Getting started: Brief Review of Linear Algebra

Linear Algebra is a big branch of mathematics concerning linear equations and their representations in vector spaces and through matrices. If you are watching this course, then you have already taken a Linear Algebra course.

The following is a brief review of concepts needed in the context of Data Science and Machine Learning. For a complete review, please follow the sources provided at the end of this document.

If there are any concepts in this document that you are unfamiliar or *rusty* about, please review them as soon as possible.

Vectors

A **vector** is the fundamental building block for Linear Algebra!

There are many ways of interpreting a vector:

- Ordered set of numbers (*data points*)
- Arrows pointing in space, with some length and direction (*geometrical representation*)

In Machine Learning and Data Science, a vector is often characterized as an ordered set of numbers.

For example, we can have a n-dimensional vector that contains the number of positive COVID-19 cases in Florida in the last n days.

Vector Operations

Some vector operations include:

- Addition
- Subtraction
- Scalar multiplication

Consider the vectors, $\mathbf{a}$ and $\mathbf{b}$, and the scalar c.

$$\mathbf{a} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \text{ and } c = \frac{1}{2}$$

Let's look at these operations in the virtual whiteboard.

- Inner product: $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{b} = a_1b_1 + a_2b_2 + \ldots + a_nb_n$
- Outer product: $\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1 \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{bmatrix}$

It is *conventional* to write vectors as vertical vectors.

Norms

The **L-p norm** of a vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ is:

$$\|\mathbf{x}\|_p = \left(x_1^p + x_2^p + \cdots + x_n^p\right)^{1/p}$$

- Different p norms have different (geometrical) properties. For example, the L-2 norm is commonly referred to as the **Euclidean norm**, and is denoted as $\|\mathbf{x}\|_2$ or simply $\|\mathbf{x}\|$.
  - The L-2 norm computes the *length* of the vector: $\text{length}(\mathbf{x}) = \|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$

- The L-2 norm relates to inner products by the **Cauchy-Schwarz inequality**:

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

- Within the Eulidean geometry, the **triangle inequality** is also a useful property that uses L-2 norm:

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Distance

- The **Euclidean distance** between two nonzero vectors $\mathbf{x}$ and $\mathbf{y}$ is:

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$$

It measures the length of the *shortest* straight-line between two points.

- The cosine distance (formally **cosine similarity**) measures the angle between two vectors:

$$d(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \cos(\theta), \text{ where } \theta = \angle(\mathbf{x}, \mathbf{y})$$

- The **spherical distance** between two vectors $\mathbf{x}$ and $\mathbf{y}$ is given by $R \times \angle(\mathbf{x}, \mathbf{y})$, where R is the radius of the sphere.

Unit vector

In geometric representations, a vector is often characterized by its *direction* and *length*.

When we are interested in only the direction of the vector, we usually *normalize* the vector by its L-2 norm.

- We say that a vector $\mathbf{x}$ is a **unit vector** if $\|\mathbf{x}\| = 1$.

- If $\|\mathbf{x}\| \neq 1$, we can create a unit vector in the same direction as $\mathbf{x}$ as: $\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$

Vector Correlation

- The **vector (Pearson's) correlation** between $\mathbf{x}$ and $\mathbf{y}$ is

$$r = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

(the same as cosine similarity!)

Vector Projection

- The **projection** of $\mathbf{y}$ onto $\mathbf{x}$ is defined as:

$$\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|}$$

- The **projection** of $\mathbf{x}$ onto $\mathbf{y}$ is defined as:

$$\frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{y}\|}$$

Orthogonal and Orthonormal vectors

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** if their inner product is zero:

$$\mathbf{x}^T\mathbf{y} = 0 \Rightarrow \mathbf{x} \perp \mathbf{y}$$

- Two vectors are $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are **orthonormal** if they are orthogonal and have unit norm.

- We say that $\{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}\}$ is a set of orthonormal vectors.

Span and Basis Vectors

Consider, for example, the vectors $\mathbf{x} = [1, 0]$ and $\mathbf{y} = [0, 1]$.

- The vector space $\mathcal{S} = \{\mathbf{x}, \mathbf{y}\}$ is a *spanning set* for $\mathbb{R}^2$, or $\mathcal{S}$ *spans* $\mathbb{R}^2$.

- Note that we can represent *any* vector in $\mathbb{R}^2$ as a linear combination of the vectors $\mathbf{x}$ and $\mathbf{y}$.

- The **dimension** of a vector space $\mathcal{S}$ is the cardinality (i.e. number of vectors) of a basis of $\mathcal{S}$.
  - A minimum of 2 spanning vectors are required to represent everything in $\mathbb{R}^2$. So the dimension of $\mathbb{R}^2$ is 2.
  - Since the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = 2$, $\mathcal{S}$ is *minimal*.

- We say that $\mathbf{S}$ is a *minimal spanning set* or a **basis** of $\mathbb{R}^2$.

- Since the vectors in $\mathbf{S}$ are orthonormal, we that $\mathcal{S}$ is an orthonormal basis of $\mathbb{R}^2$.

Matrices

Matrices are a generalization of vectors. One way to interpret matrices is: rectangular arrays or ordered numbers.

Example: Suppose you are describing 3 houses in terms of their squared footage, average number of rooms and age.

You can put this information in a matrix form: $\begin{bmatrix} 1214 & 4 & 65 \\ 2325 & 6 & 68 \\ 1710 & 4 & 71 \end{bmatrix}$

Special Matrices

- **Identity matrix**: $\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$

- **Diagonal matrix**: any matrix that can be written as the product of a constant with the identity matrix, $\alpha\mathbf{I} = \begin{bmatrix} \alpha & 0 & \cdots & 0 \\ 0 & \alpha & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha \end{bmatrix}$

Given the matrix $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, the vector $\mathbf{x} = [-1, 5]^T$ and the scalar d:

- Scalar-Matrix multiplication: $d\mathbf{A} = \mathbf{A}d = \begin{bmatrix} d & 2d \\ 3d & 4d \end{bmatrix}$

- Vector-Matrix multiplication: $\mathbf{Ax} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 5 \end{bmatrix} = \begin{bmatrix} 1(-1) + 2(5) \\ 3(-1) + 4(5) \end{bmatrix} = \begin{bmatrix} 9 \\ 17 \end{bmatrix}$

Which operations are valid?:

1. $\mathbf{xA}$
2. $\mathbf{x}^T\mathbf{A}$
3. $\mathbf{x}^T\mathbf{Ax}$

- Matrix-Matrix multiplication:
$\mathbf{AA} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1(1) + 2(3) & 1(2) + 2(4) \\ 3(1) + 4(3) & 3(2) + 4(4) \end{bmatrix} = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix}$

Matrix Operators:

- The **determinant** of a square $n \times n$ matrix $\mathbf{A}$ is a unique number. It measures the scaling factor by which the linear transformation $\mathbf{A}$ changes any area or volume.

The determinant of $\mathbf{A}$ is denoted by $|\mathbf{A}|$ or $\det(\mathbf{A})$. In Python, we can compute the determinant of a matrix using the module `numpy.linalg`.

- Consider the vectors stacked vertically on the matrix $\mathbf{A}$. If *any* of these vectors can be written as a linear combination of any other/s, then we say that $\mathbf{A}$ has **linearly dependent columns**.

- If the matrix $\mathbf{A}$ has *linearly dependent* columns (or rows) then $\det(\mathbf{A}) = 0$.

- If the matrix $\mathbf{A}$ has **linearly independent columns**, then $\mathbf{A}$ is said to be **left-invertible**, that is there exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$

- Similarly, if the rows of $\mathbf{A}$ are linearly independent, then $\mathbf{A}$ is **right-invertible**, that is there exists a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$

- The **trace** of a square $n \times n$ matrix $\mathbf{A}$ is defined as the sum of the elements on the main diagonal of $\mathbf{A}$:

$$\operatorname{trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Linear Vector Function or *Affine Function*

A linear matrix-vector multiplication function is defined as

$$\begin{aligned} \mathbf{f} : \mathbb{R}^n &\longrightarrow \mathbb{R}^m && (1) \\ \mathbf{x} &\longmapsto \mathbf{A}\mathbf{x} && (2) \end{aligned}$$

That is $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a $m \times n$ matrix.

- Since we are dealing with data, we generally start by assuming that the data is generated under some unknown **model** and then we pick a form for that model. For example, the linear vector function is a **linear model**.

Systems of Linear Equations

- One of the most important applications of linear vector functions is to solve *linear systems of equations*

- The system of linear equations can be written concisely in matrix notation as

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

  where $\mathbf{x} = [x_1, x_1, \cdots, x_n]^T$ is the vector of variables (unknowns).

- For example, consider the polynomial $p(\mathbf{x}) = c_0 + c_1\mathbf{x} + c_2\mathbf{x}^2 + \ldots c_p\mathbf{x}^p$. We can write it in the form of $\mathbf{A}\mathbf{c} = \mathbf{y}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}, \mathbf{c} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} \text{ and } \mathbf{y} = p(\mathbf{x})$$

where

- $\mathbf{A}$ is the p^{th}-order polynomial representation of the data points $\mathbf{x}$ (e.g., house square footage, age and number of rooms)
- $\mathbf{y}$ is the output (e.g. house price)
- $\mathbf{c} = [c_0, c_1, c_2, \cdots, c_p]^T$ are unknown coefficients.

Least Squares Solution

If $\mathbf{A}$ is invertible then

$$\mathbf{c} = \mathbf{A}^{-1}\mathbf{y}$$

- Suppose that $p = 1$ and $n >> p$, then $A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$

We only have 2 independent vectors, so they do not form a basis for $\mathbb{R}^n$.

In general, these problems do not have a solution, and we cannot use the matrix inverse to find the solution.

- Geometrically, you can write each row of $\mathbf{A}$ as a line and you will find out that all n lines will not intercept at the same point.

The **least squares solution** minimizes the sum of the squared errors:

$$f(\mathbf{c}) = \|\mathbf{Ac} - \mathbf{y}\|^2$$

- Then any minima $\hat{\mathbf{c}}$ will satisfy:

$$\frac{\partial f}{\partial c_i}(\hat{\mathbf{c}}) = 0, \quad i = 1, 2, \ldots, n$$

Or using gradient notation

$$\nabla f(\hat{\mathbf{c}}) = 0$$

By taking the derivative of $f(\mathbf{c})$ and solving for $\mathbf{c}$, the **least squares solution** is then given by:

$$\mathbf{c} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}$$

Where $\mathbf{A}^\dagger = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$ is the **pseudo-inverse** (or **Moore–Penrose inverse**).

Eigendecomposition

Every $n \times n$ matrix is a *linear transformation*.

For each $n \times n$ matrix, there are *special* vectors called **eigenvectors** that only get scaled by the linear transformation, that is, consider a vector $\mathbf{v}$ that satisfies:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

All vectors $\mathbf{v}$ that satisfy this equation are called **eigenvectors** and the λ values are called **eigenvalues**.

- This equation is called the **characteristic equation**

If $\mathbf{A}$ is an $n \times n$ matrix with linearly independent rows, then $\mathbf{A}$ has n **orthogonal** eigenvectors, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

- When we consider the normalized eigenvectors: $\tilde{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}$
- This means that the vector space $\mathcal{S} = \{\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \cdots, \tilde{\mathbf{v}}_n\}$ forms a **basis** of $\mathbb{R}^n$.

Using the characteristic equation and matrix multiplication, we can write:

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{\Lambda}$$

where $\mathbf{V} = \begin{bmatrix} \mathbf{v}_0 & \mathbf{v}_1 & \cdots & \mathbf{v}_{n-1} \end{bmatrix}$ is the *modal matrix*, which has the eigenvectors as its columns, and

$\mathbf{\Lambda}$ is a diagonal matrix that contains the associated eigenvalues, $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$.

- This is called the **Karhunen–Loève Transform (KLT)**

Further review

- "Chapter 2 - Linear Algebra" from the Deep Learning book by Ian GoodFellow et al., MIT Press, 2016.
- 3Blue1Brown, "Essence of Linear Algebra" YouTube series
- Stephen Boyd and Lieven Vandenberghe, "Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares" book, available online
- Gilbert Strang, 18.06 MITOpenCourseWare "Linear Algebra", link