

ADA511: Data science and data-driven engineering

Steffen Mæland PierGianLuca Porta Mana

2023-06-03

Table of contents

Preface	4
1 Introduction	5
2 Data: use and communication	6
2.1 Sentences – or, what is “data”?	6
2.2 Well-posed and ill-posed sentences	7
Reading list	8
3 Inference	9
3.1 What is inference?	9
3.2 Certain and uncertain inference	10
4 Truth inference and probability inference	12
4.1 Logic: the rules for certain inference	12
4.1.1 Basic sentences	12
4.1.2 Connectives	13
4.2 Truth, falsity, and their consistency	14
4.3 Inferences without uncertainty: the truth calculus	14
4.4 Making room for uncertainty: Plausibility, credi- bility, degree of belief, probability	14
4.5 Inferences with uncertainty: the probability cal- culus	14
4.5.1 The Three Fundamental Laws of inference	14
4.5.2 Bayes’s theorem	14
4.6 Common points of certain and uncertain inference	14
5 Data and information	15
5.1 Kinds of data	15
5.1.1 Binary	15
5.1.2 Nominal	15
5.1.3 Ordinal	15
5.1.4 Continuous	15
5.1.5 Complex data	15

5.1.6	“Soft” data	15
5.2	Data transformations	16
6	Allocation of uncertainty among possible data values: probability distributions	17
6.1	The difference between Statistics and Probability Theory	17
6.2	What’s “distributed”?	17
6.3	Distributions of probability	17
6.3.1	Representations	17
6.4	Summaries of distributions of probability	18
6.4.1	Location	18
6.4.2	Dispersion or range	18
6.4.3	Resolution	18
6.4.4	Behaviour of summaries under transformations of data and errors in data	18
6.5	Outliers and out-of-population data	18
6.6	Marginal and conditional distributions of probability	18
6.7	Collecting and sampling data	18
6.7.1	“Representative” samples	18
6.7.2	Unavoidable sampling biases	19
6.8	Quirks and warnings about high-dimensional data	19
7	Making decisions	20
7.1	Decisions, possible situations, and consequences	20
7.2	Gains and losses: utilities	20
7.2.1	Factors that enter utility quantification	20
7.3	Making decisions under uncertainty: maximization of expected utility	20
8	The most general inference problem	21

Preface

****WARNING: THIS IS A WORKING DRAFT. TEXT WILL CHANGE A LOT. MANY PASSAGES ARE JUST TEMPORARY, INCOHERENT, AND DISJOINTED.**

To be written.

1 Introduction

To be written: motivation and structure of this course.

2 Data: use and communication

2.1 Sentences – 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 *sentences*, also called *statements* or *propositions*¹. In some

¹ These terms are not equivalent in Logic, but sometimes we’ll use them as synonyms.

cases we can summarize or represent such sentences as numbers. But the numbers alone, by themselves, are not data.

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

Recognizing that data and information are ultimately sentences has important practical consequences:

Clarity and goal-orientation. As a data engineer you’ll have to acquire information and convey information. Acquiring information is not simply making some measurement or counting something: you must understand *what* you are measuring and *why*. If you gather data from third parties, you have to ask what exactly the data mean and how they were acquired. In designing and engineering a solution, you’ll have to understand what information or outcomes the end user exactly wants. It will often happen that you ask “wait, what do you mean by that?”; this question is not just an unofficial parenthesis in the official data-transfer workflow between you and someone else: it is an integral part of that workflow, it means that the data has not been completely transferred yet.

Artificial Intelligence Sentences are the central components of knowledge representation and inference in artificial-intelligence agents.

2.2 Well-posed and ill-posed sentences

We face problems when the sentences 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 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

 Reading material

[§ 7.1 in *Artificial Intelligence*](#)

problem with querying and giving data, because the sentence underlying such query is not clear.

We say that such sentences 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 sentences appear in scientific papers about data engineering! Not seldom we find discussions and disagreements that actually come from unclear underlying sentences, 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 sentences. 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?

The first core problem in all data-driven engineering applications – and in daily life too – is to *draw inferences*, that is, acquire information. We may wish to acquire information out of simple curiosity, or for some specific engineering 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.
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, within two significant digits, **how much is the velocity** $\sqrt{2GM/r}$, where $G = 6.67 \cdot 10^{-11} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1}$, and $M = 5.97 \cdot 10^{24} \text{ kg}$ and $r = 6.37 \cdot 10^6 \text{ m}$ are the mass and radius of the Earth, in order to launch a rocket to the Moon.
6. We’d like to know whether the rolled die will show , so we can win a bet.
7. An **aircraft’s autopilot system** needs to predict how much the **aircraft’s roll** will change by increasing the right wing’s **angle of attack** by 0.1 rad.
8. An archaeologist would like to know whether the fossil bone just dug out belonged to a Tyrannosaurus rex.

9. An automated system in an assembly line needs to predict whether an electric component of a widget will fail within the next two years.

Note how each of these inferences boils down to determining whether some sentences are true or false. In example 1. we want to know whether the sentence “*It rains today*” is true or not. In example 2. the clinician wants to know which of the sentences “*The patient has pneumonia*”, “*The patient has asthma*”, “*The patient has bronchitis*”, and so on, are true (several can be true at the same time). In example 5. the rocket engineer wants to know which among the sentences “*The velocity is 0.010 m/s*”, “*The velocity is 0.011 m/s*”, ..., “*The velocity is 130 m/s*”, and so on, is true. The sentences that underlie an inference can be extremely many and complex, and yet we must have an idea of what they are (otherwise, do we really know what our inference is about?).

3.2 Certain and uncertain inference

The example inferences above present very different levels of difficulty.

Inferences 3. and 5. are special because they can actually be drawn *exactly*, that is, we really find out which of their underlying sentences are true and false. In example 3. it is trivial that putting the next **X** in the mid-right slot makes the X-player win. In example 5. a couple of mathematical operations show that the sentence “*The velocity is 11 km/s*” is true. When we can obtain the data we want from the data we have by using “only”² logic and mathematical operations, our inference is *certain*, also called a “deduction”; in these notes we shall call it a *truth inference*. But every deduction can be basically drawn by repeatedly applying the rules of logic.

The other example inferences cannot be drawn exactly, in the sense that we cannot know for sure whether all their underlying sentences are true or false. But this doesn’t mean that we cannot say anything whatsoever. In example 6. we consider the sentence “*The die shows* ” to be more likely false than true. In

Exercise

Try to identify which sentences underlie the other example inferences.

² “Only” in quotation marks because the logical analysis and operations leading to the answer can still be computationally very expensive.

example 2. the clinician might be quite sure about the disease, after observing the symptoms. On the other hand, in example 1. we might really have no clue whether “*It rains today*” will turn out to be true or false. These inferences are *uncertain*. Certain inferences can be considered as a limit case of uncertain ones, in which the uncertainty vanishes or is extremely small.

To draw certain inferences, we follow the rules of Logic. What rules do we follow to draw uncertain inferences?

4 Truth inference and probability inference

4.1 Logic: the rules for certain inference

Consider the following trivial but certain inference:

$$\frac{\text{“My umbrella is either blue or red”} \quad \text{“My umbrella is not red”}}{\text{“My umbrella is blue”}}$$

Above the line we write the sentences representing the data we have. Below the line we infer the information that supposedly interests us.

How could we draw this obvious inference? Which rules did we follow?

Logic is a huge field that formalizes and makes rigorous the rules that a rational person or an artificial intelligence should use in drawing certain inferences. We’ll get a glimpse of it here, as a trampoline for jumping towards our data-driven engineering problems.

4.1.1 Basic sentences

We start by writing down the *basic*³ sentences that constitute our data and that underlie the inferences we want to draw. “Basic” in the sense that we will not analyse these sentences into further sub-sentences. In the trivial example above we identify two such sentences: “*My umbrella is blue*”, and “*My umbrella is pink*”. Let’s represent them by symbols:

$$\begin{aligned} b &:= \text{“My umbrella is blue”} \\ r &:= \text{“My umbrella is red”} \end{aligned}$$

³ A more technical term is “atomic”

4.1.2 Connectives

You notice that we didn't consider "*My umbrella is either blue or red*" and "*My umbrella is not red*" as basic sentences. These sentences can indeed be expressed in terms of the basic sentences b and r . We consider one way or operation to change a sentence into another related to it, and two ways or operations to combine two or more sentences together. These operations are called "connectives". Our natural language offer many more operations to combine sentences, but these three turn out to be all we need in virtually all engineering problems. The three connectives are:

Not: \neg for example,

$$\neg r = \text{"My umbrella is not red"}$$

And: \wedge for example,

$$b \wedge r = \text{"My umbrella is blue, and it is red"}$$

Or: \vee for example,

$$b \vee r = \text{"My umbrella is blue, or it is red"}$$



Note some subtleties in

4.2 Truth, falsity, and their consistency

4.3 Inferences without uncertainty: the truth calculus

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

4.5 Inferences with uncertainty: the probability calculus

4.5.1 The Three Fundamental Laws of inference

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

4.5.2 Bayes's theorem

4.6 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 = $(2^{\text{entropy}})/\text{precision}$

- *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