Preface

> . . . the objective of statistical methods is the reduction of data. A quantity of data...is to be replaced by relatively few quantities which shall adequately represent...the relevant information contained in the original data. Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by an actual sample is irrelevant. It is the object of the statistical process employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.
>
> R.A. Fisher (1922)

These words by one of the greatest Statistician, Sir R.A. Fisher, speak by themselves. I would say, this sentence contains the essence of machine learning, although many things have changed from the last century. For instance, nowadays we typically face datasets where the number of observations is far greater than the set of distinct features. At those times, probably the biggest dataset on which Fisher was working was the Iris dataset, but nowadays we deal with datasets of millions of examples, and therefore all classical theoretical results would, trivially, be satisfied (e.g. the Central Limit Theorem is one of those). On the other hand, in many modern applications, we still have problems of dimensionality, and therefore those words are essentially still very important in the Machine Learning community. With different words, Andrew Ng, Computer Scientist at Stanford University, has recently came up with this sentence:

> Coming up with features is difficult, time-consuming, requires expert knowledge. *Applied machine learning* is basically feature engineering.

In my personal interpretation, featuring engineering is the modern concept of dimensionality reduction, because both of them aim at producing feature extraction

to improve the performances of the model. But honestly it would be very reductive to comprise the term Machine Learning to dimensionality reduction.

Machine Learning has gained a remarkable popularity in the last decade, not just because of the massive amount of available data, which is produced by ourself in everyday simple actions, but also because there is a wide consensus that learning from data leads to take better decisions and generate a better understanding of the phenomenon under investigation.

In its very general terms, Machine Learning (ML) can be understood as the set of algorithms and mathematical models that allow a system to autonomously perform a specific task, providing model-related scores and measures to evaluate its performances. It is sometimes confused with predictive (numerical) analytics, which is indeed part of ML, but more related to statistical learning. The range of applications of Machine Learning methods is vast and heterogeneous, from image recognition to topic detection in text analysis, from predicting whether a patient will suffer from breast cancer to predicting the price of a stock in three months from now.

The main objective of Machine Learning consists of predicting an outcome based on a set of features. The model is trained on a set of data, in which the target variable is available, and a predictor (or learner) is obtained. This learner is then used to predict the outcome on new data, that are not available at the time of training, and typically a good predictor is the one that accurately predicts the target variable.

This pipeline describes a discriminative *supervised learning* method, where we aim to predict a (continuous) target variable y based on some features $\mathbf{X}$. In this book, we will focus on *Shrinkage* estimators, *Support Vector Machine* algorithms, *Ensemble* methods and their applications to structured and unstructured data. However, in many applications, we could be just interested in finding some relationship between the target and the features: this is what *unsupervised learning* methods do. Although we will give more attention to supervised techniques, a great deal of attention will be given to techniques for dimensionality reduction, such as the *Principal Component Analysis*, which is a method that basically rotates the dataset in such a way that the rotated features are statistically uncorrelated.

The aim of this book is to introduce the reader to the main modern algorithms, employed by practitioners, to tackle Machine Learning problems, ranging from linear models to modern methods that easily deal with non-linear relationships.

The book has been thought for a broad, not strictly technical audience: on the one hand, the book was proposed for Bocconi Unviersity students, who actually come from applied sciences, and most likely want to learn modern ML techniques to

develop modern applications into Economics, Finance, Social and Political Sciences; on the other hand, I strongly believe this book can be a very good pocket-friend for all who wants to use machine learning in their data science and analytics tasks. Indeed, the book is proposed as a sort of cookbook, where each statistical model is presented, and the corresponding code section is provided to consistently apply those concepts to real problem.

I have intentionally avoided mathematics in most places because I believe it is (sometimes) a good distractor from the main objective of this manuscript, that is to empower the beginner learner with machine learning methods. Similarly, in many parts of this book, we have favored exposition over succinctness. I am aware of the fact that most of the code presented here could be tightened up, but that was rationally choosen to illustrate the methodologies to a broad target audience.

Hence, this manuscript was designed and written to be primarly used for practitioners, without the need of going into the math of the algorithms, although I strongly encourange to deepen those concepts with the reading of technical books and specific papers. If you are interested in the mathematics behind the proposed algorithms, there exists many books concerning technical aspects, which are mentioned throughout this book.

The key fact about this book is that it guides the reader into different methods, ranging from Bagging to the modern XGBoost, which is probably the first-best choice for any practitioner in machine learning. This is actually a strong point of this book: to the best of my knowledge, no book has been written giving particular attention to recent ensemble methods, such as XGBoost or CatBoost.

Python is the high-level language on which the analysis are carried out: this is indeed the modern language of applied Machine Learning, and notably modern softwares and techniques are developed in this language. Note that Python is an open-source software, and can be downloaded at the following link: https://www.python.org. I would say, it democratizes the coding era by allowing anyone to produce, promote and mantain a software easily and efficiently. Furthermore, I believe that once learned, it will be much more easier to follow the machine learning community on its developments and improvements.

The book is structured as follows: In Chapter 1, we will describe the standard pipeline that a machine learning algorithm follows: we will cover standard preprocessing and more advanced techniques, such as PCA for dimensionality reduction, and try to understand the fundamental relationship between bias and variance in ML. All the techniques are shown with practical examples. In Chapter 2, the reader will be introduced to a crucial concept in ML, which is the one of shrinkage. This

is very useful when we have to deal with many features, such as in genetics, and techniques such as Ridge and Lasso are shown. Furthermore, we will distinguish between classification and regression techniques, introducing firstly the Logistic Regression model and then the Support Vector Machine, which are two classifiers employed when data is linearly separable. A great deal of attention will also be given to non-linear SVM.
In Chapter 3 we will cover one of the the most popular ML techniques, that is ensemble methods, ranging from Random Forest to Gradient Boosting, with different applications. We will cover the XGBoost algorithm, which is the holy grail for any Machine Learner, and a great discussion is also given to SHAP values, which are a great tool to explain any model outputs to a non-technical audience. In Chapter 4 we will speak about two of the main areas where ML can be further investigated: Natural Language Processing and Deep Learning. Both are very hot topics, and the community is continuously working hard to improve the available models. We will just introduce those topics, so I strongly suggest the interest reader to deepen his knowledge with the given references. Since this book is aimed at reaching the broadest audience, I have also added in Appendix A a crash course in Python: this is aimed at not just covering the basics, since it will also introduce the reader to more broad concepts, that necessarily one has to deal with when working with machine learning models, such as Object-Oriented Programming.
Note that this book comes with an online version available at https://mybook.egeaonline.it/login. The online version cannot be downloaded, but it is a colour version to promote code readability. To facilitate the use of the proposed methods, and to improve the readability, I decided to create a book-specific library, called egeaML, which is publicly available on GitHub at https://github.com/andreagiussani/Applied_Machine_Learning_with_Python.
Please follow the instructions available in the GitHub repository to install it. Please, do note that the user can directly install it in any notebook environment, such as jupyter or colab, by simply typing and running the following snippet code:

```
!pip install git+https://github.com/andreagiussani/Applied_Machine_Learning_with_Python.git
```

The ! operator tells the notebook this is not a Python code, but a command line script. The datasets used within the book have been made easily accessible within the repository. Furthermore, the Git repository will be updated periodically with extra material and new notebooks, so I strongly suggest the reader to check it frequently.

Acknowledgments

I would like to thank many people who have helped me in writing this book. Most of them has given support and motivation to continue this project, some of them has given instead interesting insights and suggestions, and this project would never have seen the conclusion without the help of each of them. Among many, I would like to thank Alberto Clerici, who was the first one extremely interested in this project, and who helped me in setting up the final version of this book. I am also grateful to Marco Bonetti, whose interest for statistical learning methods has greatly helped me to improve myself daily, and this has definitely ameliorated the manuscript. I would also thank Egea for having given to me the possibility to write this manuscript with extreme flexibility, and also for the support on the formulation and preparation of this book. Finally, I am also grateful to many colleagues and friends, with who I have discussed this project and gave to me important insights: among them, I would like to thank Alberto Arrigoni, with who I have had nice chats on this fascinating topic, and Giorgio Conte, who has helped me in structuring the GitHub project.

Preface

> . . . the objective of statistical methods is the reduction of data. A quantity of data...is to be replaced by relatively few quantities which shall adequately represent...the relevant information contained in the original data. Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by an actual sample is irrelevant. It is the object of the statistical process employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.
>
> R.A. Fisher (1922)

These words by one of the greatest Statistician, Sir R.A. Fisher, speak by themselves. I would say, this sentence contains the essence of machine learning, although many things have changed from the last century. For instance, nowadays we typically face datasets where the number of observations is far greater than the set of distinct features. At those times, probably the biggest dataset on which Fisher was working was the Iris dataset, but nowadays we deal with datasets of millions of examples, and therefore all classical theoretical results would, trivially, be satisfied (e.g. the Central Limit Theorem is one of those). On the other hand, in many modern applications, we still have problems of dimensionality, and therefore those words are essentially still very important in the Machine Learning community. With different words, Andrew Ng, Computer Scientist at Stanford University, has recently came up with this sentence:

> Coming up with features is difficult, time-consuming, requires expert knowledge. *Applied machine learning* is basically feature engineering.

In my personal interpretation, featuring engineering is the modern concept of dimensionality reduction, because both of them aim at producing feature extraction

to improve the performances of the model. But honestly it would be very reductive to comprise the term Machine Learning to dimensionality reduction.

Machine Learning has gained a remarkable popularity in the last decade, not just because of the massive amount of available data, which is produced by ourself in everyday simple actions, but also because there is a wide consensus that learning from data leads to take better decisions and generate a better understanding of the phenomenon under investigation.

In its very general terms, Machine Learning (ML) can be understood as the set of algorithms and mathematical models that allow a system to autonomously perform a specific task, providing model-related scores and measures to evaluate its performances. It is sometimes confused with predictive (numerical) analytics, which is indeed part of ML, but more related to statistical learning. The range of applications of Machine Learning methods is vast and heterogeneous, from image recognition to topic detection in text analysis, from predicting whether a patient will suffer from breast cancer to predicting the price of a stock in three months from now.

The main objective of Machine Learning consists of predicting an outcome based on a set of features. The model is trained on a set of data, in which the target variable is available, and a predictor (or learner) is obtained. This learner is then used to predict the outcome on new data, that are not available at the time of training, and typically a good predictor is the one that accurately predicts the target variable.

This pipeline describes a discriminative *supervised learning* method, where we aim to predict a (continuous) target variable y based on some features $\mathbf{X}$. In this book, we will focus on *Shrinkage* estimators, *Support Vector Machine* algorithms, *Ensemble* methods and their applications to structured and unstructured data. However, in many applications, we could be just interested in finding some relationship between the target and the features: this is what *unsupervised learning* methods do. Although we will give more attention to supervised techniques, a great deal of attention will be given to techniques for dimensionality reduction, such as the *Principal Component Analysis*, which is a method that basically rotates the dataset in such a way that the rotated features are statistically uncorrelated.

The aim of this book is to introduce the reader to the main modern algorithms, employed by practitioners, to tackle Machine Learning problems, ranging from linear models to modern methods that easily deal with non-linear relationships.

The book has been thought for a broad, not strictly technical audience: on the one hand, the book was proposed for Bocconi Unviersity students, who actually come from applied sciences, and most likely want to learn modern ML techniques to

develop modern applications into Economics, Finance, Social and Political Sciences; on the other hand, I strongly believe this book can be a very good pocket-friend for all who wants to use machine learning in their data science and analytics tasks. Indeed, the book is proposed as a sort of cookbook, where each statistical model is presented, and the corresponding code section is provided to consistently apply those concepts to real problem.

I have intentionally avoided mathematics in most places because I believe it is (sometimes) a good distractor from the main objective of this manuscript, that is to empower the beginner learner with machine learning methods. Similarly, in many parts of this book, we have favored exposition over succinctness. I am aware of the fact that most of the code presented here could be tightened up, but that was rationally choosen to illustrate the methodologies to a broad target audience.

Hence, this manuscript was designed and written to be primarly used for practitioners, without the need of going into the math of the algorithms, although I strongly encourange to deepen those concepts with the reading of technical books and specific papers. If you are interested in the mathematics behind the proposed algorithms, there exists many books concerning technical aspects, which are mentioned throughout this book.

The key fact about this book is that it guides the reader into different methods, ranging from Bagging to the modern XGBoost, which is probably the first-best choice for any practitioner in machine learning. This is actually a strong point of this book: to the best of my knowledge, no book has been written giving particular attention to recent ensemble methods, such as XGBoost or CatBoost.

Python is the high-level language on which the analysis are carried out: this is indeed the modern language of applied Machine Learning, and notably modern softwares and techniques are developed in this language. Note that Python is an open-source software, and can be downloaded at the following link: https://www.python.org. I would say, it democratizes the coding era by allowing anyone to produce, promote and mantain a software easily and efficiently. Furthermore, I believe that once learned, it will be much more easier to follow the machine learning community on its developments and improvements.

The book is structured as follows: In Chapter 1, we will describe the standard pipeline that a machine learning algorithm follows: we will cover standard preprocessing and more advanced techniques, such as PCA for dimensionality reduction, and try to understand the fundamental relationship between bias and variance in ML. All the techniques are shown with practical examples. In Chapter 2, the reader will be introduced to a crucial concept in ML, which is the one of shrinkage. This

is very useful when we have to deal with many features, such as in genetics, and techniques such as Ridge and Lasso are shown. Furthermore, we will distinguish between classification and regression techniques, introducing firstly the Logistic Regression model and then the Support Vector Machine, which are two classifiers employed when data is linearly separable. A great deal of attention will also be given to non-linear SVM.
In Chapter 3 we will cover one of the the most popular ML techniques, that is ensemble methods, ranging from Random Forest to Gradient Boosting, with different applications. We will cover the XGBoost algorithm, which is the holy grail for any Machine Learner, and a great discussion is also given to SHAP values, which are a great tool to explain any model outputs to a non-technical audience. In Chapter 4 we will speak about two of the main areas where ML can be further investigated: Natural Language Processing and Deep Learning. Both are very hot topics, and the community is continuously working hard to improve the available models. We will just introduce those topics, so I strongly suggest the interest reader to deepen his knowledge with the given references. Since this book is aimed at reaching the broadest audience, I have also added in Appendix A a crash course in Python: this is aimed at not just covering the basics, since it will also introduce the reader to more broad concepts, that necessarily one has to deal with when working with machine learning models, such as Object-Oriented Programming.
Note that this book comes with an online version available at https://mybook.egeaonline.it/login. The online version cannot be downloaded, but it is a colour version to promote code readability. To facilitate the use of the proposed methods, and to improve the readability, I decided to create a book-specific library, called egeaML, which is publicly available on GitHub at https://github.com/andreagiussani/Applied_Machine_Learning_with_Python.
Please follow the instructions available in the GitHub repository to install it. Please, do note that the user can directly install it in any notebook environment, such as jupyter or colab, by simply typing and running the following snippet code:

```
!pip install git+https://github.com/andreagiussani/Applied_Machine_Learning_with_Python.git
```

The ! operator tells the notebook this is not a Python code, but a command line script. The datasets used within the book have been made easily accessible within the repository. Furthermore, the Git repository will be updated periodically with extra material and new notebooks, so I strongly suggest the reader to check it frequently.

Acknowledgments

I would like to thank many people who have helped me in writing this book. Most of them has given support and motivation to continue this project, some of them has given instead interesting insights and suggestions, and this project would never have seen the conclusion without the help of each of them. Among many, I would like to thank Alberto Clerici, who was the first one extremely interested in this project, and who helped me in setting up the final version of this book. I am also grateful to Marco Bonetti, whose interest for statistical learning methods has greatly helped me to improve myself daily, and this has definitely ameliorated the manuscript. I would also thank Egea for having given to me the possibility to write this manuscript with extreme flexibility, and also for the support on the formulation and preparation of this book. Finally, I am also grateful to many colleagues and friends, with who I have discussed this project and gave to me important insights: among them, I would like to thank Alberto Arrigoni, with who I have had nice chats on this fascinating topic, and Giorgio Conte, who has helped me in structuring the GitHub project.