Foundations of data-science engineering

Steffen Mæland, PierGianLuca Porta Mana

2023-05-04

Contents

Preface

Under construction

Chapter 1

Introduction

To be written: motivation and structure of this course.

Chapter 2

Data science: use and communication

2.1 Statements

Facts, hypotheses, questions, decisions – and data – are communicated through language and sentences. You may say "well, data can be just numbers, they don't need to be communicated through sentences". But is that true?

I give you this number: "5". OK it's a number, but what's it about? what should you do with it? is that "data"? Instead, if I tell you: "The number of lectures in this course is 5" then I have given you a piece of information, a datum (even if it is actually false). Underlying any piece of information, hypothesis, or datum, there is always a *statement* – also called *sentence* or *proposition*[1] – that gives you the meaning and context of that datum.

2.2 Well-posed and ill-posed statements

We face problems when the statements that should convey information are not clear. Suppose that an electric-car model consumes 150 Wh/km and has a range of 200 km; a second car model consumes 250 Wh/km and has a range of 600 km. Someone asks you: "which model is better?". Well, it isn't clear how you should answer; what does "better" mean? If it refers to consumption, then the first car is "better". If it refers to range, then the second car is "better". If it refers to a combination of these two characteristics, or to something else, then you simply can't answer. Here we have a problem with querying and giving data, because the statement underlying such query is not clear. We say that statement is not **well-posed**, or that it is **ill-posed**.

[1] These terms are not equivalent in Logic, but we'll use them as synonyms here.

This may seem an obvious discussion to you. Yet you'd be surprised by how often unclear statements appear in scientific papers about data engineering! Not seldom we find discussions and disagreements that actually come from unclear underlying statements, that two parties interpret in different ways.

As a data engineer, you'll often have the upper hand if you are on the lookout for ill-posed statements. Whenever you face an important question, or you're given an important piece of information, or you must provide an important piece of information, always take a little time to examine whether the question or information is actually well-posed.

- *Exercise: give actual paper to analyse*

Chapter 3

Truth inference and probability inference

3.1 Truth, falsity, and their consistency

3.2 Inferences without uncertainty: the truth calculus

3.3 Making room for uncertainty:Plausibility, credibility, degree of belief, probability

3.4 Inferences with uncertainty: the probability calculus

3.4.1 The Three Fundamental Laws of inference

- *Exercise: Monty-Hall problem & variations*
- *Exercise: clinical test & diagnosis*

3.4.2 Bayes's theorem

3.5 Common points of certain and uncertain inference

No premises? No conclusions!

Chapter 4

Data and information

4.1 Kinds of data

4.1.1 Binary

4.1.2 Nominal

4.1.3 Ordinal

4.1.4 Continuous

- unbounded
- bounded
- censored

4.1.5 Complex data

2D, 3D, images, graphs, etc.

4.1.6 "Soft" data

- orders of magnitude
- physical bounds

4.2 Data transformations

- log
- probit

- logit

Chapter 5

Allocation of uncertainty among possible data values: probability distributions

5.1 The difference between Statistics and Probability Theory

Statistics is the study of collective properties of collections of data. It does not imply that there is any uncertainty.

Probability theory is the quantification and propagation of uncertainty. It does not imply that we have collections of data.

5.2 What's "distributed"?

Difference between distribution of probability and distribution of (a collection of) data.

5.3 Distributions of probability

5.3.1 Representations

- Density function
- Histogram
- Scatter plot

Behaviour of representations under transformations of data.

5.4 Summaries of distributions of probability

5.4.1 Location

Median, mean

5.4.2 Dispersion or range

Quantiles & quartiles, interquartile range, median absolute deviation, standard deviation, half-range

5.4.3 Resolution

Differential entropy

5.4.4 Behaviour of summaries under transformations of data and errors in data

5.5 Outliers and out-of-population data

(Warnings against tail-cutting and similar nonsense-practices)

5.6 Marginal and conditional distributions of probability

5.7 Collecting and sampling data

5.7.1 "Representative" samples

Size of minimal representative sample = (2^entropy)/precision

- *Exercise: data with 14 binary variates, 10000 samples*

5.7.2 Unavoidable sampling biases

In high dimensions, all datasets are outliers.

Data splits and cross-validation cannot correct sampling biases

5.8 Quirks and warnings about high-dimensional data

Chapter 6

Making decisions

6.1 Decisions, possible situations, and consequences

6.2 Gains and losses: utilities

6.2.1 Factors that enter utility quantification

Utilities can rarely be assigned a priori.

6.3 Making decisions under uncertainty: maximization of expected utility

Chapter 7

The most general inference problem

Chapter 8

Literature

Here is a review of existing methods.