ROOFS AND BOXES

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1998 Sep 9, 12:59 p.m.

Abstract

This note presents an example, (*roofs-and-boxes*), to refute the idea that sequence extrapolation is a paradigmatic problem for AI. This plausible idea was that intelligence predicted the sequence of future sensations from the past sequence of sensations. The *roofs-and-boxes* example illustrates that intelligence requires knowing about objects in the world and not just about one's history of sensations—even if one's goal is to predict future sensations.

The justification for writing this up many years after I discussed it in lectures is that almost all machine learning research does not undertake to infer structures in the world and not just classify the data. I'll explain this point after giving the example.

1 Introduction

This note presents an example, *roofs-nd-oxes*, to refute the idea that sequence extrapolation is the paradigmatic problem for AI. This plausible idea was that intelligence predicted the sequence of future sensations from the past sequence of sensations. This idea led to programs for sequence extrapolation. The first programs predicted sequences of integers generated by polynomials, and later programs dealt with sequences generated by programs that included conditional expressions.

Programs for sequence extrapolation were written by Edward Fredkin, Donald Michie, Jan Mycielski and others. I don't have the references yet.

My objection to taking this as a paradigm is that the prediction of the future in real life involves many other kinds of learning than that involved in direct sequence extrapolation. Specifically, human learning often involves the discovery of objects in the environment and their effects on experience.

The *roofs-and-boxes* example illustrates that intelligence requires knowing about objects in the world and not just about one's history of sensations—even if one's goal is to predict future sensations.

2 The Roofs and Boxes example

Consider the problem of extrapolating a sequence formed in the following way.

A billiard ball rolls frictionlessly in a rectangular arena and is frictionlessly reflected when it hits the wall with angle of reflection equal to angle of incidence. Inside the arena there are also some rectangular boxes that reflect the ball when it hits their sides. There are also some rectangular roofs that have no effect on the ball but hide it from observation.

A sequence of zeroes and ones is generated by a mechanism that observes the arena from above once per second (or nanosecond if you are impatient). If the ball is under a roof, it is invisible and a zero is generated. Otherwise a one is generated.

Now consider extrapolating this sequence from an initial segment not knowing about the roofs and boxes. None of the techniques of sequence extrapolation studied by the above-mentioned authors is applicable.

If you know that the sequence is generated by roofs and boxes you can try to fit models of the locations of the roofs and boxes. With enough data and computation, you will succeed.

If you don't have the idea of roofs and boxes, you will have to invent it. Donald Michie opined that a good cryptanalyst might come up with the idea. Looking for and analyzing repeated subsequences might help.

3 Making experiments

Here's a variant system that might be easier to analyze, because it lends itself to experiment. Suppose the observer, still seeing only zeroes and ones, has a button he can press. The effect of the button, perhaps unbeknownst to him, is to deflect the ball through an angle of 0.01 degrees. Pressing the button once will affect the sequence but usually only after some time. For a while the ball will be bouncing off the same surfaces.

An observer who has been told or has formed the hypothesis that he is facing a *roofs-and-boxes* problem can locate the roofs and boxes simply but tediously in the case when the sides of the roofs an boxes are parallel to edges of the arena. He presses his button and waits a long time to see if the zeroes and ones form an approximately periodic pattern. If so he has a measure of the distance of a box from the edge and the amount of overhang. If not he moves on until he has such a pattern. After analyzing one such pattern he moves on till he finds another. Eventually he will get the pattern of roofs and boxes and can predict the future.

How clever must one be to hypothesize that it is a *roofs-and-boxes* problem? It is a problem of scientific creativity. A scientist might fiddle for a long time before coming up with the hypothesis. Donald Michie said that a good cryptanalyst would probably succeed.

How hard would it be to write a program that could honestly discover the *roofs-and-boxes* theory?

Roofs-and-boxes illustrates the idea that even to extrapolate experience a robot must know or learn about phenomena in the world. Learning programs have to discover phenomena in the world and not just patterns in the data. To put the matter in another way, important patterns in the data usually take the form of observations of phenomena in the world.

/@steam.stanford.edu:/u/jmc/e98/roofs.tex: begun Sun Aug 2 14:08:24 1998, latexed September 9, 1998 at 12:59 p.m.