UNIVERSITÀ DI PISA

Data Mining and Machine Learning

Educational Version Project
2020

Daria M. Maggi
M. Gomez Gomez

July 15, 2020

Contents

Chapter 1

Introduction

1.1 Why Data Mining?

'We are living in the information age' is a popular saying, we are actually living in the data age. A huge of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, medicine, science and engineering almost every other aspect of daily life.

This explosively growing, widely available, and gigantic body of data makes our time truly the data age. Powerful and versatile tools are badly needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge. This necessity has led to the birth of data mining.

Data Mining can be viewed as a result of the natural evolution of information technology. The database and data management industry involved in the development of several critical functionalities (Figure **??**): *data collection and database creation, data management and advanced data analysis.*

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor situation.* The fast-growing, tremendous amount of data, collected and stores in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools.

Consequently, important decisions are often made based not on the information rich data stored in data repositories but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. Efforts have been made to develop expert system and knowledge-based technologies, which typically rely on users or domain experts to *manually* input knowledge into knowledge bases.

1.2 What is Data Mining?

Data mining should have been more appropriately named *'knowledge mining from data'*, which is unfortunately somewhat long. The knowledge discovery process is shown as an interactive sequence of the following steps:

1. **Data cleaning** to remove noise and inconsistent data.
2. **Data integration** where multiple data sources may be combined.
3. **Data selection** where data relevant to the analysis task are retrieved from the database.
4. **Data transformation** where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data mining** an essential process where intelligent methods are applied to extract data patterns.
6. **Pattern evaluation** to identify the truly interesting patterns representing knowledge based on *interestingness measures*.
7. **Knowledge presentation** where visualization and knowledge representation techniques are used to present mined knowledge to users.

1.2.1 Classification and Regression for Predictive Analysis

Classification is the process of finding a **model** or function that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of **training data**. The model is used to predict the class label of objects for which the class label is unknown.

The derived model may be represented in various forms, such as *classification rules, decision trees, mathematical formulae, neural network* (Figura 1.2.1).

Figure 1.1: Classification model. Representation classification rule, decision tree, mathemathical formula

1.2.2 Cluster Analysis

1.2.3 Outlier Analysis

1.3 Which Technologies Are Used?

Chapter 2

Getting to know your data

Chapter 3

Data Preprocessing

Chapter 4

Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods

Chapter 5

Advanced Pattern Mining

Chapter 6

Classification: Basic Concepts

Chapter 7

Classification: Advanced Methods

Chapter 8

Cluster Analysis: Basic Concepts

Chapter 9

Cluster Analysis: Advanced Methods

Chapter 10

Outlier Detection

Chapter 11

MapReduce and Hadoop

11.1 What is MapReduce?

11.2 Why MapReduce?

11.3 Addressing the Scale Issue

11.4 MapReduce Characteristics

11.5 MapReduce Architecture

11.6 What is Hadoop?

11.7 Core Hadoop Components

11.8 Hadoop Limitations

List of Figures