Preface

Today, we live in a world of connected things where tons of data is generated and it is humanly impossible to analyze all the incoming data and make decisions. Human decisions are increasingly replaced by decisions made by computers. Thanks to the field of data science. Data science has penetrated deeply in our connected world and there is a growing demand in the market for people who not only understand data science algorithms thoroughly, but are also capable of programming these algorithms. Data science is a field that is at the intersection of many fields, including data mining, machine learning, and statistics, to name a few. This puts an immense burden on all levels of data scientists; from the one who is aspiring to become a data scientist and those who are currently practitioners in this field. Treating these algorithms as a black box and using them in decision-making systems will lead to counterproductive results. With tons of algorithms and innumerable problems out there, it requires a good grasp of the underlying algorithms in order to choose the best one for any given problem.

Python as a programming language has evolved over the years and today, it is the number one choice for a data scientist. Its ability to act as a scripting language for quick prototype building and its sophisticated language constructs for full-fledged software development combined with its fantastic library support for numeric computations has led to its current popularity among data scientists and the general scientific programming community. Not just that, Python is also popular among web developers; thanks to frameworks such as Django and Flask.

This book has been carefully written to cater to the needs of a diverse range of data scientists—starting from novice data scientists to experienced ones—through carefully crafted recipes, which touch upon the different aspects of data science, including data exploration, data analysis and mining, machine learning, and large scale machine learning. Each chapter has been carefully crafted with recipes exploring these aspects. Sufficient math has been provided for the readers to understand the functioning of the algorithms in depth. Wherever necessary, enough references are provided for the curious readers. The recipes are written in such a way that they are easy to follow and understand.

This book brings the art of data science with power Python programming to the readers and helps them master the concepts of data science. Knowledge of Python is not mandatory to follow this book. Non-Python programmers can refer to the first chapter, which introduces the Python data structures and function programming concepts.

The early chapters cover the basics of data science and the later chapters are dedicated to advanced data science algorithms. State-of-the-art algorithms that are currently used in practice by leading data scientists across industries including the ensemble methods, random forest, regression with regularization, and others are covered in detail. Some of the algorithms that are popular in academia and still not widely introduced to the mainstream such as rotational forest are covered in detail.

With a lot of do-it-yourself books on data science today in the market, we feel that there is a gap in terms of covering the right mix of math philosophy behind the data science