The Analysis of Data

Volume 1

Probability

Guy Lebanon

First Edition

2012

Probability. The Analysis of Data, Volume 1.

First Edition, First Printing, 2013

http://theanalysisofdata.com

In memory of Alex Lebanon

Contents

Preface

The Analysis of Data Project

The Analysis of Data (TAOD) project provides educational material in the area of data analysis.

- The project features comprehensive coverage of all relevant disciplines including probability, statistics, computing, and machine learning.
- The content is almost self-contained and includes mathematical prerequisites and basic computing concepts.
- The R programming language is used to demonstrate the contents. Full code is available, facilitating reproducibility of experiments and letting readers experiment with variations of the code.
- The presentation is mathematically rigorous, and includes derivations and proofs in most cases.
- HTML versions are freely available on the website http://theanalysisofdata.com. Hardcopies are available at affordable prices.

Volume 1: Probability

This volume focuses on probability theory. There are many excellent textbooks on probability, and yet this book differs from others in several ways.

- Probability theory is a wide field. This book focuses on the parts of probability that are most relevant for statistics and machine learning.
- The book contains almost all of the mathematical prerequisites, including set theory, metric spaces, linear algebra, differentiation, integration, and measure theory.
- Almost all results in the book appear with a proof.

- Probability textbooks are typically either elementary or advanced. This book strikes a balance by attempting to avoid measure theory where possible, but resorting to measure theory and other advanced material in a few places where they are essential.

- The book uses R to illustrate concepts. Full code is available in the book, facilitating reproducibility of experiments and letting readers experiment with variations of the code.

I am not aware of a single textbook that covers the material from probability theory that is necessary and sufficient for an in-depth understanding of statistics and machine learning. This book represents my best effort in that direction.

Since this book is part of a series of books on data analysis, it does not include any statistics or machine learning. Such content is postponed to future volumes.

Website

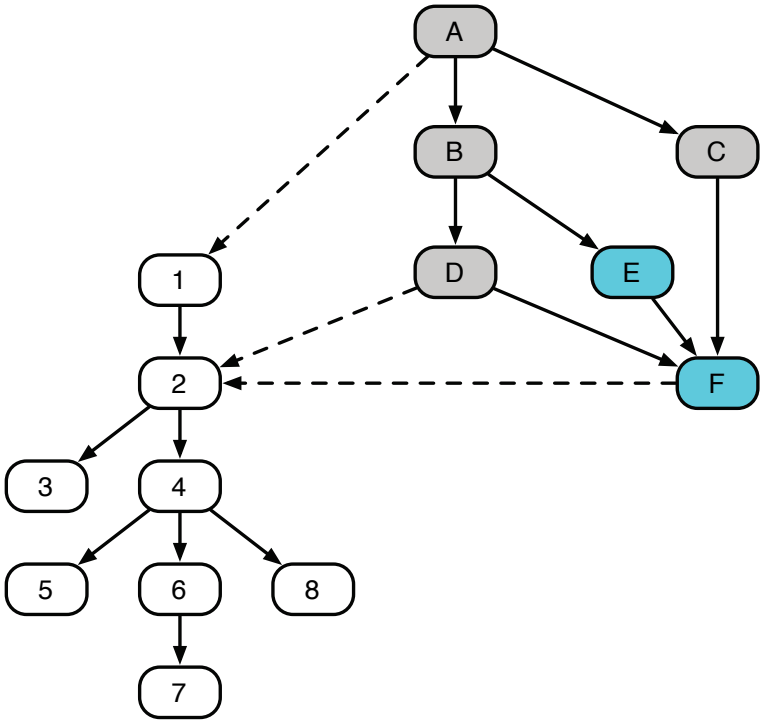
A companion website (http://theanalysisofdata.com) contains an HTML version of this book, errata, and additional multimedia material. The website will also link to additional TAOD volumes as they become available.

Mathematical Appendices

A large part of the book contains six appendices on mathematical prerequisites. Probability requires knowledge of many branches of mathematics, including calculus, linear algebra, set theory, metric spaces, measure, and Lebesgue integration. Instead of referring the reader to a large collection of math textbooks we include here all of the necessary prerequisites. References are provided in the notes sections at the end of each chapter for additional resources.

Dependencies

The diagram below indicates the dependencies between the different chapters of the book. Appendix chapters are shaded and dependencies between appendix chapters and regular chapters are marked by dashed arrows. It is not essential to strictly adhere to this dependency as many chapters require only a brief familiarity with some issues in the chapters that they depend on.



Starred sections correspond to material that requires measure theory, or that can be better appreciated with knowledge of measure theory.

R Code

The book contains many fragments of R code, aimed to illustrate probability theory and its applications. The code is included so that the reader can reproduce the results as well as modify the code and run variations of it. In order to appreciate the code and modify it, the reader will need a basic understanding of the R programming language and R graphics. A good introduction to R is available from the CRAN website at http://cran.r-project.org/doc/manuals/R-intro.pdf. Alternatively, two chapters from volume 2 of the analysis of data series (R Programming and R Graphics) are freely available online at http://theanalysisofdata.com.

The book uses a variety of R graphics packages, including `base`, `lattice`, and `ggplot2`. The most frequently used package is `ggplot2`, which is described in detail in [49] (see also http://had.co.nz/ggplot2/ and the R Graphics chapter available at http://theanalysisofdata.com.)

To ensure that the code fragments run locally as they appear in the book, install and bring into scope the required packages using the `install.packages` and `library` R functions. For example, to install and load the `ggplot2` package type the following commands in the R prompt.

```
# this is a comment
install.packages("ggplot2")  # installs the package ggplot2
library("ggplot2")  # brings ggplot2 into scope
```

The code fragments throughout the book are annotated with output displayed by the R interpreter. This output is displayed following two hash symbols which are interpreted as comments by R (see below).

```
a = pi
print(a)  # note the ## symbols preceding the output below

## [1] 3.142

print(a + 1)

## [1] 4.142
```

This format makes it easy to copy a code fragment from an HTML page and paste it directly into R (the output following the hash symbols will be interpreted as comments and not produce a syntax error).

Acknowledgements

The following people made technical suggestions that helped improve the contents: Krishnakumar Balasubramanian, Rohit Banga, Joshua Dillon, Sanjeet Hajarnis, Oded Green, Yi Mao, Seungyeon Kim, Joonseok Lee, Fuxin Li, Nishant Mehta, Yaron Rachlin, Parikshit Ram, Kaushik Rangadurai, Neil Slagle, Mingxuan Sun, Brian Steber, Gena Tang, Long Tran. In addition, many useful comments were received through a discussion board during my fall 2011 class computational data analysis at Georgia Tech. These comments were mostly anonymized, but some commentators who identified themselves appear above.

Katharina Probst, Neil Slagle and Laura Usselman edited portions of this book, and made many useful suggestions. The book features a combination of text, equations, graphs, and R code, made possible by the `knitr` package. I thank Yihui Xie for implementing `knitr`, and for his help through the `knitr` Google discussion group. Katharina Probst helped with web development and design.

Mathematical Notations

	Logic
$\forall$	for all
$\exists$	there exists
$\Rightarrow$	implies
$\Leftrightarrow$	if and only if
$\stackrel{\text{def}}{=}$	defined as

For example, $\forall a > 0, 1/a > 0$ reads "for every $a > 0$ we have $1/a > 0$" and $\forall a > 0, \exists b > 0,\ a/b = 1 \Rightarrow b/a = 1$ reads "for all $a > 0$ there exists $b > 0$ such that, whenever $a/b = 1$, we also have $b/a = 1$".

	Sets and Functions (Chapter A)
Ω	sample space
ω	an element of the sample space Ω
I_A	indicator function $I_A(x) = 1$ if $x \in A$ and 0 otherwise
δ_{ij}	Kronercker's delta: $\delta_{ij} = 1$ if $i = j$ and 0 otherwise
$\mathbb{N}$	natural numbers $\{1, 2, 3, \ldots\}$
$\mathbb{Z}$	integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Q}$	rational numbers
$A \times B$	Cartesian product of two sets
A^k	repeated Cartesian product (k times) of a set: $A \times \cdots \times A$
A^c	complement of a set: $\Omega \setminus A$
2^A	power set of the set A
$\lvert A \rvert$	number of elements in a finite set
$f^{-1}(A)$	pre-image of the function f: $\{x : f(x) \in A\}$
$f \equiv g$	$f(x) = g(x)$ for all x
$x^{(n)}$	sequence of elements in a set
$A_n \nearrow A$	convergence of a sequence of increasing sets
$A_n \searrow A$	convergence of a sequence of decreasing sets

	Combinatorics (Section 1.6)
$n!$	factorial function $n(n-1)\cdots 2 \cdot 1$
$(n)_r$	$n!/(n-r)!$
$\binom{n}{r}$	$n!/(r!(n-r)!)$

	Metric spaces (Chapter B)
$\mathbb{R}$	real numbers
$\mathbb{R}^d$	set of d-dimensional vectors of real numbers
$B_a(\boldsymbol{x})$	open ball of radius a centered at $\boldsymbol{x}$
$\boldsymbol{x}$	vector in $\mathbb{R}^d$
x_i	the i-component of the vector $\boldsymbol{x}$
$\boldsymbol{x}^{(n)}$	sequence of vectors in $\mathbb{R}^d$
$x_i^{(j)}$	the i-component of the vector $\boldsymbol{x}^{(j)}$
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$	inner product $\sum_i x_i y_i$
$\|\boldsymbol{x}\|$	Euclidean norm $\sqrt{\sum_i x_i^2}$
$d(\boldsymbol{x}, \boldsymbol{y})$	Euclidean distance $\sqrt{\sum_i (x_i - y_i)^2}$
$\boldsymbol{x}^{(n)} \to \boldsymbol{x}$	convergence of a sequence
$f_n \to f$	pointwise convergence of a sequence of functions
$f_n \nearrow f$	pointwise convergence of an increasing sequence of functions
$f_n \searrow f$	pointwise convergence of a decreasing sequence of functions
$O(f)$	big O growth notation
$o(f)$	little o growth notation

Note in particular that we denote vectors in bold face, for example $\boldsymbol{x}$, and the scalar components of such vectors using subscripts (non-bold face) $\boldsymbol{x} = (x_1, \ldots, x_d)$. We refer to sequence of vectors using superscripts, for example $\boldsymbol{x}^{(n)}$ and their scalar components as $\boldsymbol{x}^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)})$.

	Probability (Chapters 1, 2, 4, 8)
P	probability function
P_E	probability conditioned on the event E
F_X	cumulative distribution function (cdf) corresponding to the RV X
f_X	probability density function (pdf) corresponding to the RV X
p_X	probability mass function (pmf) corresponding to the RV X
E	expectation
Var	variance
Cov	covariance
std	standard deviation
$\sim$	distributed according to
$\overset{\text{iid}}{\sim}$	independent identically distributed (iid) sampling
$\mathbb{P}_k$	simplex of probability functions over $\Omega = \{1, \ldots, k\}$
i.o.	infinitely often
$\xrightarrow{\text{as}}$	convergence with probability 1
$\xrightarrow{\text{p}}$	convergence in probability
$\rightsquigarrow$	convergence in distribution

	Matrices (Chapter C)
$A^\top$	matrix transpose
A^n	the matrix A raised to the n-power
A^n_{ij}	the i, j element of the matrix A^n
A^{-1}	inverse of matrix A
I	identity matrix
$\mathsf{tr}\, A$	trace of the matrix A
$\det A$	determinant of the matrix A
$\mathsf{row}\, A$	row space of the matrix A
$\mathsf{col}\, A$	column space of the matrix A
$\mathsf{dim}\, A$	dimension of a linear space
$\mathsf{rank}\, A$	rank of the matrix A
$\mathsf{diag}(\boldsymbol{v})$	diagonal matrix whose diagonal is given by the vector $\boldsymbol{v}$
$A \otimes B$	Kronecker product of two matrices A, B

All vectors are assumed to be column vectors unless stated explicitly otherwise. For example, if $\boldsymbol{x}$ is an arbitrary vector and A a matrix the expression $\boldsymbol{x}^\top A \boldsymbol{x}$ represents a scalar.

	Differentiation (Chapter D)
df/dx	derivative
d^2f/dx^2	second order derivative
$\partial f/\partial x_i$	partial derivative
∇f	gradient vector of partial derivatives of $f : \mathbb{R}^k \to \mathbb{R}$
$\nabla \boldsymbol{f}$	Jacobian matrix of partial derivatives of $\boldsymbol{f} : \mathbb{R}^k \to \mathbb{R}^m$
$\partial^2 f/\partial x_i \partial x_j$	second order partial derivative
$\nabla^2 f$	Hessian matrix of second order partial derivatives of $f : \mathbb{R}^k \to \mathbb{R}$

We consider the gradient vector ∇f as a row vector. This ensures that the notations ∇f and $\nabla \boldsymbol{f}$ are consistent.

	Measure and Integration (Chapters E, F)
$\sigma(\mathcal{C})$	σ algebra generated by the set of sets $\mathcal{C}$
μ	measure
$\int d\mu$	Lebesgue integral with respect to measure μ
$\int d\,\mathsf{P}$	Lebesgue integral with respect to probability measure P
$\int dF_X$	Lebesgue integral with respect to probability measure corresponding to cdf F_X
$\int dx$	Lebesgue integral with respect to Lebesgue measure or Riemann integral
$\mathcal{B}(\mathbb{R}^d)$	Borel σ-algebra over metric space $\mathbb{R}^d$
$\mathcal{F}_1 \otimes \mathcal{F}_1$	product σ-algebra
$\mu_1 \times \mu_2$	product measure
a.e.	almost everywhere (except on a set of measure zero)