settings of classification problems, such as one-class classification: to confirm that a sample is of the same class as the available set. This setting differs from the binary classification because it is sometimes impossible to define or reliably describe the second class. This setting is closely linked to the unsupervised problem of probability density modelling. Many other task-specific variants of classification problems can be carried out, for instance, one may want the option to reject an input sample as not belonging to any of the previously observed classes, or to detect outliers and novelties in the data, and so on.

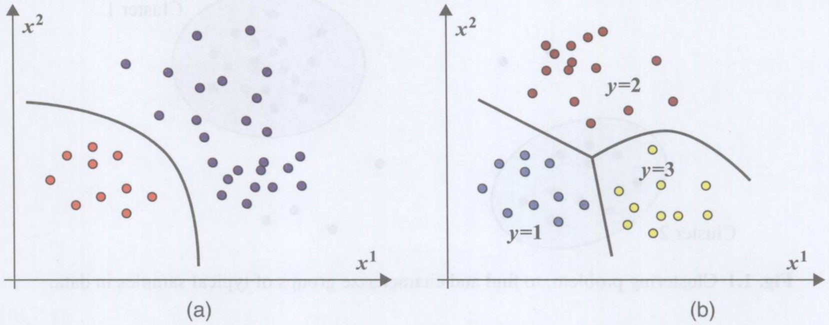


Fig. 1.3 Classification problems: binary (a) and multi-class (b).

With a one-dimensional continuous output, one faces the *regression* problem (modelling the conditional mean). With high dimensional outputs, the regression problem is called *multivariate*. Solving a regression problem by drawing a line through a point cloud is somewhat quite clear intuitively (Figure 1.4). It is also clear that the problem is distinct from data interpolation or approximation, as the data samples can be noisy and the line which explains the general input-output dependence does not need to pass through all the data points precisely. There are many other interesting problems one may associate with regression estimation, such as the estimation of

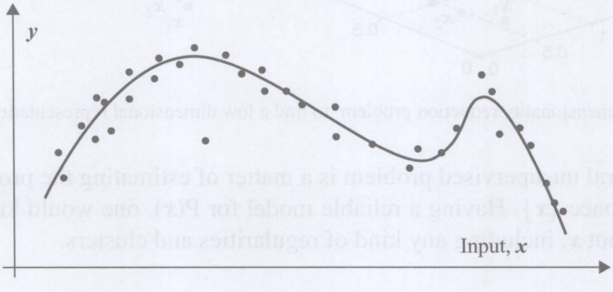


Fig. 1.4 A regression problem.

the uncertainty of the regression model, the estimation of the noise level in data, the behaviour of the regression model in the extrapolation areas, etc.

Other learning problems. In addition to the described tasks approached by machine learning algorithms, real-life data analysis brings some more interesting settings. Let us describe them briefly. A situation when only a small sample of labelled data (outputs are known) is accompanied with the large set of unlabelled data (inputs with unknown outputs) often encountered. An illustrative example of this situation concerns image classification. Suppose one has a remote sensing image of the region with several patches marked by a human expert as urban or rural zones. The task at hand is to build a classifier for the image to provide a segmentation of the image into urban and rural areas. Because the whole image is available, one can extract many patches and use them as unlabelled data together with the few labelled ones. This setting is known as *semi-supervised learning*. The information one obtains from the unlabelled part of the dataset mainly concerns the geometrical properties or the structure of the input space (Figure 1.5).

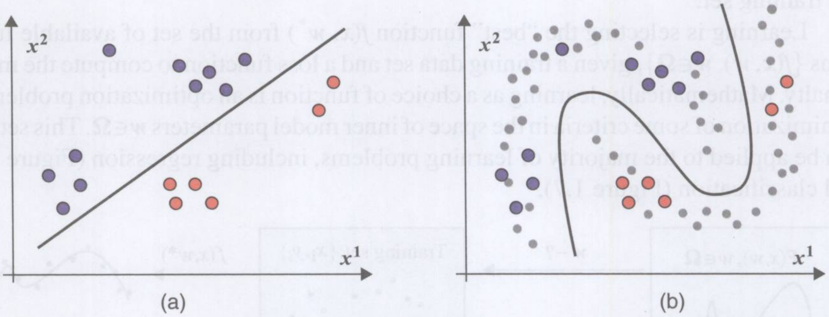


Fig. 1.5 Classification problem solved without (a) and with (b) unlabelled data. Unlabelled data, shown with grey dots in the right, are useful to bring information about the structure of the input space.

A special case of semi-supervised learning is the one where prediction is required only for the finite set of inputs that are known in advance. This setting can be considered as conceptually different from the traditional one, as one needs to build a model that operates, not in the whole input space, but rather on a very specific (finite or countable) subspace of it. It is known as a problem of *transduction*, or a problem of predicting the outputs of particular samples from a particular training set, opposed to the conventional inductive-deductive scheme of building a general model from particular samples and subsequently applying it for prediction [Vapnik, 1998].

Some other types of problems considered in machine learning arise when one needs to modify the behaviour of a system based on the feedback of whether a single modification trial is successive or not. This problem is known as *reinforcement learning*. It is mainly met in robotics and is not considered in this book. Another problem, more closely related to geospatial data modelling, is that of *active learning*. It is formulated in the following fashion: given a training set and a pool of unlabelled samples, one has to select a small number of samples from the pool that, being labelled (trough trials or by an expert) and added to the training set, would bring the highest possible improvement in performance to the current algorithm. This setting is much related to the problem of the *optimization of the monitoring network*.