Session 07: Statistics fundamentals

wifi: GA-Guest, yellowpencil

GENERAL ASSEMBLY

```
cd ~/Documents/ga-ldn-ds37
git commit -am “your commit message here”
git pull
```

Today’s session plan

1800-1820	Standup & Review
1820-1845	Linear algebra review
1845-1900	Linear algebra in machine learning
1900-1920	Break
1920-2000	Descriptive statistics fundamentals
2000-2100	Exercises
Homework: Linear algebra and numpy practise	

At the end of the session, you will be able to ...

Identify a normal distribution within a dataset

Compute summary statistics using numpy

Compute dot products, vector norms and matrix multiplication by hand and with numpy

Data Science Part Time

Review

Computers Out:

Pandas review

Open the notebook ds37-07-01.ipynb and work through the exercises.

Data Science Part Time

Linear algebra

What’s linear algebra?

Linear algebra is a branch of mathematics that deals with linear equations, including the use of vectors and matrices to solve linear problems in high dimensional space.

Let’s explore what we mean by this.

Scalars, vectors and matrices

A scalar is a single number or quantity.

A vector is an ordered sequence of numbers.

A matrix is a rectangular array of numbers,

Vectors

We usually represent vectors as lowercase single letters with an arrow.

$$\vec{u} = \begin{bmatrix} 1 & 3 & 7 \end{bmatrix}$$

Matrices

A matrix is a rectangular array of numbers with **m** rows and **n** columns. Each number in the matrix is an entry. Entries can be denoted a_{ij} where **i** denotes the row number and **j** denotes the column number. Note that, because each entry is a lowercase single letter, a matrix is an array of scalars:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Computers Out:

Vectors and matrices in Numpy

Let's open up ds37-07-02.ipynb to start creating vectors and matrices in Numpy.

Sigma notation

We can perform calculations between scalars, vectors and matrices.

However, the rules for how we do this are slightly different compared to when we're working with scalars (single numbers) only.

To understand how to perform some of these calculations, we first need to understand **sigma notation**.

Sigma notation is a method used to write out a long sum in a concise way.

In some ways, it's a bit like a **for loop**.

Sigma notation

Imagine we're adding a sequence of numbers, where there's a clear pattern in the sequence:

$$1+2+3+4+5$$

A shorter way of writing this is to let i represent the number in the sequence and write:

$$\sum_{i=1}^{5} i$$

Sigma notation

Let’s break down what we mean here: this is the sum of i from i=1 up to i=5

$$\sum_{i=1}^{5} i$$

Sigma notation

Now let's try reading and expanding a sum that's written using sigma notation.

$$\sum_{n=1}^{5} n^2$$

This is the sum of n^2 from $n=1$ up to $n=5$ or:

$$1^2 + 2^2 + 3^2 + 4^2 + 5^2$$

$n=1$ $n=2$ $n=3$ $n=4$ $n=5$

Solo Exercise:

Sigma notation

Evaluate the following sums by hand and **then by writing for-loops** in Python:

(a) $\sum_{n=1}^{5} n^3$ (b) $\sum_{n=1}^{5} 3^n$ (c) $\sum_{r=1}^{4} (-1)^r r^2$ (d) $\sum_{k=1}^{4} \frac{(-1)^{k+1}}{2k+1}$

Sigma notation

Now imagine we're adding the elements of the vector u = [2, 4, 6, 8, 10]. We can rewrite this sum as:

$$u_1 + u_2 + u_3 + u_4 + u_5$$

A shorter way of writing this is to let i represent the position or index of the number in the sequence* and write:

$$\Sigma^5_{i=1}\, u_i$$

Sigma notation

Let’s break down what we mean here: this is the sum of the elements of u from element i=1 up to element i=5

$$\sum_{i=1}^{5} u_i$$

Solo Exercise:

Using sigma notation

By hand, work out the sum of the following series **and then using a for-loop in Python**.

v = [2, 4, 5, 7, 8]

w = [9, 1, 1, 0, 5]

$$\sum_{i=1}^{n} v_i w_i$$

Vector addition and subtraction

We sum or subtract the corresponding elements of a vector.

$$\vec{v} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}, \vec{w} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$\vec{v} + \vec{w} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1+1 \\ 3+0 \\ 7+1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 8 \end{bmatrix}$$

Scalar multiplication

We scale a vector with scalar multiplication, multiplying a vector by a scalar (single quantity)

$$2 \cdot \vec{v} = 2\begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 \\ 2 \cdot 3 \\ 2 \cdot 7 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 14 \end{bmatrix}$$

Scalar multiplication

The **dot product** of two n-dimensional vectors is:

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} v_i w_i$$

Solo Exercise:

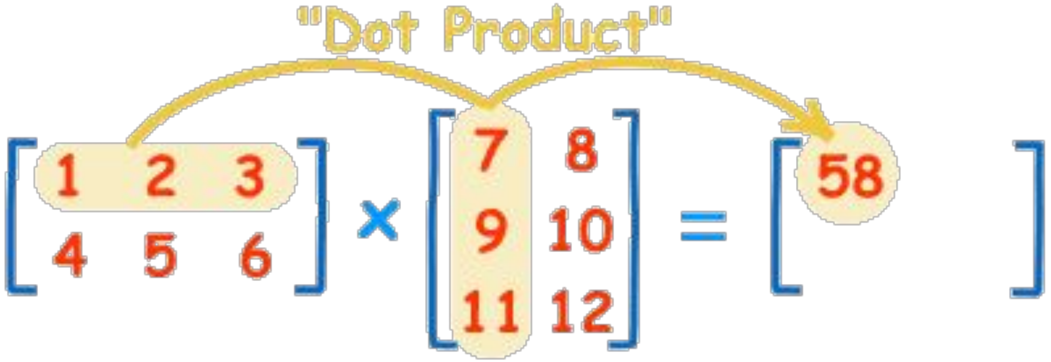
Dot products

Calculate the dot product of these two vectors:

$$\vec{v} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}, \vec{w} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Matrix multiplication

Matrix multiplication is valid when the left matrix has the same number of columns as the right matrix has rows. Each entry is the dot product of corresponding row and column vectors.



Solo Exercise:

Matrix multiplication

Calculate the following **by hand**, and then using Python.

$$\vec{a} = \begin{bmatrix} 5 \\ 8 \\ 2 \end{bmatrix} \quad \vec{b} = \begin{bmatrix} 6 \\ 0 \\ 5 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1 & 7 & 1 \\ 7 & 8 & 4 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 2 & 4 \\ 3 & 1 \\ 1 & 4 \end{bmatrix}$$

(a) $a + b$

(b) $a - b$

(c) $3b$

(d) $a \cdot b$

(e) $C \times D$

Vector norms

The **size** of a vector is found by calculating the vector **norm**.

$$||\vec{v}|| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}$$

Solo Exercise:

Vector norms

On paper, show that

$$||\vec{v}|| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2} = \sqrt{v^T v}$$

Where $\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ and v^T is the **transpose** of v, the **row** vector $[v_1, v_2, v_3, \ldots, v_n]$

Solo Exercise:

Distance between points

Draw a set of axes and the points: $p_1 = [1,2]$ and $p_2 = [4,6]$

Calculate the straight line distance between these two points.

Now imagine our points are in 3D, with z coordinates as well as x and y coordinates.

$p_1 = [1,2,1]$ and $p_2 = [4,6,3]$

What’s the straight line distance now?

Distance between two points

You probably used something like this formula to work out the distance between the two points

$$\|\vec{p_1} - \vec{p_2}\| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

We calculate the prediction **error** of a model using the exact same formula.

Solo Exercise:

How wrong is your model?

Imagine you have a model that predicts house prices.

For two houses with **actual values** (in thousands of pounds) a = [100, 75] the model makes predictions p = [80, 120].

What’s the **size** of the error of the model?

Solo Exercise:

How wrong is your model?

Now imagine that our same model makes **five** predictions.

For five houses with **actual values** (in thousands of pounds) a = [100, 75, 240, 375, 80] the model makes predictions p = [80, 120, 250, 350, 95].

What's the **size** of the error of the model now?

How wrong is your model?

Asking ‘what’s the **size** of the error of our model’ is the same as asking ‘what’s the **distance** between our model’s predictions and the actual values’. We often use the **mean squared error** to show the distance between our **model’s predictions (y-hat)** and the **actual values** (y).

$$MSE = \frac{1}{n} \| \hat{y}(\mathbf{X}) - \vec{y} \|^2$$

Solo Exercise:

A simple model

Imagine I’ve given you five chocolate bars*

Their weights are (in g): 101, 98, 120, 70, 75

What’s your best guess for the weight of the next chocolate bar I’ll give you?

How would you write this mathematically?

*I am not going to give you any chocolate bars

Intro to Python

Let’s Review

At the end of the session, you will be able to ...

Identify a normal distribution within a dataset

Compute summary statistics using numpy

Compute dot products, vector norms and matrix multiplication by hand and with numpy

Coming up next session...

- Designing experiments, missing data, hypothesis testing