Linear Algebra *in* Data Science

An unconventional approach

Norman Matloff

2024-11-27

Table of contents

Preface

Welcome to my magnum opus! :-) I've written a number of books, but consider this one to be the most important.

Subtlety in the Title

Let's start with the title of this book, "Linear Algebra *in* Data Science." It would be more typical of "math support for field X" books to use "for" rather than "in," but the use of the latter aims to emphasize the fact that:

Linear algebra is absolutely fundamental to the Data Science field. For us data scientists, it is "our" branch of math. Mastering this branch, which is definitely within the reach of all, pays major dividends.

Philosophy

This is an unconventional linear algebra textbook, doing everything "backwards." The presentation of each concept begins with a problem to be solved, almost always from Data Science, leading up to a linear algebra solution. Basically, the math sneaks up on the reader, who suddenly realizes they've just learned a new general concept!

Who is this book for?

Of course the book should work well as a course textbook. The "applications first" approach should motivate student, and the use of Quarto enables easy conversion to Powerpoint by instructors.

I hope the book's emphasize on the Why? and How? especially appeals to do-it-yourselfers, those who engagement in self-study is motivated by intellectual curiosity rather than a course grade.

Prerequisite background

Basic data science:

- Calculus.
- Some exposure to R is recommended, but the text can be read without it.
- Basics of random variables, expected value and variance.

For a quick
R, see my fa
first 8 lesso

1 Matrix Multiplication

1.1 A Random Walk Model

Let's consider a *random walk* on {1,2,3,4,5} in the number line. Time is numbered 1,2,3,... Our current position is termed our *state*. The notation $\mathrm{X_k} = \mathrm{i}$ means that at time k we are in state/position i.

Our rule will be that at any time k, we flip a coin. If we are currently at position i, we move to either i+1 or i-1, depending on whether the coin landed heads or tails. The exceptions are k = 1 and k = 5, in which case we stay put if tails or move to the adjacent position if heads.

We can summarize the probabilities with a *matrix*, a two-dimensional array:

$$
P_1 = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix}
$$

For instance, look at Row 2. There are 0.5 values in Columns 1 and 3, meaning there is a 0.5 chance of a move $2 \rightarrow 1$, and a 0.5 chance of a move $2 \rightarrow 3$.

We use a subscript 1 here in P_1, meaning "one step." We go from, say, state 2 to state 1 in one step with probability 0.5. P_1 is called the *one-step transition matrix* (or simply the *transition matrix*) for this process.

What about the two-step transition matrix P_2? From state 3, we could go to state 1 in two steps, by two tails flips of the coin. The probability of that is $0.5^2 = 0.25$. So the row 3, column 1 element in P_2 is 0.25. On the other hand, if from state 3 we flip tails then heads, or heads then tails, we are back to state 3. So, the row 3, column 3 element in P_2 is $0.25 + 0.25 = 0.5$.

The reader should verify the correctness here: