ADA511 Data science and data-driven engineering

Steffen Mæland, PierGianLuca Porta Mana

2023-05-27

Table of contents

Preface **4**

1 Introduction **5**

2 Data: use and communication **6**
2.1 Statements – or, what is "data"? 6
2.2 Well-posed and ill-posed statements 6
Reading list 7

3 Inference **8**
3.1 What is inference? 8

4 Truth inference and probability inference **9**
4.1 Truth, falsity, and their consistency 9
4.2 Inferences without uncertainty: the truth calculus 9
4.3 Making room for uncertainty:Plausibility, credibility, degree of belief, probability 9
4.4 Inferences with uncertainty: the probability calculus 9
4.4.1 The Three Fundamental Laws of inference 9
4.4.2 Bayes's theorem 9
4.5 Common points of certain and uncertain inference 9

5 Data and information **10**
5.1 Kinds of data 10
5.1.1 Binary 10
5.1.2 Nominal 10
5.1.3 Ordinal 10
5.1.4 Continuous 10
5.1.5 Complex data 10
5.1.6 "Soft" data 10
5.2 Data transformations 10

6 Allocation of uncertainty among possible data values: probability distributions **11**
6.1 The difference between Statistics and Probability Theory 11
6.2 What's "distributed"? 11
6.3 Distributions of probability 11
6.3.1 Representations 11

6.4 Summaries of distributions of probability . . . 11
6.4.1 Location . . . 11
6.4.2 Dispersion or range . . . 12
6.4.3 Resolution . . . 12
6.4.4 Behaviour of summaries under transformations of data and errors in data 12
6.5 Outliers and out-of-population data . . . 12
6.6 Marginal and conditional distributions of probability . . . 12
6.7 Collecting and sampling data . . . 12
6.7.1 "Representative" samples . . . 12
6.7.2 Unavoidable sampling biases . . . 12
6.8 Quirks and warnings about high-dimensional data . . . 12

7 Making decisions **13**
7.1 Decisions, possible situations, and consequences . . . 13
7.2 Gains and losses: utilities . . . 13
7.2.1 Factors that enter utility quantification . . . 13
7.3 Making decisions under uncertainty: maximization of expected utility . . . 13

8 The most general inference problem **14**

Preface

To be written.

1 Introduction

To be written: motivation and structure of this course.

2 Data: use and communication

2.1 Statements – or, what is "data"?

What is "data"?

"Data" (from Latin "given") is used more or less in the same sense as "information", and in these notes we'll use the two words as synonyms.

"Data" is often presented as numbers; but it's obviously more than that. I give you this number: "8". Is it "data"? what is it about? what should you do with it? We can hardly call this number a piece of information, since we have no clue what we could do with it. Instead, if I tell you: "*The number of official planets in the solar system is 8*", then we can say that I've given you data. So "data" is not just numbers. A number is not "data" unless there's some verbal, non-numeric context associated with it – even if this context is only implicitly understood.

Data can also be completely non-numeric. A clinician saying "*The patient has fully recovered from the disease*" (we imagine to know who's the patient and what was the disease) is giving us a piece of information that we could further use, for instance, to make prognoses about other, similar patients. The clinician's statement surely is "data". It is essentially non-numeric data, even if in some situations we can represent it as "1", say, while "0" would represent "not recovered".

From these two examples, and with some further thought, we realize that "data" – and in general any piece of information or hypothesis – can universally be represented and communicated by *statements* , also called *sentences* or *propositions*[1]. In some cases we can summarize or represent such statements as numbers. But the numbers alone, by themselves, are not data.

So our conclusion is that *information* or *data* is represented by *statements*.

2.2 Well-posed and ill-posed statements

We face problems when the statements that should convey information and data 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 says "I think the second

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

model is better; what do you think?". It isn't clear how we should answer; what does "better" mean? If it refers to consumption, then the first car model is "better". If it refers to range, then the second model is "better". If it refers to a combination of these two characteristics, or to something else, then we 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 such statement are **not well-posed**, or that they are **ill-posed**.

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.*

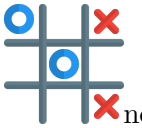
- *[TODO] Exercise: give actual paper to analyse*

Reading list

3 Inference

3.1 What is inference?

One core problem in all data-driven engineering applications – and in daily life too – is that *we would like to know something that we don't know*. In other words, we would like to have more information. We may wish to acquire information out of simple curiosity; or we may need it for some specific reason or goal, as we'll discuss later. Examples:

1. We'd like to know whether it'll rain today, so we can decide whether to get an umbrella or rain clothes, or not to.
2. A clinician would like to know which disease affects a patient, so as to decide for the optimal treatment.
3. The X-player of this game of Xs & Os:  needs to know where put the next **X** in order to win.
4. The computer of a self-driving car needs to know whether a particular patch of colours in the visual field is a person, so as to slow down the car and stop.
5. A rocket engineer needs to know how much is $\sqrt{2\,G\,M\,r}$, where $G = 6.67 \cdot 10^{-11}\,\mathrm{m^3\,s^{-2}\,kg^{-1}}$, and M and r are the mass and radius of the Earth, in order to launch a rocket to the Moon.
6. An aircraft's autopilot system needs to predict how altering a wing's attack edge will change the aircraft's roll.

4 Truth inference and probability inference

4.1 Truth, falsity, and their consistency

4.2 Inferences without uncertainty: the truth calculus

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

4.4 Inferences with uncertainty: the probability calculus

4.4.1 The Three Fundamental Laws of inference

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

4.4.2 Bayes's theorem

4.5 Common points of certain and uncertain inference

No premises? No conclusions!

5 Data and information

5.1 Kinds of data

5.1.1 Binary

5.1.2 Nominal

5.1.3 Ordinal

5.1.4 Continuous

- unbounded
- bounded
- censored

5.1.5 Complex data

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

5.1.6 "Soft" data

- orders of magnitude
- physical bounds

5.2 Data transformations

- log
- probit
- logit

6 Allocation of uncertainty among possible data values: probability distributions

6.1 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.

6.2 What's "distributed"?

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

6.3 Distributions of probability

6.3.1 Representations

- Density function
- Histogram
- Scatter plot

Behaviour of representations under transformations of data.

6.4 Summaries of distributions of probability

6.4.1 Location

Median, mean

6.4.2 Dispersion or range

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

6.4.3 Resolution

Differential entropy

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

6.5 Outliers and out-of-population data

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

6.6 Marginal and conditional distributions of probability

6.7 Collecting and sampling data

6.7.1 "Representative" samples

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

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

6.7.2 Unavoidable sampling biases

In high dimensions, all datasets are outliers.

Data splits and cross-validation cannot correct sampling biases

6.8 Quirks and warnings about high-dimensional data

7 Making decisions

7.1 Decisions, possible situations, and consequences

7.2 Gains and losses: utilities

7.2.1 Factors that enter utility quantification

Utilities can rarely be assigned a priori.

7.3 Making decisions under uncertainty: maximization of expected utility

8 The most general inference problem