Data Science from Scratch

First Principles with Python

Stephen CUI

March 20, 2023

Chapter 1

Linear Algebra

1.1 Vectors

Abstractly, vectors are objects that can be added together to form new vectors and that can be multiplied by scalars (i.e., numbers), also to form new vectors.

Concretely (for us), vectors are points in some finite-dimensional space. Although you might not think of your data as vectors, they are often a useful way to represent numeric data.

The simplest from-scratch approach is to represent vectors as lists of numbers. A list of three numbers corresponds to a vector in three- dimensional space, and vice versa.

We'll accomplish this with a type alias that says a Vector is just a list of floats.

Vectors add componentwise(分量式). This means that if two vectors v and w are the same length, their sum is just the vector whose first element is `v[0] + w[0]`, whose second element is `v[1] + w[1]`, and so on. (If they're not the same length, then we're not allowed to add them.)

For example, adding the vectors `[1, 2]` and `[2, 1]` results in `[1 + 2, 2+ 1]` or `[3, 3]`.

We can easily implement this by zip-ing the vectors together and using a list comprehension to add the corresponding elements.

We'll also sometimes want to componentwise sum a list of vectors—that is, create a new vector whose first element is the sum of all the first elements, whose second element is the sum of all the second elements, and so on.

We'll also need to be able to multiply a vector by a scalar, which we do simply by multiplying each element of the vector by that number.

A less obvious tool is the dot product. The dot product of two vectors is the sum of their componentwise products. If w has magnitude 1, the dot product measures how far the vector v extends in the w direction. Another way of saying this is that it's the length of the vector you'd get if you projected v onto w.

Using this, it's easy to compute a vector's sum of squares, which we can use to compute its magnitude (or length)

We now have all the pieces we need to compute the distance between two vectors, defined as:

$$\sqrt{(v_1 - w_1)^2 + \cdots + (v_n - w_n)^2}$$

Using lists as vectors is great for exposition but terrible for performance.
In production code, you would want to use the NumPy library, which includes a high-performance array class with all sorts of arithmetic operations included.

1.2 Matrices

A matrix is a two-dimensional collection of numbers. We will represent matrices as lists of lists, with each inner list having the same size and representing a row of the matrix. If `A` is a matrix, then `A[i][j]` is the element in the ith row and the jth column. Per mathematical convention, we will frequently use capital letters to represent matrices.

Matrices will be important to us for several reasons.

1. we can use a matrix to represent a dataset consisting of multiple vectors, simply by considering each vector as a row of the matrix.
2. we can use an $n \times k$ matrix to represent a linear function that maps k-dimensional vectors to n-dimensional vectors.
3. matrices can be used to represent binary relationships.