Chapter 1

Introduction

Currently machine learning (ML) and data mining (DM) are the largest part of artificial intelligence (AI) research, and AI itself would be one of the largest research area in computer science and also being emerged as an important part of the entire science and more generally our society. Behind the technologies of AI, there have been developed various approaches and methods of ML for a variety of applications, although the history of ML itself is not so long, comparing with other research fields, like basic sciences and engineering.

ML research started in 1970s to 80s, and the community was not becoming so large for a long time, even after the DM research started around 1980s to 90s. However after around 2005 to 2010, the research area related with ML and DM has become drastically huge, affecting a lot of impacts on a variety of science and engineering fields, and also our society and life. Originally the application of ML was very limited and closely related with scientific or academic research areas, such as natural language processing, speech recognition, computer vision, robotics and medical engineering. However due to the development of internet, the applications were becoming more diverse, such as web searching, recommendation, fraud detection, etc. Also ML has accelerated pure science research, including drug discovery process, biological experiment management and physical or chemical simulations. Furthermore by the recent explosive data increase, the applications of ML are not limited to science and engineering but also abundant fields of our society, such as autonomous driving, industrial applications, finance, gaming, cooperate management, construction and agriculture.

There would be a lot of ways to classify a variety of ML and DM methods into some sort of order. In this book, we first focus on input data types. We segment ML and DM methods into seven categories, from a viewpoint of input data types. More in detail we categorize data types into the following sevens: *vectors*, *sets*, *sequences (strings)*, *trees*, *graphs*, *nodes in a graph* and *integrated data*. The last integrated data means that data can be a combination of one or more datasets, such as sequences plus graphs. We will define and also explain each of these seven types in Chapter 2. Then each of the seven data types is described by one chapter: Chapters 3 to 9, in the above order. In each chapter, we further classify the methods for the corresponding data type, by either problem settings or their techniques. That is, problem settings are supervised, unsupervised and semi-supervised learning. This was done for *vectors* and *nodes in a graph* (Chapters 3 and 8). On the other hand, the techniques are frequent pattern mining, probabilistic models (statistical learning) and kernel learning. This way of classification by techniques was done for *sets*, *sequences*, *trees*, *graphs* (Chapters 4 to 7). Then we explain one or more standard methods for each classified setting, focusing on already established methods. Also in the last of each chapter, we raise one or more application examples of bioinformatics.

These data types, such as vectors and graphs, look totally different from each other, and also maybe because of this, ML methods for different data types would be thought to be different from each other. They however definitely have some common ideas, motivations and derivation, etc., especially techniques behind different methods. One example is the formulation of principal component analysis, kernel K-means clustering (both for vectors), spectral clustering (for graphs), and canonical correlation analysis (for integrated data), all take the form of so-called Rayleigh quotient and eventually their setting can be solved by a (generalized) eigenvalue problem. Also there is a reverse case, in which different methods have been developed for the same problem setting for the same data type, while they have been developed by different ideas. We will explain these cases on what point they are different and also how reasonably they have been developed. We think that describing such shared points of different methods or unique features of each method will clarify each corresponding method.

Theoretical work are usually classified mainly by their significance into different levels, such as *Theorems*, *Lemmas*, etc. In this book theoretical results are all presented as *Propositions*, since they are rather already established.

After the seven chapters corresponding to the seven data types, we extensively explain practical manners of evaluating the results obtained by applying ML methods to actual data. Even if ML methods can be analyzed theoretically, it would be reasonable to investigate how well predictions by ML methods can be succeeded, to understand the performance of the ML methods themselves and also the hardness of the problem or the data, to which the methods are applied.

The last chapter is Appendix, which describes the terms/methods used in this book. Although the corresponding sections of Appendix are cited in the main text as much as possible, readers can refer Appendix if coming across the terms not described well in the main text. The last part of Appendix is the detail of several derivations in the main text.

The idea of this book is to cover major and standard approaches of already established problem settings in ML and DM extensively, showing their relevance and differences. Then ML and DM areas covered by this book are major, while much more ML topics are not considered in this book, because of rather being new or too complicated for an introductory text book. These uncovered topics include reinforcement learning, transfer learning, query learning, learning to rank, statistical Bayesian learning for numerous settings, such as Bayesiann belief networks, etc. There exist good books and reviews already for each of these problem settings, and interested readers can refer to them.