

Chapter 1

Introduction

Last Modified: 2008-09-03 20:38

The traditional research paradigm in the sciences was hypothesis-driven. Starting with a hypothesis, the researchers designed experiments, collected data, and after data analysis, either validated the hypothesis or iterated the whole process via an updated hypothesis. The same was true in business applications, where data was collected to either validate or repudiate certain hypotheses about the customers. Over the last decade or so, this hypothesis-driven view has largely been replaced with a data-driven view of scientific research and business intelligence. In almost all fields of scientific endeavor, large research teams are systematically collecting data on questions of great import. However, a key element of this data amassing is the lack of prior hypotheses guiding the effort. With cheap, massive disk space, the idea is to gather as much data as possible. Hypothesis generation then becomes a matter of data analysis and mining, feeding this inversion of science and business analysis, i.e., rather than going from hypothesis to data, we use data to generate and validate hypotheses. Thus emerges the art and science of data mining.

The data-driven approach is readily witnessed in scientific areas like Bioinformatics, which was initially borne of the effort to understand the patterns in biological sequences. The poster child was the sequencing of the Human Genome, which has led to a revolution in sequencing of other genomes and more recently meta-genomes. Microarray technologies revolutionized what data could be collected about mRNA expression in cells, and with mass spectrometry and other proteomics technologies, we can now gather data about protein interactions. Analyzing and mining these diverse sources of data has led to the birth of systems biology. Astronomy offers another great example of the data-driven science. For example, the US National Virtual Observatory¹ provides a platform to access the numerous digital sky surveys over the web, allowing scientists and even amateurs, to analyze these data for exotic objects and phenomena. Ecological-informatics, geo-informatics, chem-informatics, and materials-informatics have similar aims in their corresponding fields. The data-driven trend was adopted even early on in economics and finance, and in business analysis. One must also not forget that the World Wide Web is itself a massive networked database, and analysis and search comprises a multi-billion dollar industry.

The fundamental goal of data mining is to extract patterns and to build models from the available datasets to discover useful and novel insights, or simply put, to transform the vast amounts of data into nuggets of knowledge. In a way, data mining seeks to make data a first-class object to be studied in its own right, looking for fundamental theories and models of data that span the disciplinary boundaries. Data mining aims at developing new methods and tools to study and characterize precisely these universal aspects of data. It is also worth emphasizing that the data encompasses the knowledge that is derived from it, as well

¹<http://www.us-vo.org/>

as the process (or workflow) that led to those knowledge nuggets, since post-discovery such knowledge and its context, becomes data for subsequent steps.

In this book we aim to cover the fundamental methods of data mining, and to also give a flavor of some of the exciting avenues of current research. The book is divided into four parts:

Exploratory Data Analysis aims to explore the attributes of the data individually or jointly to get a feel for measures like centrality (e.g., mean) and dispersion (e.g., variance) of the data values. Given that in data mining one typically encounters massive datasets with thousands of attributes and millions of points, another goal of exploratory analysis is data reduction methods that allow one to focus on the core characteristics of the data. For instance, feature selection and feature construction methods are used to select the most important dimensions/attributes. Sampling methods try to select a representative subset of points for further mining. Discretization methods try to reduce the number of values for the attributes, and so on. This part begins with uni-variate and multi-variate statistical analysis methods for both numeric and categorical attributes. It then discusses the challenges and anomalies of high-dimensional spaces. Finally, we discuss dimensionality reduction methods like Principal Component Analysis (PCA) and Singular Value Decomposition (SVD).

Frequent Pattern Mining deals with extracting informative and useful patterns in massive, complex datasets representing intricate and subtle interactions between diverse entities from a variety of sources. Pattern mining encompasses a whole range of tasks like mining frequent co-occurrences or itemsets, sequences, trees, and graphs, which will be the focus of this part.

Clustering is process used to partition points into *natural groups* or clusters, such that points within a group are very similar, whereas points across clusters are as dissimilar as possible. Depending on the characteristics of the data and the end-goals, different types of clustering paradigms have been proposed. This part starts with the basic partitioning approaches to clustering. It then discusses probabilistic, hierarchical, density-based, spectral and subspace clustering methods.

Classification Here we are given a *training* database of points that are labeled with predefined categories or classes, and the goal is to build a model that can predict the class for a point, for which we do not know the class beforehand. This part starts with partition-based classification, and then discusses the other paradigms like probabilistic classification, linear discriminant analysis, support vector machines and kernel methods.