Executive Summary

What exactly constitutes Data Science is not universally agreed upon, but it is certainly inseparable from Machine Learning. Some would even consider Data Science as a subfield of Machine Learning—its intersection with application domains. This book adopts a different viewpoint: Data Science is seen as incorporating the field of Machine Learning.

A necessarily incomplete outline of this vast field is presented in the first of the book's five parts. It starts by considering the scenario of supervised learning, in which a to-be-learned function f is available only through point values $y_i = f(x^{(i)})$ at datapoints $x^{(1)}, \ldots, x^{(m)}$. In Statistical Learning Theory, these datapoints are assumed to be realizations of some hidden random variable. Chapter 1 introduces the main notions attached to this theory, in particular the PAC-learning framework. Chapter 2 scrutinizes the concept of VC-dimension, in anticipation of its connection to the problem of binary classification, where the labels y_i take only two values. Chapter 3, of a technical nature, makes this connection precise by establishing the fundamental theorem of PAC-learning. Chapter 4 continues to probe the problem of binary classification but drops the statistical setting. It proposes some tools—in particular, support vector machines—to separate datapoints and it also acquaints the readers with kernel methods. Chapter 5 takes a careful look at the associated reproducing kernel Hilbert spaces. Chapter 6 concludes the tour of supervised learning by way of a few peeks at the regression problem, featuring real-valued labels y_i. Chapter 7 turns to the scenario of unsupervised learning, in which the labels are absent: the task examined there consists in exploiting similarity information about the datapoints to cluster them in a meaningful way. Finally, Chapter 8 presents common techniques to deal with the hindering high-dimensionality of datapoints.

Readers in search of a more detailed exposition to Machine Learning are referred to the books by Shalev-Shwartz and Ben-David (2014) and Mohri et al. (2018). For more targeted reading, they can also consult the books by Hastie et al. (2009), Scholkopf and Smola (2001), and Vershynin (2018).

https://doi.org/10.1017/9781009003933.004 Published online by Cambridge University Press