Foundations of data science

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

Truth inference and probability inference

2.1 Sentences

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 *sentence* – also called *statement* 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 sentences 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*

2.3 Truth, falsity, and their consistency

2.4 Inferences when there's no uncertainty: the truth calculus

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

2.6 Inferences when there's uncertainty: the probability calculus

Chapter 3

Literature

Here is a review of existing methods.