

Introduction to Data Science 1MS041

¹, Benny Avelin¹, and Raazesh Sainudiin¹

¹Department of Mathematics, Uppsala University, Uppsala, Sweden

This work is licensed¹ under [CC BY NC SA 4.0](#)
December 20, 2022

¹Partly based on [CseBook](#) by [lamastex](#) under [CC BY NC SA 4.0](#)

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 performance 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 rely 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

1	Probability Model	1
1.1	Experiments	1
1.2	Probability	3
1.2.1	Consequences of our Definition of Probability	5
1.2.2	More on Sigma Algebras	8
1.3	Conditional Probability	9
1.3.1	Bayes' Theorem	10
1.3.2	Independence and Dependence	12
1.4	Extension of probability*	14
2	Random Variables	15
2.1	Basic Definitions	15
2.2	Discrete Random Variables	18
2.3	Continuous Random Variables	20
2.3.1	Viewing a deterministic real variable as a random variable	22
2.4	Transformations of random variables	23
2.4.1	Transformations of discrete random variables	23
2.4.2	Transformations of continuous random variables	24
2.5	Expectations and L^p spaces	30
2.6	Multivariate Random Variables	31
2.6.1	Discrete random vectors	33
2.6.2	Continuous random vectors	35
2.6.3	Properties of expectations	36
2.6.4	L^p is a normed vector space	37
2.6.5	Conditional Random Variables	41
2.6.6	Mixed random variables	43
2.7	Examples Of Modeling	44
2.7.1	Email spam filtering	44
2.7.2	Number of website requests during a day	45
2.7.3	Summary	46

3	Concentration and Limits	48
3.1	Concentration inequalities	48
3.1.1	Random variables that are not exponentially integrable*	57
3.2	Convergence of Random Variables	58
3.2.1	Properties of Convergence of RVs**	62
3.3	Law of Large Numbers	63
3.4	Central Limit Theorem	64
4	Risk	66
4.1	The supervised learning problem	67
4.1.1	Mathematical description of the learning problem "find f "	68
4.1.2	Finding the regression function $r(x) = \mathbb{E}[Y X]$	69
4.1.3	The pattern recognition problem (classification)	71
4.2	Maximum Likelihood Estimation	73
4.2.1	Maximum Likelihood and regression	74
5	Fundamentals of Estimation	77
5.1	Introduction	77
5.2	Point Estimation	77
5.2.1	Some Properties of Point Estimators	79
5.3	Non-parametric DF Estimation	84
5.4	Plug-in Estimators of Statistical Functionals: Direct estimation	87
6	Random Variable Generation	91
6.1	Congruential Generators	92
6.2	Sampling	95
6.3	Practice exercises	97
7	Finite Markov Chains	99
7.1	Introduction	99
7.1.1	Advanced intro*	100
7.1.2	Non advanced introduction	101
7.2	Random Mapping Representation and Simulation	105
7.3	Irreducibility and Aperiodicity	106
7.4	Stationarity	107
7.5	Reversibility	108
7.5.1	Random Walks on Graphs	108
7.6	Computer exercises	110
8	Pattern recognition	112
8.1	Linear Classifiers	112
8.1.1	Linearly Separable Dataset	113
8.1.2	The perceptron algorithm	113

8.2	Kernelization	116
8.2.1	Other types of Kernels	119
8.3	Theoretical guarantees	119
8.3.1	Guarantees with a held out testing set	120
8.3.2	Other test metrics	121
8.4	Empirical Risk Minimization for Linear Classifiers	122
8.4.1	A classifier with finitely many hyperplanes (without testing)	122
8.5	Preliminaries for VC theory	125
8.6	VC theory	126
8.7	Vapnik Chervonenkis dimension	130
8.8	What if you don't care about $\inf R(\phi)$?	132
8.9	Bibliography	134
9	Regression	135
9.1	Guarantees with a held out testing set	136
9.1.1	R^2	137
9.2	Bibliography	140
10	High dimension	141
10.1	Introduction: Volume of the unit ball in d dimensions	141
10.2	The geometry of high dimension	144
10.3	Properties of the unit ball	145
10.4	Uniform at random from a ball and sphere	148
10.4.1	Generating points uniformly at random from a circle	148
10.4.2	Uniform at random on the unit sphere in high dimension	150
10.4.3	Uniform at random from the unit ball B_1 ?	151
10.5	High dimensional annulus theorem	151
10.6	Bibliography	152
11	Dimensionality reduction	153
11.1	Random Projection and Johnson – Lindenstrauss Lemma	153
11.2	SVD (Singular Value Decomposition)	155
11.2.1	The power method	161
11.3	PCA	162
11.4	SVD in Action	162
11.4.1	Factor Analysis	162
11.4.2	Example on compressing data	163
11.4.3	Anomaly detection and reconstruction error	165
11.5	Theoretical analysis	165
11.6	Reconstruction error	166
11.7	Bibliography	167

12 Group Assignments	168
12.1 Group Assignment 1	168
12.2 Group Assignment 2	168
12.3 Group Assignment 3	168
Index	171

List of Figures

1.1	Reference to the Venn digram will help you understand this idea behind the proof of the total probability theorem in Theorem 1.16 for the four event case.	11
2.1	PDF and DF of a $\text{Normal}(\mu, \sigma^2)$ RV for different values of μ and σ^2	27
3.1	Examples of distributions that are sub-exponential and sub-Gaussian	57
3.2	PDF $f_{X_n}(x) := \mathbf{1}_{(0,1)}(x)(1 - \cos(2\pi nx))$ of the RV X_n [the left sub-figure] and its DF $F_n(x) := \int_{-\infty}^x \mathbf{1}_{(0,1)}(v)(1 - \cos(2\pi nv))dv$ [the right sub-figure], for $n = 1$ [red '-'], $n = 10$ [blue '-'], and $n = 100$ [green '-'], respectively. One can see clear convergence of the DFs F_n to $\mathbf{1}_{(0,1)}(x)x$, the DF of the $\text{Uniform}(0, 1)$ RV, while the corresponding PDFs $f_n(x)$ keep oscillating wildly with n across $[0, 2]$ about $\mathbf{1}_{(0,1)}(x)$, the PDF of the $\text{Uniform}(0, 1)$ RV X . Thus giving a counter-example to the claim that convergence in DFs does not imply convergence in PDFs.	60
8.1	Linearly separable data with labels $+1$ or red and -1 or blue.	113
8.2	Linearly non-separable data in two dimensions.	116
8.3	Linearly separable in three dimensions after $(x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2)$	117
11.1	The distribution of relative error on the Olivetti faces dataset using only $k = 20$ and $k = 400$ respectively.	155
11.2	Sample data for SVD	155
11.3	The data from Fig. 11.2 projected onto the normal of the plane defined by v_1	157
11.4	10 sample images from Mnist	163
11.5	The data from Fig. 11.4 projected onto the plane defined by the first 10 singular vectors.	164