Applied Linear Algebra in Data Analysis: Course Notes

Sivakumar Balasubramanian
CMC Vellore

Update on July 10, 2024

Contents

CONTENTS

List of Figures

Part I

Linear Algebra

Chapter 1

Vectors

1.1 n-Vectors

A collection of an ordered list of n numbers is called an n-vector. We will use bold lower case alphabets to represent such vectors, and we will represent these as a column of numbers, which is referred to as a *column vector*. We will look at *row vectors* at a later stage. Consider the following example:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The elements of the n-vector $x_1, x_2, \ldots, x_n$ are called the *components* of the vector $\mathbf{x}$; x_i is the i^{th} component of the vector $\mathbf{x}$. If these components are all real numbers, the set of all such n-vectors is the set $\mathbb{R}^n$.

Where do we come across such n-vectors? In many places, such as in physics, engineering, economics, medicine, etc. Any application where we deal with multiple pieces of information that can be represented as a list of numbers can be represented as an n-vector. When we deal with systems with multiple inputs, multiple output, or multiple states, we can represent these as n-vectors. We talk about the state of a system in a later chapter.

1.1.1 Some common n-vectors

We will often come across some special n-vectors in this document course and in many applications. We will define some of these vectors here.

- **Zero vector:** The n-vector whose components are all zeros is called the *zero vector*. $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$
- **One vector:** The n-vector whose components are all ones is called the *one vector*. $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- **Unit vectors:** The n-vectors whose components are all zeros except for one component which is 1. These are called the *standard basis vectors* and are denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. The n-vector $\mathbf{e}_i$ has all

components as zeros except for the i^{th} component which is 1. For example, the unit vectors in $\mathbb{R}^2$ are:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

We now look at some examples of n-vectors that we come aross in applications.

Example 1.1. Basic clinical information during a hospital visit. When a patient visits a hostpial, several clinically relevant variables are captured, for instance:

Index	Variable	Units
1	Sex	None (0: Male, 1: Female)
2	Age	Years
3	Height	cm
4	Weight	kg
5	Heart rate	count
6	Systolic pressure	mm of Hg
7	Diastolic pressure	mm of Hg
8	Temperature	celcius

The following are some examples of n-vectors generated from three different patients visiting the hospital.

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 67 \\ 152 \\ 56 \\ 132 \\ 102 \\ 37.1 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 36 \\ 172 \\ 97 \\ 156 \\ 97 \\ 36.5 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 22 \\ 162 \\ 56 \\ 121 \\ 78 \\ 38.2 \end{bmatrix}$$

Example 1.2. Time series data. We often collect data over time, often at regular intervals. For example, consider the example of an attending nurse taking the temperature of a patient admitted to the hospital for an infectious disease. The nurse records the temperature of the patient every hour, without fail for the 48 hours the patient spent in the hosptial. This temperature record will have a total of 49 measurements, which can conveniently think of as a n-vector, in this case a 49-vector. Instead of writing down the entire 49-vector, we depict it as a time series plot.

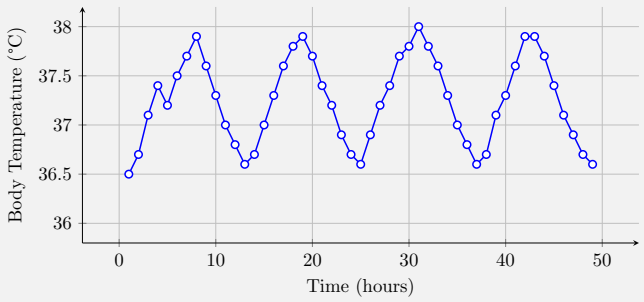


Figure 1.1: Body temperature recorded at multiple time points.

1.2 Visualizing n-vectors

The n-vectors can be visualized as points in n-dimensional space. For example, A 1-vector or just single real number or a *scalar* can be thought of as a point on the real line. The 1-vector $x = 2.45$ is shown in Figure 1.2 is the red point. But we will find it useful to visualize a 1-vector as an arrow starting at the origin and ending at the point on the real line. The arrow is shown in blue in Figure 1.2.

The elements of $\mathbb{R}^2$ are points on the plane, and we can visualize them as points in the plane. The 2-vectors $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ are shown in Figure 1.3a. A similar visualization is shown for $\mathbb{R}^3$ (Figure 1.3b), and for $\mathbb{R}^4$ and beyond you simply pretend that you can visualize things in your head like your instructor does.

1.3 Some Commonly Used n-vectors

We will now define a some commonly used n-vectors that we will use in the course.

- **Zero vector:** The n-vector whose components are all zeros is called the *zero vector*. $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$
- **One vector:** The n-vector whose components are all ones is called the *one vector*. $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- **Unit vectors:** The n-vectors whose components are all zeros except for one component which is 1. These are called the *standard basis vectors* and are denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. The n-vector $\mathbf{e}_i$ has all components as zeros except for the i^{th} component which is 1. For example, the unit vectors in $\mathbb{R}^2$ are:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

1.4 Operations on n-vectors

There are many operations we can perform on n-vectors, but we will only focus on two operations for this:

- **Scalar multiplication:** Given a scalar $c \in \mathbb{R}$ and an n-vector $\mathbf{x}$. The scalar multiplication operation produces another n-vector $c\mathbf{x}$ whose components are $cx_1, cx_2, \ldots, cx_n$.

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \longrightarrow 2\mathbf{x} = \begin{bmatrix} 2\,(1) \\ 2\,(2) \end{bmatrix} = \begin{bmatrix} 2 \\ 8.2 \end{bmatrix}$$

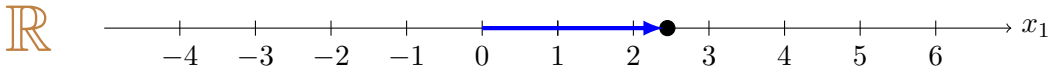
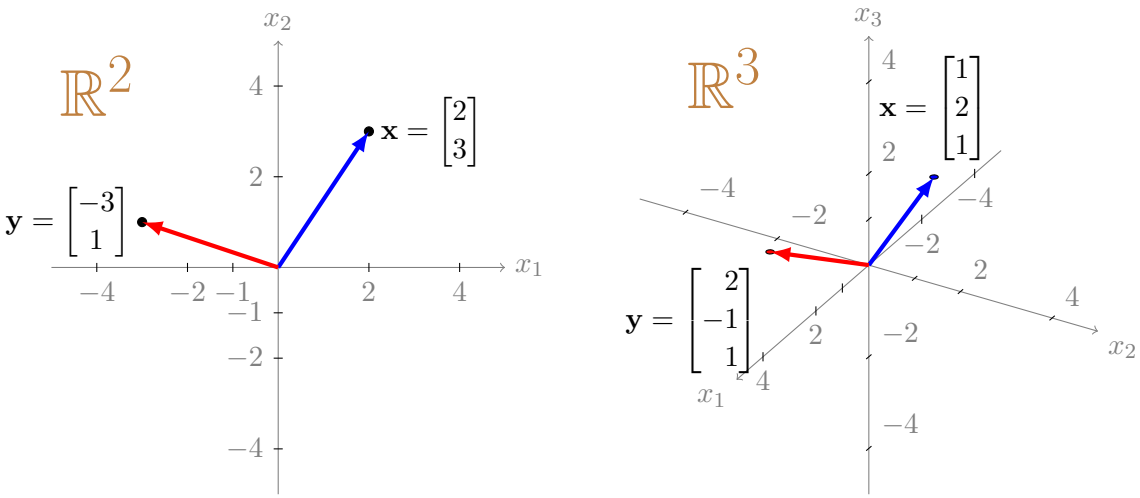


Figure 1.2: The real line $\mathbb{R}$ contains the 1-vectors.



(a) The $\mathbb{R}^2$ plane contains the 2-vectors. **(b)** $\mathbb{R}^3$ contains the 3-vectors.

Figure 1.3: The $\mathbb{R}^2$ and $\mathbb{R}^3$ sets.

The geometric interpretation scalar multiplication is shown in Figure 1.4. Scalar multiplication stretches or shrinks the vector without rotating it. When the scalar is positive the direction of the scaled vector is the same as the original vector, and when the scalar is negative the direction is opposite. When the scalar is zero, the scaled vector is the zero vector $\mathbf{0}$.

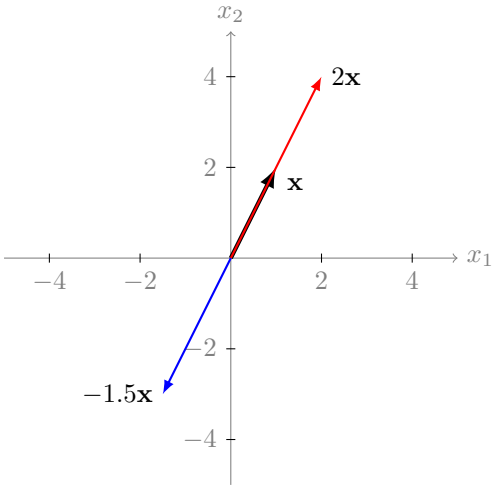


Figure 1.4: Scalar multiplication of a vector.

- **Vector Addition:** Given two n-vectors $\mathbf{x}$ and $\mathbf{y}$, the vector addition operation, represented by $\mathbf{x}+\mathbf{y}$, producs another n-vector whose components are $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$.

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \longrightarrow \mathbf{x} + \mathbf{y} = \begin{bmatrix} 1+2 \\ 3+1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

The geometric interpretation the vector addition operation is shown in Figure 1.5. Geometrically, the vector addition operation follows the parallelogram law of addition, where the resulting vector $\mathbf{x}+\mathbf{y}$ is a diagonal of the parallelogram formed by the two vectors $\mathbf{x}$ and $\mathbf{y}$. Another way to think about this, is that you first move along $\mathbf{x}$ to its end point, and starting from there then move along $\mathbf{y}$ to its end point or vice versa.

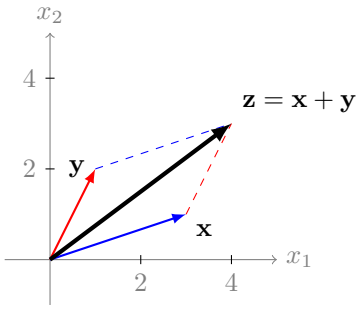


Figure 1.5: Vector addition.

You can add more than two vectors to obtain a new vector, like below:

$$\mathbf{w} = \mathbf{x} + \mathbf{y} + \mathbf{z}$$

Geometrically, we can first apply the parallelogram law to $\mathbf{x}$ and $\mathbf{y}$, and then apply the parallelogram law to $\mathbf{x} + \mathbf{y}$ and $\mathbf{z}$ to get $\mathbf{w}$.

1.5 Vector spaces

Vector spaces are *sets* with some special properites. More specifically, a vector space is a set V of elements called *vectors* that are closed under two operations called *addition* and *scalar multiplication*. This simply means that if you perfom these operations using elements from the set V, the result is also an element of the set V. A vector space must satisfy the following properties:

- **Closure under addition:** For any two vectors $\mathbf{x}, \mathbf{y} \in V$, the sum $\mathbf{x} + \mathbf{y} \in V$.
- **Closure under scalar multiplication:** For any scalar $c \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, the product $c\mathbf{x} \in V$.
- **Additive identity:** There exists a vector $\mathbf{0} \in V$ such that for any vector $\mathbf{x} \in V$, $\mathbf{x} + \mathbf{0} = \mathbf{x}$.
- **Additive inverse:** For any vector $\mathbf{x} \in V$, there exists a vector $-\mathbf{x} \in V$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
- **Commutativity of addition:** For any two vectors $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
- **Associativity of addition:** For any three vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
- **Distributive property:** For any scalar $c \in \mathbb{R}$ and any two vectors $\mathbf{x}, \mathbf{y} \in V$, $c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$.
- **Distributive property:** For any two scalars $c, d \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, $(c + d)\mathbf{x} = c\mathbf{x} + d\mathbf{x}$.
- **Associativity of scalar multiplication:** For any two scalars $c, d \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, $(cd)\mathbf{x} = c(d\mathbf{x})$.
- **Multiplicative identity:** For any vector $\mathbf{x} \in V$, $1\mathbf{x} = \mathbf{x}$.

These properties are satisfied by the set $\mathbb{R}^n$ of n-vectors, and hence $\mathbb{R}^n$ is a vector space. Geometrically, the concept of a vector space corresponds to flat spaces that contain the origin. This will become more clear when we talk about subspaces. Notice that definition of the vector space given above does not make any specific mention of n-vectors. The definition is general and can be applied to any set of elements that satisfy the properties listed above. The following are some examples of vector spaces with the addition and scalar multiplication operations defined on them.

Example 1.3. Set of $m \times n$ matrices. The set M of all $m \times n$ matrices of real numbers is a vector space.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}, \ a_{ij} \in \mathbb{R}$$

We define scalar multiplication and addition of matrices as follows:

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{11} & ca_{12} & \cdots & ca_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{11} & ca_{12} & \cdots & ca_{1n} \end{bmatrix}, \ c \in \mathbb{R}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \end{bmatrix}, \ \mathbf{A}, \mathbf{B} \in M$$

Since each element of $c\mathbf{A}$ and $\mathbf{A} + \mathbf{B}$ is a real number, M is a vector space.

Example 1.4. Set of polynomials of order $\leq n$. Now we look at strange example of a vector space. The set P_n of all polynomials of degree at most n with real coefficients, defined over an interval $[a, b]$.

$$p(x) = \sum_{k=0}^{n-1} a_k x^k, \ x \in [a, b], \ a_k \in \mathbb{R}$$

The set P_n contains all polynomials of the form shown above. We define scalar multiplication and addition of polynomials as follows:

$$cp(x) = c \sum_{k=0}^{n-1} a_k x^k = \sum_{k=0}^{n-1} c a_k x^k, \ p(x) \in P$$

$$p(x) + q(x) = \sum_{k=0}^{n-1} a_k x^k + \sum_{k=0}^{n-1} b_k x^k = \sum_{k=0}^{n-1} (a_k + b_k) x^k, \ p(x), q(x) \in P_n$$

The set P_n is a vector space because the sum and product of any two polynomials from P_n is also a polynomial of degree at most n with real coefficients.

Example 1.5. Set of continuous functions. The set $C[0, 1]$ of all continuous functions $f(x)$ over the time interval $x \in [0, 1]$ is a vector space. We define scalar multiplication and addition of functions as follows:

$$cf(x) = cf(x), \ f(x) \in C(0, 1)$$

$$f(x) + g(x) = f(x) + g(x), \ f(x), g(x) \in C(0, 1)$$

The set $C(0, 1)$ is a vector space because the sum and product of any two continuous functions from $C(0, 1)$ is also a continuous function.

1.6 Subspaces – "Little" Vector Spaces

These are little subspaces in the sense that they are subsets of a larger vector space that are themselves vector spaces. More formally, a subspace U of a vector space V is a subset of V that is itself a vector space. The subspace U of the vector space V must satisfy the following properties:

- **Closure under addition:** For any two vectors $\mathbf{x}, \mathbf{y} \in U$, the sum $\mathbf{x} + \mathbf{y} \in U$.
- **Closure under scalar multiplication:** For any scalar $c \in \mathbb{R}$ and any vector $\mathbf{x} \in U$, the product $c\mathbf{x} \in U$.

One immediate consequence of the above definition is that hte zero element of the vector space V must be present in every subspace of V. If the zero element is not in a subset, then it cannot be a subspace. Geometrically subspace are flat structures (or surfaces or manifolds) in $\mathbb{R}^n$ (or the parent vector space) that contain the origin, and extends infinitely. Let's look at some examples of subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$, which are easier to visualize.

Example 1.6. A straight line through the origin. We know that $\mathbb{R}^2$ is a vector space. Now consider the set of all points in $\mathbb{R}^2$ that lie on a straight line passing through the origin, defined as follows:

$$S = \left\{ \mathbf{x} \,:\, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2,\, x_1 = m \cdot x_2,\, m \in \mathbb{R} \right\}$$

How do we verfiy this is a subspace of $\mathbb{R}^2$? The definition above shows that any x in S comes from $\mathbb{R}^2$, which means its a subset of $\mathbb{R}^2$. Figure 1.6a shows the set S for $m = 2$. How do we verify if S is a subspace of $\mathbb{R}^2$? We need to now verify that S satisfies the properties of a vector space.

1. First, let's check if S contains the zero vector. If it does not contain the zero vector, then it cannot be a subspace. The elements from S are of the form $\begin{bmatrix} x \\ mx \end{bmatrix}$, thus if we choose $x = 0$, then we get $\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in S$. So, S contains the zero vector. This means that S can be a subspace space of $\mathbb{R}^2$.
2. Let's verify vector scaling. Scaling the element $\begin{bmatrix} x \\ mx \end{bmatrix} \in S$ by a scalar c we get,

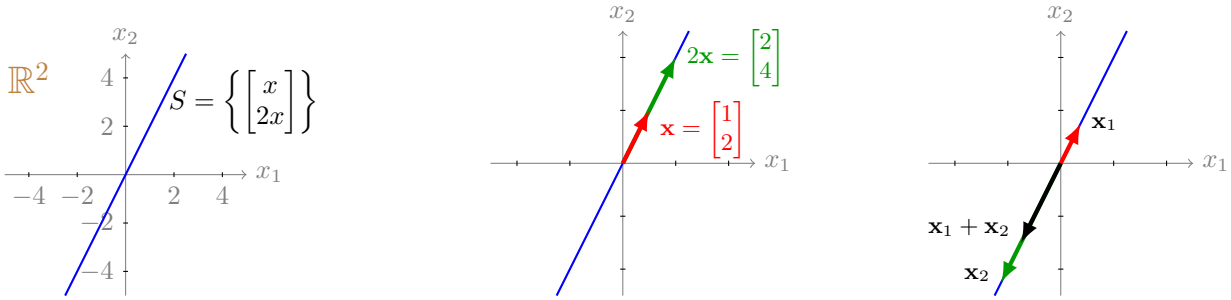
   $$c \begin{bmatrix} x \\ mx \end{bmatrix} = \begin{bmatrix} cx \\ cmx \end{bmatrix} = \begin{bmatrix} cx \\ m\,(cx) \end{bmatrix} = \begin{bmatrix} y \\ my \end{bmatrix}, \quad \text{where } y = cx \in \mathbb{R}$$

   This means that $c \begin{bmatrix} x \\ mx \end{bmatrix}$ belongs to S, this the set S is closed under scalar multiplication. This still means that S can be a subspace of $\mathbb{R}^2$.
3. Let's verify vector addition. Adding two elements $\begin{bmatrix} x_1 \\ mx_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ mx_2 \end{bmatrix} \in S$ we get,

   $$\begin{bmatrix} x_1 \\ mx_1 \end{bmatrix} + \begin{bmatrix} x_2 \\ mx_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ mx_1 + mx_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ my_1 \end{bmatrix}, \quad \text{where } y_1 = x_1 + x_2 \in \mathbb{R}$$

This means that $\begin{bmatrix} y_1 \\ my_1 \end{bmatrix}$ belongs to S, this the set S is closed under vector addition. This means that S is a subspace of $\mathbb{R}^2$.

Since, the subset S is closed under both vector addition and scalar multiplication, it is a subspace of $\mathbb{R}^2$.



(a) A subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.6: Example of a subspace of $\mathbb{R}^2$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

Example 1.7. A straight line not through the origin. Consider the set of all points in $\mathbb{R}^2$ of the following form:

$$S = \left\{ \mathbf{x} : x = \begin{bmatrix} x \\ mx + c \end{bmatrix} \in \mathbb{R}^2,\ m, c \in \mathbb{R} \right\}$$

This is shown in the Figure 1.7a.

How do we verfiy this is a subspace of $\mathbb{R}^2$? The definition above shows that any x in S comes from $\mathbb{R}^2$, which means its a subset of $\mathbb{R}^2$. Figure 1.7a shows the set S for $m = -\frac{1}{2}$ and $c = 1$. How do we verify if S is a subspace of $\mathbb{R}^2$? We need to now verify that S satisfies the properties of a vector space.

1. First, let's check if S contains the zero vector. If it does not contain the zero vector, then it cannot be a subspace. The elements from S are of the form $\begin{bmatrix} x \\ mx + c \end{bmatrix}$, thus if we choose $x = 0$, then we get $\begin{bmatrix} 0 \\ c \end{bmatrix} \in S$. So, S does not contain the zero vector, which implies that S is not a subspace space of $\mathbb{R}^2$. We need not check the other two conditions; but we will test them just to see which of these two fails.

2. Scaling the element $\begin{bmatrix} x \\ mx + c \end{bmatrix} \in S$ by a scalar d we get,

$$d \begin{bmatrix} x \\ mx + c \end{bmatrix} = \begin{bmatrix} dx \\ dmx + dc \end{bmatrix} = \begin{bmatrix} dx \\ m\,(dx) + dc \end{bmatrix} \neq \begin{bmatrix} y \\ my + c \end{bmatrix}, \quad \text{where } y = dx \in \mathbb{R}$$

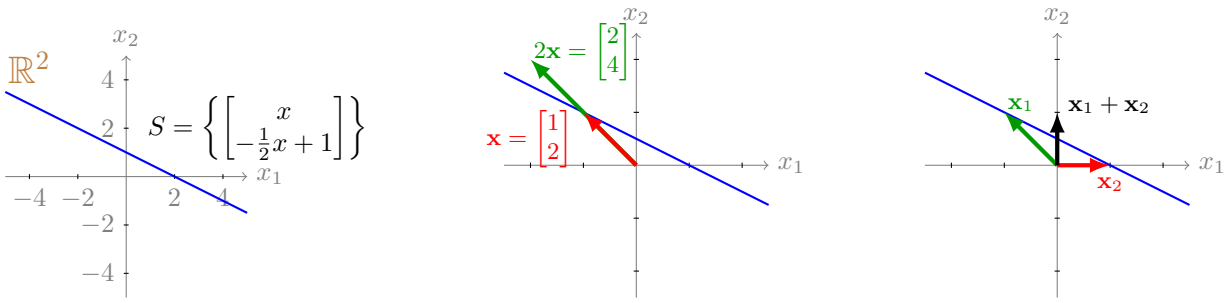
This means that $d\begin{bmatrix} x \\ mx+c \end{bmatrix} \notin S$. Thus, the set S is closed under scalar multiplication. Another confirmation that it is not a subspace. This can be seen in Figure 1.7b, which shows that when we choose an elment from $\mathbf{x}$ (red arrow) from S (blue line), the scaled version of this vector leaves the set S, i.e., the tip of the green arrow does not stay on the blue line.

3. Let's verify vector addition. Adding two elements $\begin{bmatrix} x_1 \\ mx_1+c \end{bmatrix}, \begin{bmatrix} x_2 \\ mx_2+c \end{bmatrix} \in S$ we get,

$$\begin{bmatrix} x_1 \\ mx_1+c \end{bmatrix} + \begin{bmatrix} x_2 \\ mx_2+c \end{bmatrix} = \begin{bmatrix} x_1+x_2 \\ mx_1+mx_2+2c \end{bmatrix} \neq \begin{bmatrix} y_1 \\ my_1+c \end{bmatrix}, \quad \text{where } y_1 = x_1+x_2 \in \mathbb{R}$$

This means that $\begin{bmatrix} x_1 \\ mx_1+c \end{bmatrix} + \begin{bmatrix} x_2 \\ mx_2+c \end{bmatrix} \notin S$. Thus the set S is closed under vector addition. We see this geometrically in Figure 1.7c, where the sum of two vectors in S does not stay in the set S. Even though the tips of the green and red arrow are on the blue line, the tip of the black arrow is not on the blue line.

Since, the subset S is closed under both vector addition and scalar multiplication, it is a subspace of $\mathbb{R}^2$.



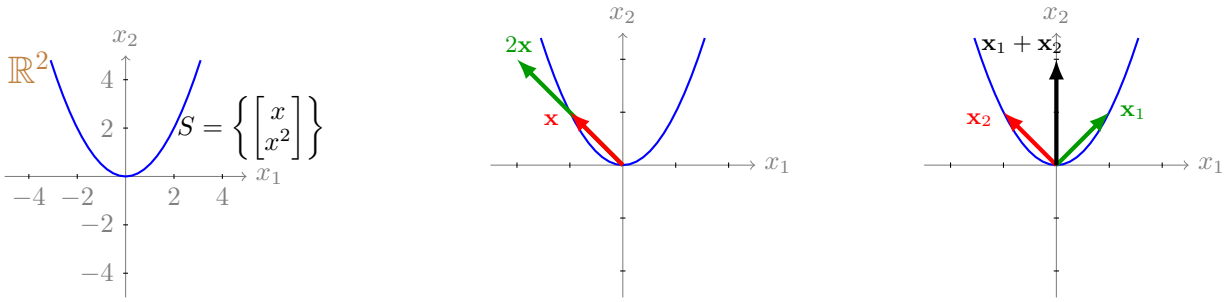
(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.7: Example of a subspace of $\mathbb{R}$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

Example 1.8. A parabola through the origin. Consider the set of all points in $\mathbb{R}^2$ of the following form:

$$S = \left\{ \mathbf{x} : x = \begin{bmatrix} x \\ \frac{1}{2}x^2 \end{bmatrix} \in \mathbb{R}^2, m, c \in \mathbb{R} \right\}$$

This is not a subspace of $\mathbb{R}^2$. This is geometrically depicted in Figure 1.8a, Figure 1.8b and Figure 1.8c. You are encouraged to verify this algebraically by checking the properties of a vector space.



(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.8: Example of a subspace of $\mathbb{R}$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

1.7 Linear combination

Linear combination is an *algebraic operation* performed on a set of vectors. We can combine the two fundamental operations on vectors into a single operation called the *linear combination* of a set of vectors. Given a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in \mathbb{R}^n$ and scalars $c_1, c_2, \ldots, c_n \in \mathbb{R}$, the linear combination of the vectors is given by:

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m \in \mathbb{R}^n \tag{1.1}$$

Notice that the linear cominations of single vector $\mathbf{v}_1$ are simply different scaled versions of the vector $c_1\mathbf{v}_1$. Linear cominations are the bread-and-butter of linear algebra and we will encounter them again and again. An informal way to think of a linear combination of a set of vectors as process of mixing the set of vectors together with the corresponding scalar c_i determining the amount of a vector in the mixture. There are other types of combinations of vectors, which we will not discuss further in this book.

- **Affine combination:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad \sum_{i=1}^{m} c_i = 1$
- **Convex combinations:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad c_i \geq 0, \ \sum_{i=1}^{m} c_i = 1$
- **Conic combinations:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad c_i \geq 0$

1.8 Linear independence of a set of vectors

- Linear independence is a property of a set of vectors.
- A set is either linear independent or its not.
- No element of a linearly independent set can be represented as linear combination of the other elements in the set.
- Linearly independent set does not have any redundancy.

Linear independece is a *property* of a set of vectors; a set of vector is either linearly independent or linearly dependent. The concept of linear independence is easy to understand the but the algebraic condition for independence can seem a bit unintuitive. A set of vectors is said to be linearly independent if no vector in the set can be expressed as a linear combination of the other vectors in the set. This means that there is

some unique information contained in each element of the set, which cannot be obtained from the other elements of the set, i.e., there is no redundancy, so to speak.

More formally, a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^{m}$ is said to be linearly independent if and only if the only way to produce the zero vector $\mathbf{0}$ through the linear combination of the set V is by setting all the scalars to zero, i.e.,

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m = \mathbf{0} \quad \text{if and only if} \quad c_1 = c_2 = \cdots = c_m = 0 \tag{1.2}$$

To understand this better, let's assume that the set V is linear dependent and let's assume that the vector $\mathbf{v}_m$ can be represented as the linear combination of the vectors $\mathbf{v}_1, \mathbf{v} - 2, \cdots \mathbf{v}_{m-1}$. This means that there exist a set of scalar $\alpha_i,\ 1 \leq i \leq m-1$, such that

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_{m-1}\mathbf{v}_{m-1} = \mathbf{v}_m$$

Multiplying both sides by a scalar $c_m \neq 0$ we get,

$$\begin{aligned} c_m\alpha_1\mathbf{v}_1 + c_m\alpha_2\mathbf{v}_2 + \cdots + c_m\alpha_{m-1}\mathbf{v}_{m-1} &= c_m\mathbf{v}_m \\ \implies c_m\alpha_1\mathbf{v}_1 + c_m\alpha_2\mathbf{v}_2 + \cdots + c_m\alpha_{m-1}\mathbf{v}_{m-1} - c_m\mathbf{v}_m &= \mathbf{0} \end{aligned}$$

This implies that there exist a set of scalar $c_i = c_m\alpha_i,\ 1 \leq i \leq m-1$, and c_m such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m = \mathbf{0}$, where not all c_i are zero. So when a set is linearly dependent, then there are scalars c_i, not all zero, such that the linear combination of the vectors from V with these scalars produces the zero vector.

Now, let's assume that the set V is linearly independent, that is no vector in the set V can be expressed as a linear combination of other vectors in that set. And let's assume that there are scalars c_i, not all zero, such that the linear combination of the vectors from V with these scalars produces the zero vector, i.e.,

$$\begin{aligned} c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m &= \mathbf{0} \\ \implies \frac{c_1}{c_m}\mathbf{v}_1 + \frac{c_2}{c_m}\mathbf{v}_2 + \cdots + \frac{c_{m-1}}{c_m}\mathbf{v}_{m-1} &= \mathbf{v}_m,\ c_m \neq 0 \end{aligned}$$

But this a contradiction because we have just expressed $\mathbf{v}_m$ is expressed as a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_{m-1}$.

Example 1.9. Consider the set of vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$, such that $\mathbf{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}$ in $\mathbb{R}^2$. This set is linearly independent. Let's verify this algebraically. Let's assume that there exist scalars c_1, c_2 such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \mathbf{0}$. This implies that, $c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}c_1\\c_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \implies c_1 = c_2 = 0$. Thus the set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly independent.

Example 1.10. Consider the set of vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$, such that $\mathbf{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}, , \mathbf{v}_3 = \begin{bmatrix}23\\-5\end{bmatrix}$ in $\mathbb{R}^2$. This set is not linearly independent, i.e., its linear dependent. Let's assume that there exist scalars c_1, c_2, c_3 such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$. This implies that, $c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\end{bmatrix} + c_3\begin{bmatrix}23\\-5\end{bmatrix} = \begin{bmatrix}c_1 + 23c_3\\c_2 - 5c_3\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \implies c_1 = -23c_3$ and $c_2 = 5c_3$. If we choose c_3 to be a non-zero value, we have a set of non-zero scalars such that linear combination of the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ produces the zero vector. Thus, this set is linearly dependent.

Example 1.11. Consider the set of vectors $\{\mathbf{v}_1\}$, $\mathbf{v}_1 = \begin{bmatrix}-1\\1\\1\end{bmatrix}$ in $\mathbb{R}^3$. This set is linearly independent.

Let's assume that there exist scalars c_1 such that $c_1\mathbf{v}_1 = \mathbf{0}$. This implies that, $\begin{bmatrix} -c_1 \\ c_1 \\ c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \implies c_1 = 0$. Thus, the set $\{\mathbf{v}_1\}$ is linearly independent.

Example 1.12. Consider the set $\{\mathbf{0}\}$ in $\mathbb{R}^3$. This set is linearly dependent. Here, $c_1\mathbf{0} = \begin{bmatrix} 0c_1 \\ 0c_1 \\ 0c_1 \end{bmatrix} = \mathbf{0}$. Any non-zero c_1 will produce the zero vector. Thus, the set $\{\mathbf{v}_1\}$ is linearly dependent. In fact, any set that contains the zero vector is linearly dependent. (Why? Can you show this algebraically?)

1.9 Span of a set of vectors

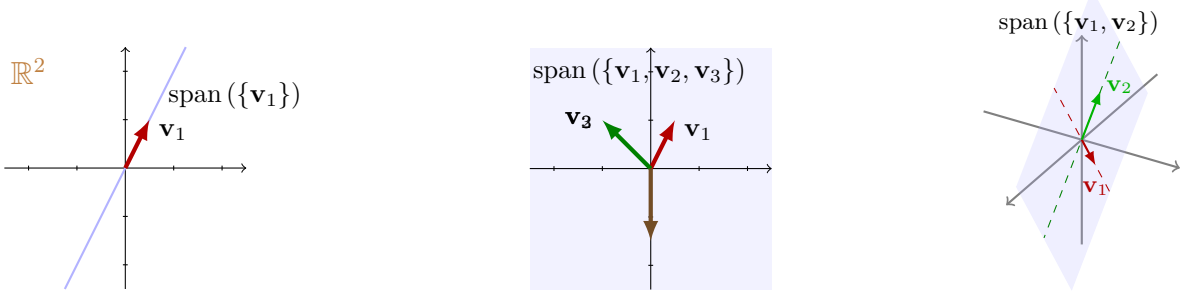
- The span of a set of vectors V is another set.
- It is generated through the linear combination of the elements of V.
- The span of a set of vector V is a subsapce of the originak vectors space the elements of V are from.

So, linear combinations of a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^m$ ($\mathbf{v}_i \in \mathbb{R}^n$) is a way of generating new vectors not in that set. All we need to do is choose a random set of real numbers $\{c_i\}_{i=1}^m$, and "mix" the vectors $\mathbf{v}_i$ from the set using these as weights. Clearly there are infinite number of vectors we could generate through this process, and we can put them all together in a set. And this set has a name – the *span* of the set V. The span of a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^m$ is denoted by span (V) and is defined as:

$$\text{span}\,(V) = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m \,:\, c_i \in \mathbb{R}\} \subseteq \mathbb{R}^n \tag{1.3}$$

Its clear that this will be a subset of $\mathbb{R}^n$, but it turns out it is also a subspace of $\mathbb{R}^n$. Why? Can you verify this fact algebraically? (*Hint*: Just follow the steps in Examples 1.6-1.8).

Geometrically, this means that the span (V) will be a flat surface in $\mathbb{R}^n$. Which means that the linear combination operation generates vectors that lie on a flat surface spanned by the vectors emmployed in the linear combination.



(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.9: Span of a set of vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$.

1.10 How big is a vector?

The size of a vector is an extension of the idea of the magnitude of a real number. The magnitude of a real number $a \in \mathbb{R}$ tells us how big the number is irrespective of its sign:

$$|a| = \begin{cases} a, & a \geq 0 \\ -a, & a < 0 \end{cases} \tag{1.4}$$

The "magnitude" or size of an element of a vector space (such as $\mathbb{R}^n$) is called the *norm* of the vector. The norm is a generalization of the magnitude of a real number to a vector. The norm of a vector is a function that maps a vector to a non-negative real number, and satisfies the following properties:

- **Non-negativity:** For any vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\| \geq 0$.
- **Definiteness:** The norm of a vector is zero if and only if the vector is the zero vector, i.e., $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- **Homogeneity:** Scaling a vector by a scalar c, scales the norm of the vector by $|c|$. For any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar $c \in \mathbb{R}$, $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$.
- **Triangle inequality:** For any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

According to this definition, the magnitude of real numbers (Eq. 1.4) is a norm of the vector $\mathbb{R}$. The most common norm of a vector is the *Euclidean norm* or the *2-norm* of a vector. The Euclidean norm of a vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is defined as:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{1.5}$$

We are well-versed with this as the length of a vector in $\mathbb{R}^2$ and $\mathbb{R}^3$. The properties of non-negativity, definiteness, and homogeneity are easy to verify. The triangle inequality is a bit more involved.

The subscript 2 in Eq. 1.5 is used to indicate that it is the 2-norm, which is a special case of a general class of norms in $\mathbb{R}^n$ – the *p-norm*. The *p-norm* is defined as the following:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \; p \in \mathbb{Z}, \; p \geq 1 \tag{1.6}$$

Apart from the *2-norm*, the *1-norm* and the ∞*-norm* are also commonly used norms, which are defined as the following:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|\mathbf{x}\|_\infty = \max_i |x_i| \tag{1.7}$$

The *1-norm* is the sum of the absolute value of the elements of the vector, and the ∞*-norm* is the maximum of the absolute value of the elements of the vector. The *1-norm* is also called the *Manhattan norm* or the *Taxicab norm* because it measures the distance between two points in a city if you can only travel along the grid of streets.

Example 1.13. Let's calculate the 1-norm, 2-norm, and ∞-norm of the some vectors:

1. $\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \longrightarrow \|\mathbf{x}_1\|_1 = 5, \quad \|\mathbf{x}_1\|_2 = \sqrt{1+1+9} = \sqrt{11}, \quad \|\mathbf{x}_1\|_\infty = 3.$

2. $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \longrightarrow \|\mathbf{x}_1\|_1 = 1, \quad \|\mathbf{x}_1\|_2 = 1, \quad \|\mathbf{x}_1\|_\infty = 1.$ All the p-norms of the unit vectors are 1. No wonder we call them "unit" vectors.

3. $\|\mathbf{0}\|_1 = \|\mathbf{0}\|_2 = \|\mathbf{0}\|_\infty = 0.$ All p-norms will produce 0, otherwise it is not a norm (remember the definiteness property?)

Problem 1.1. Why does the ∞-norm measure have this weird looking definition compared to the other p-norms?

Solution. Consider the vector $\mathbf{x} \in \mathbb{R}^n$, and $x_{max} = \max_{0 \le i \le n} |x_i|$; let's also assume that the j^{th} element of $\mathbf{x}$ has the maximum absolute value, i.e. $x_{max} = |x_j|$. The p-norm is defined as the following:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = x_{max} \left(1 + \sum_{\substack{1 \le i \le n \\ i \ne j}} \left| \frac{x_i}{x_j} \right|^p \right)^{1/p} = x_{max} \, (N)^{1/p}$$

where, N is a real number between 1 and n, because $|\frac{x_i}{x_j}| \le 1$ (why?). Now, if we increase the value of p to infinity, then the term $\lim_{p \to infty} (N)^{1/p} = 1$. Thus, we have $\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p = x_{max} = \max_i |x_i|$. □

1.10.1 Geometry of the p-norms

In the case of real numbers, the set of all numbers with a magnitude of 1 is the set $\{-1, 1\}$. We can plot these points in the real line as below.

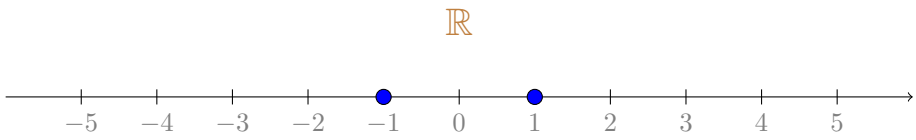


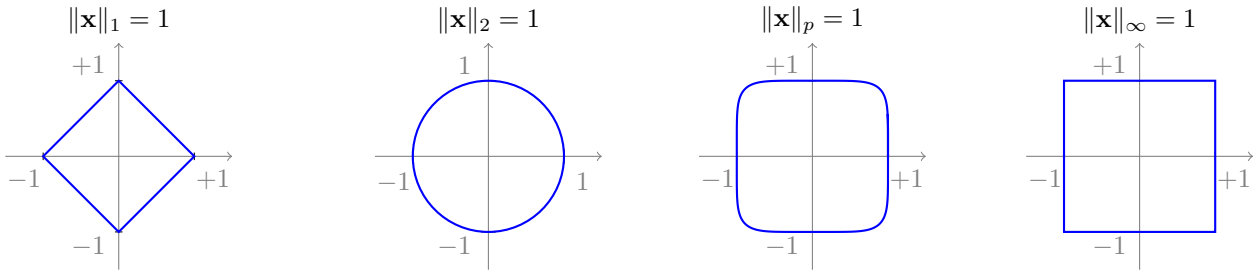
Figure 1.10: The set of all real numbers with magnitude 1. This set contains two numbers $\{-1, 1\}$.

In $\mathbb{R}^2$, the set of all vectors from $\mathbb{R}^2$ with a *2-norm* of 1 is the unit circle. The following figure shows the set of all points in $\mathbb{R}^2$ with unit 1, 2, p, and ∞ norm.

Problem 1.2. Can you explain why the different norms have these shapes?

Problem 1.3. Can you write a Python program to generate the above plots for different values of $p = 1, 2, 3, 10$ and ∞?

Problem 1.4. Can describe what these 1, 2, p and ∞ norms will look like in $\mathbb{R}^3$?



(a) 1-norm **(b)** 2-norm **(c)** p-norm **(d)** ∞-norm

Figure 1.11: Locus of all points with unit 1, 2, p, and ∞ norms in $\mathbb{R}^2$.

1.11 How similar are two vectors?

The idea of how similar two or more vectors are is an important topic in data analysis, in particular in classification problems in machine learning. Vectors that are "similar" somehow belong to the same "category" or "class", while vectors that are "dissimilar" belong to different categories or classes. There are various ways to measure the similarity between two vectors. We will look at two methods in this section where similarity is mesasured by computing the distance between two vectors or by computing the angle between two vectors.

1.11.1 Distance between two vectors

The logic here is that similar vectors correspond to points that are close together, while disimilar vectors are father away. We can make use of the norm to compute the distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Since the the difference between these two vectors $\mathbf{x} - \mathbf{y}$ is also another vector, we can compute the distance between vectors $\mathbf{x}$ and $\mathbf{y}$ as the norm of the vector $\mathbf{x} - \mathbf{y}$ (Figure 1.12a).

$$\text{Distance between } \mathbf{x} \text{ and } \mathbf{y} = d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$$

We could use any of the p-norms to compute this or come-up with a new norm depending on the application we are dealing with. Take look at the clusters of points shown in Figure 1.12b, we would agree that the different colored points each form a cluster, since the points of the same color are closer to each other than points from another color.

Example 1.14. Test scores in ALADA. The ALADA course has three segments: linear algebra, optimization, and probability/statistics. Let's assume that the final exam contains three sections with a maximum of 25 marks students. The scores from these three segments are stored in a 3-vector of the form $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^3$, where x_1, x_2, x_3 are the marks obtained for linear algebra, optimization, and probability/statistics section, respectively. Consider the scores from the 5 students that took the course:

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 5 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 8 \\ 20 \\ 22 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} 9 \\ 20 \\ 21 \end{bmatrix}, \quad \mathbf{x}_5 = \begin{bmatrix} 24 \\ 24 \\ 23 \end{bmatrix}, \quad \mathbf{x}_6 = \begin{bmatrix} 24 \\ 23 \\ 22 \end{bmatrix}$$

The distance between the scores of these students tells us something about the ability of the students in the course. Let $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ be the Euclidean distance between the scores of student i and j;

notice that $d_{ij} = d_{ji}$. The distance between the different scores is given by the following table.

		$\mathbf{x}_i$					
		1	2	3	4	5	6
$\mathbf{x}_j$	1	0.0	3.7	26.5	26.0	36.5	35.3
	2	θ_{12}	0.0	25.3	24.8	35.3	34.1
	3	θ_{13}	θ_{23}	0.0	1.4	16.5	16.3
	4	θ_{14}	θ_{24}	θ_{34}	0.0	15.7	15.3
	5	θ_{15}	θ_{25}	θ_{35}	θ_{45}	0.0	1.4
	6	θ_{16}	θ_{26}	θ_{36}	θ_{46}	θ_{56}	0.0

The following observations can be made from the table:

- Students 1 and 2 have similar scores, compared to the other students.
- Students 3 and 4 have very similar scores, compared to the other students.
- Students 5 and 6 have very similar scores, compared to the other students.
- Students 1 and 2, are closer to 3 and 4 than to 5 and 6.

Notice that you could have also used the other norms to creata table similar to the above one. Its left as an exercise for you generate a similar table using the 1-norm and the ∞-norm to define the distance between the scores of the students.

Using norms instead of norms of differences. Another way to understand the score vectors is to directly compute the norms of $\mathbf{x}_i$ and see what information they convey about the students' performance in the ALADA final exam.

$\|\mathbf{x}_1\|_2$	$\|\mathbf{x}_2\|_2$	$\|\mathbf{x}_3\|_2$	$\|\mathbf{x}_4\|_2$	$\|\mathbf{x}_5\|_2$	$\|\mathbf{x}_6\|_2$
5.5	5.8	30.8	30.4	41.0	39.9

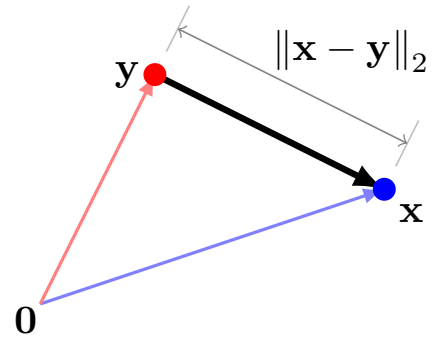
The size of the score vectors tells us that students 1 and 2 performed worst among the six students, while students 5 and 6 performed the best; the performacne of students 3 and 4 was somehwere in the middle. Would we have reached similar conclusions if we had used the 1-norm or the ∞-norm to compute the norms of the score vectors?

1.11.2 Angle between two vectors

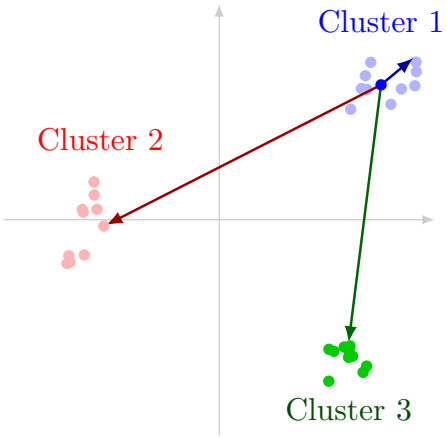
This approach is based on the idea that the direction of the vector repesenting a point contains information about the point. Thus, vectors that point in a similar direction could be considered similar. But how do we measure the angle between two vectors in $\mathbb{R}^n$? This is where the concept of the *standard inner product* (or the dot product from vectors from high school math and physics). The standard inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is defined as:

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

The superscript '$\top$' represents the transpose operation. We will not worry about what it means in the next chapter. The standard inner product takes in two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and returns a scalar value $\mathbb{R}$; it can be both positive and negative. We compute it by simply taking the two vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$,



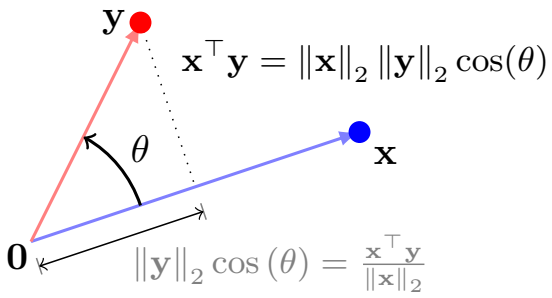
(a) Distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. This figures depicts the 2-norm, but any p-norm or valid norm function could be used to quantify the distance between two vectors or points.



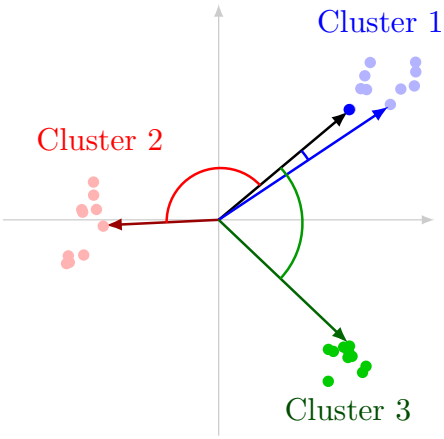
(b) Distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. This figures depicts the 2-norm, but any p-norm or valid norm function could be used to quantify the distance between two vectors or points. could be used to quantify the distance between

Figure 1.12

and multiply the two of them element-wise $x_i y_i$, $1 \leq i \leq n$ and add the n products together $\sum_{i=1}^{n} x_i y_i$ to obtain the inner product.



(a) Angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. The standard inner product provides a measure of the cosine of the angle between the two vectoes.



(b) The relative angle between the points of the same colors is smaller than that of points of different colors.

Figure 1.13

The standard inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is related to the cosine of the angle θ between the two vectors, and the 2-norm of the two vectors.

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \cdot \cos(\theta)$$

The angle θ between the two vectors $\mathbf{x}$ and $\mathbf{y}$ can be computed as:

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}\right), \quad \|\mathbf{x}\|_2 \neq 0, \|\mathbf{y}\|_2 \neq 0$$

If the 2-norms of the $\mathbf{x}$ and $\mathbf{y}$, then $\mathbf{x}^\top \mathbf{y}$ is simply the cosine of the angle between the vectors.

Example 1.15. Let's look at the data from Example 1.14, but this time using the angle between the vectors to understand the performance of the students in the ALADA final exam. The angle between the scores (in degrees) of the students is given by the following table. Let $\theta_{ij} = \cos^{-1}\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}\right)$ be the angle between the scores of student i and j; notice that $\theta_{ij} = \theta_{ji}$.

		$\mathbf{x}_i$					
		1	2	3	4	5	6
$\mathbf{x}_j$	1	0.0	38.5	35.1	33.3	31.7	32.1
	2	θ_{12}	0.0	16.7	15.0	9.2	9.9
	3	θ_{13}	θ_{23}	0.0	2.5	21.1	22.2
	4	θ_{14}	θ_{24}	θ_{34}	0.0	18.7	19.9
	5	θ_{15}	θ_{25}	θ_{35}	θ_{45}	0.0	1.2
	6	θ_{16}	θ_{26}	θ_{36}	θ_{46}	θ_{56}	0.0

It looks like the angles do not do a good job of capturing the similarity between the scores of the students like the distance between the scores, in particular θ_{12} is quite large, while θ_{34} and θ_{56} are quite small. Why do you think this is so?

1.12 Standard and other inner products

$\mathbf{x}^T\mathbf{y}$ is the standard inner product, which of course means there are non-standard inner products. But before we look at generalizing the concept of an inner product, let's look at some properties of the standard inner product.

- **Connection to the 2-norm**. The standard inner product of a vector $\mathbf{x}$ with itself is the square of the 2-norm of the vector, i.e., $\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2$.
- **Cauchy-Bunyakovski-Schwartz Inequality**:

$$\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \tag{1.8}$$

The concept of an inner product is a general one. An inner product $\langle \cdot, \cdot \rangle$ is a function that maps two vectors from $\mathbb{R}^n$ to a scalar value, and satisfies the following properties:

- **Positive definiteness:** $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- **Symmetry:** For any vectors $\mathbf{x}, \mathbf{y}$, $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.
- **Linearity:** For any vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$, and any scalars $\alpha, \beta \in \mathbb{R}$, $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle + \beta\langle \mathbf{y}, \mathbf{z} \rangle$.

We will come across other inner products in due course, but we will stick to the standard inner product for most problem in $\mathbb{R}^n$ in this course. A class of inner products in $\mathbb{R}^n$ is of the form $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is a $n \times n$ *positive definite* matrix. The standard inner product is a special case of this class of inner products where $\mathbf{A} = \mathbf{I}$, the identity matrix. That may sound like too much jargon for now, but we will come to these concepts in the upcoming chapters.

Problem 1.5. Consider the vector space $\mathbb{R}^n$. Is the the following a valid inner product of $\mathbb{R}^n$?

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^n w_1 x_i y_i, \;\; w_i \in \mathbb{R}, \;\; w_i > 0$$

Solution. To find out if this is a valid inner product, we need to verify by if it satisfies the properties of an inner product.

- **Positive definiteness**: The first property is positive definiteness, which is satisfied because $w_i > 0$.

$$\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{n} w_i x_i^2 \geq 0, \text{ since, } w_i > 0 \text{ and } x_i^2 \geq 0$$

  Notice, that if any of the w_i is zero or negative, then $\langle \mathbf{x}, \mathbf{y} \rangle$ will not be a valid inner product (why?).

- **Symmetry**: From the commutativity and associativity of multiplication of real numbers, we have $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.
- **Linearity**: This is also satisfied. You should verify this yourself.

The given function is a valid inner product of $\mathbb{R}^n$. □

Problem 1.6. Consider the vector space $\mathbb{R}^n$. Is the the following a valid inner product of $\mathbb{R}^n$?

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} w_i x_i y_i, \; w_i = \begin{cases} +1, & i \text{ is odd} \\ -1, & i \text{ is even} \end{cases}$$

Solution. To find out if this is a valid inner product, we need to verify by if it satisfies the properties of an inner product.

- **Positive definiteness**: This is not satisfied. Can you provide an example where positive definiteness fails?

The given function is a *not* valid inner product of $\mathbb{R}^n$. □

1.13 Orthogonality of vectors

The concept of orthogonality is a generalization of the concept of perpendicularity in $\mathbb{R}^2$ and $\mathbb{R}^3$. Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are said to be orthogonal if their standard inner product is zero, i.e., $\mathbf{x}^\top \mathbf{y} = 0$.

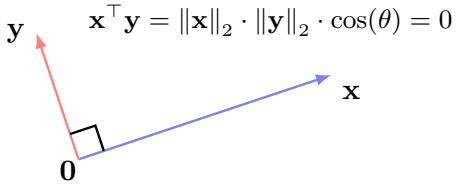


Figure 1.14: Orthogonal vectors in $\mathbb{R}^2$.

Geometrically, when two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then when we move along the direction of one vector, we are not moving along the direction of the other vector. If directions of vector convey some information about something, then vectors that are orthogonal to each other are vectors that convey mutually exclusive information,i.e. the two vectors share nothing in commong. The concept of orthogonality is a very important concept in linear algebra and we will encounter it again and again.

Note that this definition of orthogonality also implied that $\mathbf{0}$ is othorognal to all vectors!

Problem 1.7. Explain why the following statement about the unit vectors of $\mathbb{R}^n$ is true.

$$\mathbf{e}_i^\top \mathbf{e}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

All the unit vectors of $\mathbb{R}^n$ are orthogonal to each other, and have unit length.

Problem 1.8. Show if two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other, then the following is true.

$$\|\mathbf{x}-\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2$$

1.14 Basis of a vector space

A set of vectors $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$, $\mathbf{v}_i \in \mathbb{R}^n$ is called a basis for $\mathbb{R}^n$ if it is a linearly independent set and if it spans $\mathbb{R}^n$. This effectively means the following two things. If V is a basis for a vector space, then,

1. " spans $\mathbb{R}^n$ " $\longrightarrow$ It can be used to generate every element of $\mathbb{R}^n$ through a linear comibination operation.

2. " linearly independent " $\longrightarrow$ There is a unique linear combination of the elements of V that produces every element of $\mathbb{R}^n$.

A basis is the smallest possible set for generating a vector space.

Note that although the above definition and description is done usign $\mathbb{R}^n$, the concept of a basis applies to any vector space and all its subspaces.

Example 1.16. Consider the set of vectors $V = \{\mathbf{e}_1\} \subset \mathbb{R}^2$. This set V forms a basis for the subspace $S_1 = \left\{ \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \ : \ \alpha \in \mathbb{R} \right\}$. This is because, the span $(V) = S_1$ (verify this) and the set V is linearly independent, because $\beta\mathbf{e}_1 = \mathbf{0}$ only if $\beta = 0$.

Example 1.17. Consider the set of vectors $V = \{\mathbf{e}_1, 3\mathbf{e}_1\} \subset \mathbb{R}^2$. This set V does not form a basis for the subspace $S_1 = \left\{ \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \ : \ \alpha \in \mathbb{R} \right\}$. The span $(V) = S_1$ (verify this). But the set V is linearly dependent, because $-3\mathbf{e}_1 + 1\,(3\mathbf{e}_1) = \mathbf{0}$, thus a non-zero set of coefficients result in the zero vector.

Example 1.18. Consider the set of vectors $V = \{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^2$. This set V is a basis for $\mathbb{R}^2$. The span $(V) = \mathbb{R}^2$ (verify this). And this set V is linearly independent, because $\beta_1\mathbf{e}_1 + \beta_2\mathbf{e}_2 = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, thus the only was we get the zero vector through the linear combination is if $\beta_1 = \beta_2 = 0$.

Example 1.19. Consider the set of vectors $V = \left\{ \mathbf{e}_1, \mathbf{e}_2, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\} \subset \mathbb{R}^2$. This set V is a basis for $\mathbb{R}^2$. The span $(V) = \mathbb{R}^2$ (verify this). And this set V is linearly dependent, because $\mathbf{e}_1 + \mathbf{e}_2 - 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbf{0}$; a non-zero linear comination produces the zero vector. Thus, V is not a basis for $\mathbb{R}^2$.

How many different basis does a vector space have? For instance, how many different basis does $\mathbb{R}^2$ have? For a set to be a basis, all we need to ensure is that the set spans $\mathbb{R}^2$ and is linearly independent. Thus, there are infinitely many basis for $\mathbb{R}^2$. For instance, the sets $\{\alpha_1\mathbf{e}_1, \alpha_2\mathbf{e}_2\}$ with $0 \neq \alpha_1, \alpha_2 \in \mathbb{R}$ are all basis for $\mathbb{R}^2$. Since there are inifinite number of choices for α_1, α_2, we have an infite number of basis for $\mathbb{R}^2$. The same argument applies to $\mathbb{R}^n$.

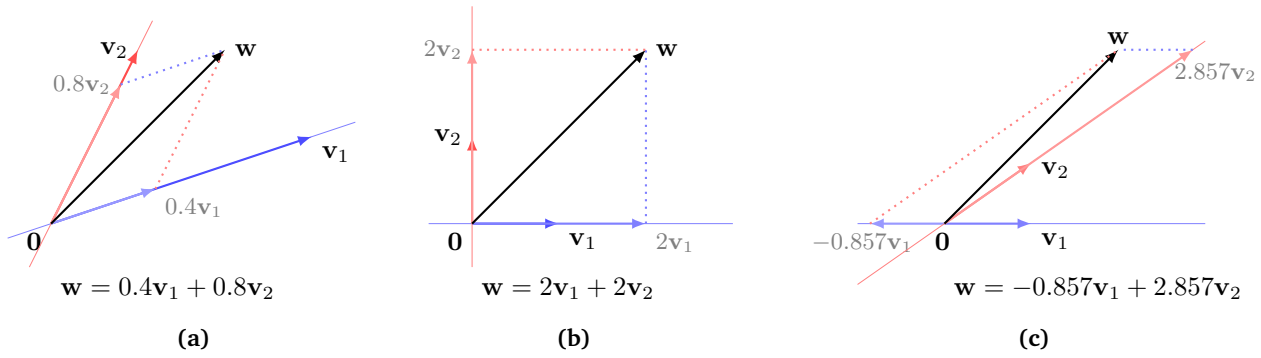


Figure 1.15: Representation of $\mathbf{w}$ in three different basis of $\mathbb{R}^2$. (a) and (c) are some arbitrary basis, while (b) is an orthonormal basis.

1.14.1 Orthonormal basis

Among the infinite number of basis for a vector space, there is a special class of basis called the *orthonormal basis*. Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_n\}$, then for an orthonormal basis, the following properties hold:

$$\mathbf{v}_i^\top \mathbf{v}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{1.9}$$

This means that all basis vector have unit 2-norm, and are orthogonal to each other. We will come across orthonormal basis often in this course, and for a good reason. Orthonormal basis and easy to work with and its easy to compute the representation of a vector in an orthonormal basis. Let V be an orthonormal basis for $\mathbb{R}^n$, and let $\mathbf{w} \in \setminus^n$ with the following representation,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$$

Then, the coefficients α_i can be computed as the following:

$$\mathbf{v}_j^\top \mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_j^\top \mathbf{v}_i = \alpha_j, \quad \text{because } \mathbf{v}_i^\top \mathbf{v}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

A special orthonormal basis for the $\mathbb{R}^n$ is the *standard basis* $\{\mathbf{e}_1, \mathbf{e}_2 \cdots \mathbf{e}_n\}$.

1.15 Dimension of a vector space

Although there can be infinite number of basis for a vector space – they all have the same number of elements or basis vectors. This number is called the *dimension* of the vector space that they span. This is also sometimes called the *degrees of freedom* of the vector space. We represent the dimension of a vector space V as $\dim(V)$, e.g. $\dim(\mathbb{R}^n) = n$.

The dimension of a vector space also tells us that the maximum number of linearly independent vectors we can choose from that vector space. For instance, in $\mathbb{R}^n$ we can only choose n vectors that can form a linearly independent set. If we already have a linearly independent set with n elements, then adding even one more vector (any vector) to the set will make it linearly dependent. Proper subspaces of a vector space will have dimensions less than the vector space itself.

Example 1.20. The dimension of $\mathbb{R}^n$ is n. The following are subspaces of $\mathbb{R}^n$ and their dimensions. Consider the set $V = \{\mathbf{v}_1\}$ with $\mathbf{v}_1 \neq \mathbf{0}$.

- $\dim(\text{span}(V)) = 1$
- Now, let's add another vector $\mathbf{v}_2$ to V to get $V_1 = \{\mathbf{v}_1, \mathbf{v}_2\}$ which is still linearly independent, then
$$\dim(\text{span}(V_1)) = 2$$
- Let's now add the vector $\mathbf{v}_1 - \mathbf{v}_2$ to V_1 to get $V_2 = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2\}$, then
$$\dim(\text{span}(V_2)) = 2$$
Why?
- Consider the $V_k = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$ which is linearly independent.
$$\dim(\text{span}(V_k)) = ?$$
Can you find out the answer and explain?

1.16 Linear functions

We will conclude this chapter with a brief introduction to linear functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be linear if it satisfies the following property $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\forall \alpha, \beta \in \mathbb{R}$:

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}) \tag{1.10}$$

This means that $f(\mathbf{0}) = 0$ for all linear functions. if a function does not satisfy this property, then it is not linear.

The standard inner product $\mathbf{w}^\top\mathbf{x}$ with a fixed vector $\mathbf{w}$ is a linear function of $\mathbf{x}$,

$$f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

An interesting fact about linear functions is that every possible $f : \mathbb{R}^n \to \mathbb{R}$ can be presented as a inner product operation with a fixed vector $\mathbf{w} \in \mathbb{R}^n$. This means the following: if f is a linear function $\mathbb{R}^n$ to $\mathbb{R}$, then there exists a vector $\mathbf{w} \in \mathbb{R}^n$ such that $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}, \forall \mathbf{x} \in \mathbb{R}^n$. This might seem like a strange fact at first. Let's now look at how we could find the vector $\mathbf{w}$ if the function f is linear. This is very simple. We first compute the value of the function for the n unit vectors of $\mathbb{R}^n$, i.e. $f(\mathbf{e}_1), f(\mathbf{e}_2), \cdots f(\mathbf{e}_n)$. Then, the vector $\mathbf{w}$ is simply the vector of these values, i.e. $\mathbf{w} = \begin{bmatrix} f(\mathbf{e}_1) \\ f(\mathbf{e}_2) \\ \vdots \\ f(\mathbf{e}_n) \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$. For any given vector $\mathbf{x}$, let the representation of $\mathbf{x}$ in the standard basis be $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$; this simply means that $\mathbf{x} = \sum_{i=1}^n x_i\mathbf{e}_i$.

$$f(\mathbf{x}) = f\left(\sum_{i=1}^n x_i\mathbf{e}_i\right) = \sum_{i=1}^n x_i f(\mathbf{e}_i) = \sum_{i=1}^n x_i w_i = \mathbf{w}^\top\mathbf{x}$$

Here is another interesting consequence of linearity. If we know the value of a linear function for a set of vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_k\}$, then know the output of the function for the set span (X). Can you think of why it is so?

1.17 Applications

The concepts convered in this chapter are enough to understand the following important and useful applications:

- k-nearest neighbors (k-NN) classification/regression algorithm
- k-means clustering algorithm

We will now look two applications in data analysis based on the concepts we have discussed in this chapter. We will present the bare-bones version of two commonly used algorithms in data analysis – the k-nearest neighbors (k-NN) classification algorithm and the k-means clustering algorithm. Numerous variations and improvements of these algorithms exist, which are beyond the scope of this course.

1.17.1 k-nearest neighbors (k-NN) classification and regression algorithms

The k-NN algorithm is a simple and intuitive algorithm for classification and regression problems. But before that, what are classification and regression problems? Both of them are problems were we one is interested in finding a function or a map from a set of features (or inputs) to a target value. The features are often the form of n-vectors, and the target value is a scalar value. Classificaiton and regression problems differ in the nature of the target value. In a classification problem, the target value is a class label (from set of finite values), while in a regression problem, the target value is a real-valued number (from an interval on the real line).

Examples of classification problems

Example 1.21. Disease diagnosis. We are often interested in knowing whether or not a patient presenting with a set of symptoms at the hospital has a particular disease, based on clinical symptoms, and clinical lab tests. This is a typical example of a classification problem encountered in medicine. Given a set of features of a patient, the goal of the classifier is to determine whether the patient has a particular disease or not.

- **Inputs**: Set of demographic data, clinical tests, imaging data, etc.
- **Output**: Disease label (e.g., positive, negative, 0: no disease/1: disease, etc.)

Example 1.22. Treatment prognosis: Let's assume there is a treatment that can be used for curing a particular disease. This treatment works well on a group of patients who recovery full after the treatment, while some only recover partially, and the rest do not recover at all. Given a patient with this disease, the goal of the treatment prognosis classifer is to predict the effectiveness of the treatment for patient. This classifier would take the features of the patient as input and predict the effectiveness of the treatment as one of three possible labels – *full recovery*, *partial recover* or *no recovery*.

- **Inputs**: Set of demographic data, clinical tests, imaging data, severity of the disease, etc.

- **Output**: Recovery label (e.g., full recovery, partial recovery, or no recovery)

Example 1.23. Spam email detection: This is a classifier that we see in action on a daily basis. Our email managers/service provides automatically send certain emails to the spam folder to weed out the unless emails from the useful ones. Given an email, the goal of such a classifier is to determine whether the email is spam or not.

- **Inputs**: Features extracted from the email content, sender, subject, etc.
- **Output**: Spam label (e.g., spam, not spam)

Example 1.24. Handwritten digit recognition: Given an image of a handwritten digit, the classifier here needs to determine the corresponding digit of the image.

- **Inputs**: Image features extracted from the image of a handwritten digit.
- **Output**: Digit label (e.g., 0, 1, 2, ..., 9)

Examples of regression problems

Example 1.25. Growth prediction: We are interested in knowing the effect iof the addition of protien supplements to the daily diet of childen over a period of three months. The goal here is to develop a model that can predict the increase in height of the children after six months given the certain level of protien supplement in their diet.

- **Inputs**: Demographic details and amount of protien supplement in the diet.
- **Output**: Increase in child's height after six months.

Example 1.26. Clinical score estimation: There are clinical tests that are considered gold standard for know the health status of a patient. These gold standard tests are often time consuming, expensive, and difficult to administer on a regular basis. Thus, we are often interestd in estimating the outcome of this gold standard test using other clinically relevant variables that are easily measured.

- **Inputs**: A set of clinically relevant parameters that are easy to measure.
- **Output**: Estimate of the value of the gold standard test.

k-NN classification algorithm

The k-NN classifier is a very simple and intuitive algorithm. Let's assume that we have a dataset with N data points/samples $(\mathbf{x}_i, y_i)\,,\ 1 \leq i \leq N$, where $\mathbf{x}_i$ is the vector of features and y_i is the know label for the i^{th} sample. The feature vector is an elements from $\mathbb{R}^n$, i.e. $\mathbf{x}_i \in \mathbb{R}^n$. The labels taken on values from a finite set with L distinct labels, $y_i \in \{1, 2, \ldots, L\}$. We would like to use this available dataset, which will often be called the *training dataset*, to learn to correctly classify new samples that we may encounter in the future, where we have the feature vector $\mathbf{x}_{new}$ but we do not know the corresponding label y_{new}; we want a prediction of the label from the k-NN algorithm.

Although, the k-NN algorithm is called a *supervised learning* algorithm, there is really no learning in this algorithm. The algorithm keeps the entire training dataset in its memory, and assign the label to a new feature vector to be the label of the most similar feature vectors from the training dataset. The similarity can be measured through a multitude of ways, but the most common way is to use the Euclidean distance between the feature vectors. The only (hyper)parameter required for the k-NN algorithm is the value of

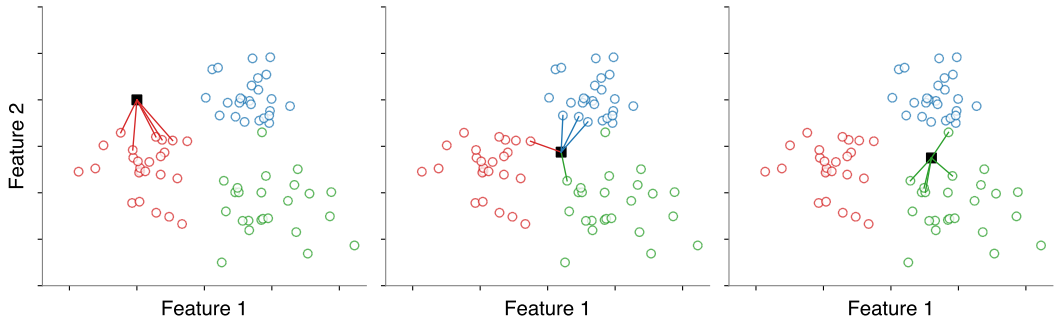


Figure 1.16: Demonstration of the k-NN classification algorithm. There are three classes or labels, which are shown in different colors. Three test points (black filled sqaure) are considered in the three plots shown in this figure. For each point the $k = 5$ nearest neighbours are depicted through lines joining the test point with the nearest neighbours. The colors of the line also indicate the class of that neighbour.

k, which is the number of nearest neighbors to consider. The label of the new feature vector is assigned by taking a vote from the k-nearest neighbors. The label with the highest number of votes is assigned to the new feature vector. The outline of the k-NN algorithm is given below in Algorithm 1.1.

Algorithm 1.1: k-Nearest Neighbors (k-NN) Algorithm

Data: Training data $(\mathbf{x}_i, y_i)\ \ 1 \leq i \leq N$; test point $\mathbf{x}_{new}$; number of neighbors k.
Result: Predicted label for test point $\mathbf{x}_{new}$.

foreach $\mathbf{x}_j$ *in* $(\mathbf{x}_i)_{i=1}^{N}$ **do**
  Compute the distance $d(\mathbf{x}_j, \mathbf{x}_{new})$ between $\mathbf{x}_j$ and $\mathbf{x}_{new}$;
end
Sort the training instances/samples $(\mathbf{x}_i, y_i)$ by distance in ascending order;
Select the k closest samples to $\mathbf{x}_{new}$; let $\mathcal{K}$ be the set of indices between 1 to N corresponding to these k closest samples;
let $Y_{\mathcal{K}}$ be the set of labels of these k closest samples;
foreach *label in* $Y_{\mathcal{K}}$ **do**
  Compute the frequency of each label;
end
Return the label with the highest frequency;

This is depicted in Figure 1.16, which shows the example of a 3-class classification problem. The three classes are shown in different colors. The $k = 5$ nearest neighbors of three test points are shown. The label of the test point is assigned based on the majority vote of the k nearest neighbors; whenever there is a tie, a tie-breaking rule is applied to choose the label of the test point. Although, simple and intuitive, the k-NN is a computationally expensive algorithm, especially when the number of samples in the training dataset is large. The advantage, however, is that there is no training process required and there is only only hyperparameter to choose.

k-NN regression algorithm

The k-NN regression algorithm uses the same principle as the k-NN classification algorothm. The only difference is that the output of the algorithm is the average of the y_i of the k nearest neighbors. This algorithm is detailed in Algorithm 1.2.

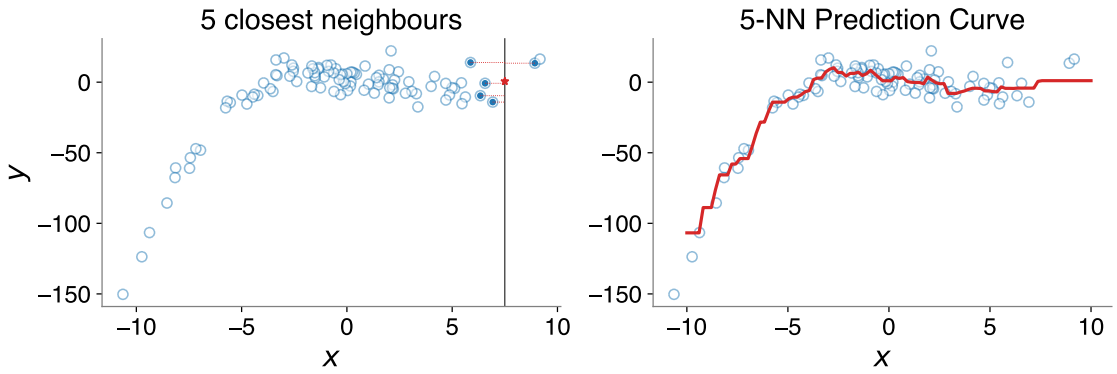


Figure 1.17: Demonstration of the k-NN regression algorithm. In this example, $\mathbf{x} \in \mathbb{R}$. The left plot demonstrates the algorithm, where the vertical black line is the $\mathbf{x}_{new}$, the filled blue circles are the 5 closest neighbours. The red star along the black line is the predicted value for $\mathbf{x}_{new}$. The right plot shows the 5-NN prediction curve for the given data in red.

Algorithm 1.2: k-Nearest Neighbors (k-NN) Regression Algorithm

Data: Training data $\{(\mathbf{x}i, y_i)\}\, i = 1^N$; test point $\mathbf{x}new$; number of neighbors k.
Result: Predicted value for test point $\mathbf{x}new$.

1 **foreach** $\mathbf{x}_j$ *in* $\{\mathbf{x}_i\}_{i=1}^N$ **do**
2 | Compute the distance $d(\mathbf{x}_j, \mathbf{x}_{new})$ between $\mathbf{x}_j$ and $\mathbf{x}_{new}$;
3 **end**
4 Sort the training instances/samples $(\mathbf{x}_i, y_i)$ by distance in ascending order;
5 Select the k closest samples to $\mathbf{x}_{new}$; let $\mathcal{K}$ be the set of indices between 1 to N corresponding to these k closest samples;
6 let $Y_{\mathcal{K}}$ be the set of values of these k closest samples;
7 Compute the average of the values in $Y_{\mathcal{K}}$;
8 Return the computed average as the predicted value;

1.17.2 k-mean clustering algorithm

The k-mean is a popular clustering algorithm, which unlike the k-NN algorithms, is an *unsupervised learning* algorithm. In a *supervised learning* algorithm, we have a dataset that has an output label or numerical value of interest, which can be used to learn the association between the given features and the output label/numerical value. However, we will often come across datasets where there is no such pre-existing output label or numerical value or it is unknown. In such cases, we are simply interested in interesting patterns (clusters or groups) in the data; we need to mathematically define what we mean by "interesting" to find such patterns. The k-means algorithm lumps data points into k clusters or groups, where the elements of the cluster/group are considered to be similar; remember the discussion on measuring similarity between vectors in Section 1.11.

Let's assume that our dataset consists of N samples, each of which is a feature vector, $\mathbf{x}_i \in \mathbb{R}^n$, $1 \leq i \leq N$. We want to group the data points into k clusters, with $k < N$. Given the data points $\mathbf{x}_1 \ldots \mathbf{x}_N$, the k-means algorithm produces two outputs:

1. **Cluster means**: A set of k points $\mathbf{m}_j \in \mathbb{R}^n$, $1 \leq j \leq k$ that are supposed to be representatives of the k clusters identified. You can think of $\mathbf{m}_j$ to be a typical member of the j^{th} cluster.

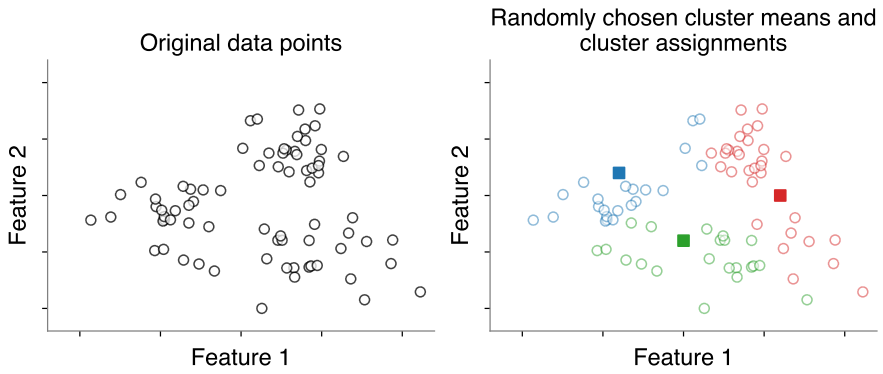


Figure 1.18: The clustering problem tackled by the k-means algorithm. The lft plot shows

2. **Cluster assignment**: Each data point $\mathbf{x}_i$ is given a cluster assignment j, such that $\mathbf{x}_i$ is closest to $\mathbf{m}_j$. This is an N-tuple of the form $(c_i)_{i=1}^N$, with $1 \leq c_i \leq k$. If $c_i = j$, then the data point $\mathbf{x}_i$ belongs to the j^{th} cluster.

This is demonstrated in Figure 1.18, where the left plot shows the dataset with N samples, each belonging to $\mathbb{R}^2$. The right plot shows three randomly chosen cluster means $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3$ and the cluster assignment of the data points to the closest $\mathbf{m}_j$. The three cluster means are shown in different colors and the corresponding data points in those clusters are shown in the same color (but a lighter shade). The goal for the k-means algorithm is to find the optimal cluster means and the clustern assignment such that the spread of points within each cluster is minimized across all clusters. We can measure this spread as the following,

$$J_{clust} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \tag{1.11}$$

where, $C_j = \{i : 1 \leq i \leq N, c_i = j\}$ is the set of indices of the data points that belong to the cluster j.

Minimizing J_{clust} for a given dataset is a computationally intensive problem. Note that the cluster means and the cluster assingments both depend on each other, and optimally choosing both of them to minimize J_{clust} is not easy.

The k-means simplifies the computational problem because, choosing the optimal cluster means for a fixed cluster assigment, and choosing the optimal cluster assingment for fixed cluster means is easy to do. The k-mean algorithm optmizes the cluster means and cluster assingments, while fixing the other, and iterates these steps until the cluster means and cluster assingments converge. This algorithm is detailed in Algorithm 1.3.

Algorithm 1.3: k-means Clustering Algorithm

Data: Dataset $\mathbf{x}_1 \dots \mathbf{x}_N$; number of cluster k tp be identified.
Result: A set of k cluster means $\mathbf{m}_1, \dots, \mathbf{m}_k$ and a N-tuple of cluster assingments $(c_i)_{i=1}^N$.

Choose a random set of k cluster means $\mathbf{m}_1, \dots, \mathbf{m}_k$;
repeat
  Update cluster assignment: For the current means find the best cluster assingment. $c_i = j$, such that $\|\mathbf{x}_i - \mathbf{m}_j\|_2 < \|\mathbf{x}_i - \mathbf{m}_l\|_2, \forall\ 1 \leq l \leq k, l \neq j$;
  Update cluster means: For the new cluster assignment, find the best cluster means. $\mathbf{m}_j = \frac{1}{|C_j|} \sum_{i \in C_j}$, where C_j is the set of indices of data points for the j^{th} cluster, and $|\cdot|$ is a function that returns the number of elements in a set;
until *until convergence*;
Return the cluster means $\mathbf{m}_1 \dots \mathbf{m}_k$ and the cluster assignment $(c_i)_{i=1}^N$;

This is demonstrated in Figure 1.18. The top left plot shows the dataset with N samples without any cluster assignment; all data points are colored black. The top right plot shows the first step in the iteration where the cluster means were chosen randomly; these are shown in different colored filled triangles. The corresponding optimal cluster assingment of points that are closest to each mean are also depicted in this plot.

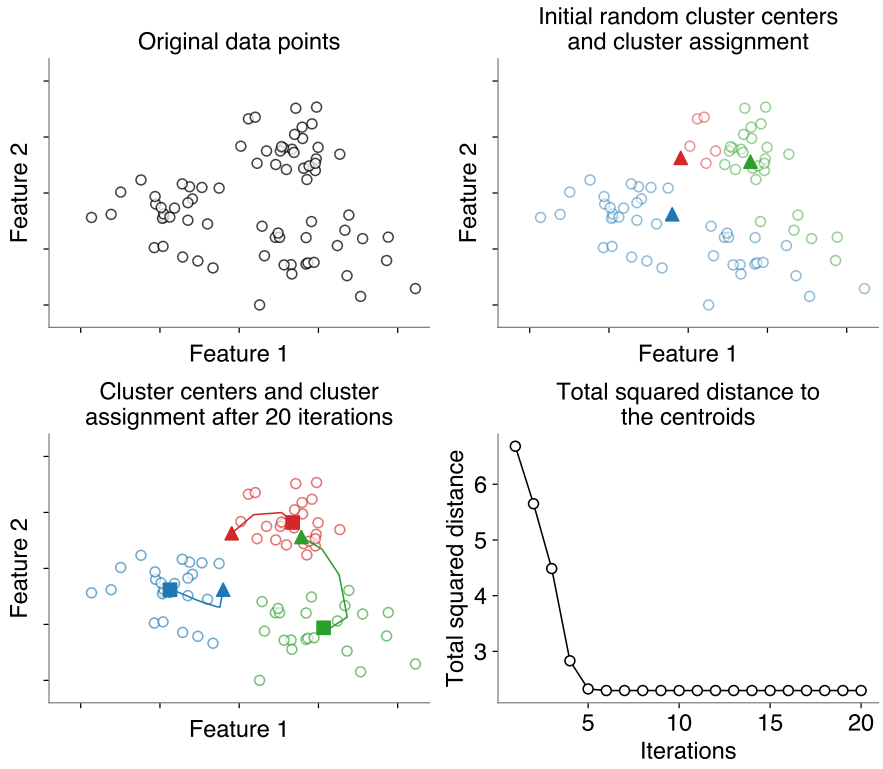


Figure 1.19: The clustering problem tackled by the k-means algorithm. The left plot shows

The bottom left plot shows trajectory of the cluster means as the k-means algorithm iterates repeating the process of updating the cluster means and cluster assingments. The two steps in of the k-means algorithm are guaranteed to reduce J_{clust}, and thus with each iteration the value of J_{clust} for the current cluster means and cluster assingments will reduce. This is depicted in the bottom right polot, which shows

the trend of J_{clust} as a function of the iteration number.

1.18 Exercise

1. Is this set of vectors $\left\{ \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$ independent? Explain your answer.

2. Show that the set $\{\mathbf{0}\}$, $\mathbf{0} \in \mathbb{R}^n$ a subspace of $\mathbb{R}^n$? This is called the trivial subspace of $\mathbb{R}^n$. What is the dimension of this subspace?

3. Prove the following for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

   (a) **Triangle Inequality**:
   $$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

   (b) **Backward Triangle Inequality**:
   $$\|\mathbf{x} - \mathbf{y}\| \geq |\|\mathbf{x}\| - \|\mathbf{y}\||$$

   (c) **Parallelogram Identity**:
   $$\frac{1}{2}\left(\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2\right) = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

4. Consider a set of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. When is $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{x} + \mathbf{y}\|$? What can you say about the geometry of the vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x} - \mathbf{y}$ and $\mathbf{x} + \mathbf{y}$?

5. Find the 1, 2 and ∞ norms of the following vectors from $\mathbb{R}^3$:

   (a) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ (c) $\mathbf{e}_3$ (d) $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ (e) $\mathbf{e}_1 - \mathbf{e}_2 + \mathbf{e}_3$

6. If $S_1, S_2 \subseteq V$ are subspaces of a vectors space V then, is $S_1 \cap S_2$ a subspace? Is $S_1 \cup S_2$ a subspace? Explain your answers.

7. Consider a vector $\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^\top$. Express the following in-terms of inner product between a constant vector $\mathbf{u}$ and the given vector $\mathbf{v}$, and in each case specify the vector $\mathbf{u}$.

   (a) $\sum_{i=1}^n v_i$
   (b) $\frac{1}{n}\sum_{i=1}^n v_i$
   (c) $\frac{1}{5}\sum_{i=3}^5 v_i$

8. Which of the following are linear functions of $\{x_1, x_2, \ldots, x_n\}$?

   (a) $\min_i \{x_i\}_{i=1}^n$
   (b) $\left(\sum_{i=1}^n x_i^2\right)^{1/2}$
   (c) x_6

9. Consider a linear function $f : \mathbb{R}^n \to \mathbb{R}$. Prove that every linear function of this form can be represented in the following form.
   $$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{i=1}^n w_i x_i, \qquad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n$$

10. An *affine* function f is defined as the sum of a linear function and a constant. It can in general be represented in the form,

$$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \beta, \qquad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n,\ \beta \in \mathbb{R}$$

Prove that affine functions are not linear. Prove that any affine function can be represented in the form $\mathbf{w}^\top \mathbf{x} + \beta$.

11. Consider a basis $B = \{\mathbf{b}_i\}_{i=1}^n$ of $\mathbb{R}^n$. Let the vectors $\mathbf{x}$ and $\mathbf{x}_b$ be the representations in the standard and B basis respectively.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n x_i \mathbf{e}_i \quad \text{and} \quad \mathbf{x}_b = \begin{bmatrix} x_{b1} \\ x_{b2} \\ \vdots \\ x_{bn} \end{bmatrix} = \sum_{i=1}^n x_{bi} \mathbf{b}_i$$

Evaluate the $\|\mathbf{x}\|_2^2$ and $\|\mathbf{x}_b\|_2^2$. Determine what happens to $\|\mathbf{x}_b\|_2^2$ under the following conditions on the basis vectors: **[Marks: 2]**

(a) $\|\mathbf{b}_i\| = 1, \forall i$

(b) $\left\|\mathbf{b}_i^\top \mathbf{b}_j\right\| = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

12. **[Programming]** Let's build a simple classifier using the concepts you've learned in this chapter. We will pretend that we are doing this for classifying or detecting the presence or absence of a disease called - *vector space sickness*. We wish to diagnoze using two clinical tests - (1) *Subspace assay* and (2) *Basis balance scale*. Both these serious clinical tests generate numerical outcomes that can takeon any real number value.

The department has been conducting a large scale clinical study for the last 5-6 years collecting data from participants, from different background, with and without vector space sickness by administering the subspace assay and the basis balance test. The data from this study is stored as a CSV file with four columns: (a) `subjectno` – subject numbers, (b) `x1` – value of the subspace assay test, (c) `x2` – value of the basis balance scale, and (d) `vss` – presence (1) or absence (0) of the vector space sickness condition. Each row of this CSV file corresponds to a individual subject who participated in the study. Your goal here is to look at the data from this experiment and propose a classifier to determine if a person has vector space sickness if we are given their scores on the subspace assay and the basis balance scale clinical tests. A group of Master's students participated in the study. Unfortunately, almost half of these students were diagnosed with vector space sickness. This data is stored in `expt1.csv`. Read this data, make a scatter plot in 2D (`x1` versus `x2`) with the data points for participants with vector space sickness in blue and the one without in red. Look at the data, and propose how you could use the measurements `x1` and `x2` to distinguish between participants from the two groups. Implement the classifier you've proposed and find out how well it performs in correctly classifying the two groups.

13. **[Programming]** Consider a set of measurements made from adult male subjects, where their height, weight and BMI (body mass index) were recorded and stored as vectors of length three; the first element is the height in cm, second is the weight in Kg, and the last is the BMI. Consider the

following four subjects,

$$\mathbf{s}_1 = \begin{bmatrix} 167 \\ 102 \\ 36.6 \end{bmatrix} ; \mathbf{s}_2 = \begin{bmatrix} 180 \\ 87 \\ 26.9 \end{bmatrix} \mathbf{s}_3 = \begin{bmatrix} 177 \\ 78 \\ 24.9 \end{bmatrix} ; \mathbf{s}_4 = \begin{bmatrix} 152 \\ 76 \\ 32.9 \end{bmatrix}$$

You can use the distance between these vectors $\|\mathbf{s}_i - \mathbf{s}_j\|_2$ as a measure of the similiarity between the four subjects. Generate a 4×4 table comparing the distance of each subject with respect to another subject; the diagonal elements of this table will be zero, and it will be symmetric about the main diagonal.

(a) Based on this table, how do the different subjects compare to each other?

(b) How do the similarities change if the height had been measured in m instead of cm? Can you explain this difference?

(c) Consider the weighted norm presented in one of the earlier problems.

$$\|x\|_{\mathbf{w}} = \left(w_1 x_1^2 + w_2 x_2^2 + \cdots + w_n x_n^2\right)^{\frac{1}{2}}$$

Will this fix the problem? What would be a good choice for $\mathbf{w}$ to address the problems with comparing distance between vectors due to unit change?

(d) Can the angle between two vectors be used as a measure of similarity between vectors? Does this suffer from the problem of the 2-norm $\|\cdot\|_2$?

Part II

Optimization

Part III

Probability and Statistics

Part IV

Least Squares