Introduction to Data Science 1MS041

[1], Benny Avelin[1], and Raazesh Sainudiin[1]

[1]Department of Mathematics, Uppsala University, Uppsala, Sweden

December 20, 2022

Preface

These set of lecture notes began as notes for the course Introduction to Data Science during the spring and summer of 2021. This course was designed as a mathematical introduction to datascience, covering most of the basics one would need to start their education in data science, in the sense of giving the reader a strong mathematical foundation on which to stand in the future. Our belief is that in order to develop new algorithms for data science / AI problems, one needs a strong mathematical intuition.

Another aim with these notes have been to bridge a gap between math, theoretical computer science and modern approaches concerning concentration of measure. Most of this material can be found in other texts, but scattered and with wildly differing levels of rigor.

Another novel point of these notes is that we focus quite a lot on separating the statistical model (assumptions about the data) from the estimation procedure (computer algorithms). This idea is hidden in most of modern data science and often goes against traditional parametric estimation, where the assumption is fairly often that the true underlying parameter one tries to estimate, is among those searched for. For instance, in linear regression, one assumes that the truth is linear and we are just trying to find that. In modern data-science where the goal is one of prediction (mostly) one instead does not assume that the truth is linear but one tries to approximate it with a linear function, and if the fit is good one is happy.

The concentration of measure phenomenon permeates these notes and we will use it to arrive at non-asymptotic estimates (finite sample bounds) in many practically useful cases, from performance metrics of classification to compression of data using dimensionality reduction. The goal has been to provide quantitative estimates for almost all problems in these notes, while some have been left out, they can be approached using the methods developed in these notes.

These notes suit students with little knowledge of probability theory, it is however easier to digest if you are familiar with the mathematical way of thinking, i.e. in that of abstraction.

Topics

- Axiomatic probability: Chapters 1 and 2. These chapters cover the mathematical basics needed for the rest of the notes. We have chosen a fairly rigorous way of presenting axiomatic probability which is very flexible and after you get to know it, very easy to use as there is very little ambiguity.

- Concentration of measure: Chapter 3. This is the main backbone of these notes, all chapters following rely on the results obtained here (except: Chapter 7). We have decided to only touch on the simplest concentration inequalities, i.e. Hoeffding's inequality and similar.

- Risk: Chapter 4. This chapter concerns the concept of Risk and how you can phrase common problems, like regression, pattern recognition and parameter estimation as risk minimization problems. All estimation problems that appear later in these notes will be a risk minimization problem, and specifically empirical risk minimization problems.

- Fundamentals of estimation: Chapter 5. Covers the traditional statistical terminology surrounding parameter estimation. Like consistency, bias or asymptotic properties.

- Random variable Generation: Chapter 6. Introduces the concept of pseudo-randomness, some ways to produce it on the computer, and how to use it to sample from arbitrary distributions.

- Finite Markov Chains: Chapter 7. This chapter introduces Markov chains as a means of modelling more than just i.i.d. samples. It is also where you will see a natural interpretation of the σ-algebra as history. Markov chains are essential in many sequential problems and is the simplest form of time-series.

- Pattern recognition and Regression: Chapters 8 and 9. This chapter covers pattern recognition and regression from the perspective of a-priori perfomance or a-posteriori testing. That is, what can we say about the performance of an algorithm without a test-set and what can we say once we have a test-set? These chapters relies on the fairly advanced topic of VC-dimension and growth functions. The a-posteriori testing is most important for a first course, as well as understanding the difference between guaranteeing performance beforehand or afterwards.

- High dimension and Dimensionality reduction: Chapters 10 and 11. These set of notes end with another look at concentration from the perspective of dimension and utilize this to perform dimensionality reduction. We also cover singular value decomposition and its use in image compression/data.

Contents

List of Figures