A Mathematical Challenge

Arising from AI governance

Richard Pinch
Heilbronn Institute for Mathematical Research
rgep@chalcedon.co.uk

Abstract

I claim that there are significant gaps between the ethical, professional, regulatory and legal requirements likely to be imposed on AI systems on the one hand; and the theory, techniques, tools and tradecraft available to designers, developers and users on the other.

I propose a simple challenge to those working with decision systems which may help to identify some of the mathematical problems that arise in bridging these gaps.

Introduction

Ethical applications of AI systems are subject to requirements such as safety, security, privacy, interpretability, contestability and sustainability. These require the ability to reason about their behaviour and to quantify uncertainty. Experience working with professional bodies suggests that there is in general a lack of theories, techniques, tools and tradecraft to meet these requirements.

In order to calibrate these gaps, I propose a simple challenge, not in itself an important question but one which identifies some (but not all) of the questions – and gaps – that arise in meeting more realistic challenges from AI governance.

Overview

We consider a decision support system, such as a classifier, from three aspects, each of which requires a principled discussion.

- Data
  - We need to be able to make assertions about the universe from which the data is taken.
- Function
  - We need to be able to make formal statements about what decisions are to be made; in particular we will need to be able to quantify such aspects as uncertainty, risk and confidence.
- Engine
  - We need to describe the decision-making engine in sufficiently precise terms to reason about and quantify its behaviour.

In this poster, I consider the engine largely as a mathematical abstraction.

The decision engine

We consider the decision engine as a family of functions, paramatrised by weights, defined on some standard (large) space of inputs, and having outputs in some (much smaller) space of values.

The training process consists of finding a set of weights that gives the best approximation to known training values with respect to some score or loss function.

We assume that the spaces involved are vector spaces over the real numbers; and that the arithmetic involved is performed using some specific hardware approximation.

Challenge

I pose the following challenge:

> **Consider a decision engine with a training data set T. Let E denote the engine trained on the data set T; let E' be the engine trained on the data set T' consisting of T in reverse order.**
> **Show that E and E' are functionally the same.**

Discussion

Of course this is not as important as questions of practical governance — although I suggest that it is not entirely unimportant either. My position is that *the theoretical understanding of a type of decision engine required to answer the challenge is a prerequsite for making principled claims about the practical behaviour of such decision systems*.

Some points that a successful answer will need to address:

- analysis of the optimisation algorithm implicit in training and the nature of the loss landscape;
- notions of distance between sets of weights;
- descriptions of functional behaviour;
- the relation between similarity of weights and behaviours;
- the effects of implementation in hardware arithmetic.

Acknowledgements

I acknowledge helpful discussions with colleagues at HIMR; the IMA Professional Affairs and AI groups; RSS and Alliance for Data Science Professionals.

I am solely responsible for the opinions expressed in this poster.