MATH 3332: Data Analytic Tools

HU-HTAKM
Website: https://htakm.github.io/htakm_test/

Last major change: October 18, 2025
Last small update: October 18, 2025

Contents

Chapter 1

Introduction

Machine Learning is a type of artificial intelligence that focuses on the development of algorithms and models to perform tasks without being explicitly programmed to do so. Machine learning algorithms create a model based on sample data, known as training data. There are three common types of machine learning:

1. Supervised learning: Classification, Regression
   Data: N input-output pairs

   $$(\mathbf{x}_i, \mathbf{y}_i) \qquad \text{for } \mathbf{x}_i \in X, \mathbf{y}_i \in Y,\ i = 1, \cdots, N.$$

   Goal: Find a function map $f : X \to Y$ such that:

   $$f(\mathbf{x}_i) \approx \mathbf{y}_i \qquad \text{for } i = 1, \cdots, N.$$

   If a new input $\mathbf{x}$ is provided, $f(\mathbf{x})$ should accurately predict the label of $\mathbf{x}$.

2. Unsupervised learning: Clustering, Self-supervised Learning
   Data: N inputs without labels

   $$\mathbf{x}_i \qquad \text{for } \mathbf{x}_i \in X,\ i = 1, \cdots, N.$$

   Goal: Different applications have their own goals. For example, in the case of denoising, find a function map f such that:

   $$f(\mathbf{x}_i + \boldsymbol{\varepsilon}_i) = \mathbf{x}_i \qquad \text{for } i = 1, \cdots, N,$$

   where $\boldsymbol{\varepsilon}_i$ are noise vectors.

3. Reinforcement learning (Reinforcement learning algorithms are usually iterative algorithms).

In general, we want to find a good function f that maps the training data well while generalizing to other inputs.

Definition 1.1. The set of all 'good' candidate functions (models) is called the **hypothesis space**.

In our case studies, we will follow Pedro Domingos' definition of machine learning:

$$\text{Learning} = \text{Representation} + \text{Evaluation} + \text{Optimization}.$$

1. Representation: Mainly focuses on 'vector' representations.
   (a) How can we effectively represent the input data $\mathbf{x}_i$?
   (b) How can we represent the function f?
2. Evaluation: Evaluate the problem.
   (a) How do we define 'the best' function in the hypothesis space?
   We need to define a function $f' : f \to \mathbb{R}$ to compare.
   (b) How do we define 'the best' representation of the input data?
3. Optimization: Find the optimal model.
   (a) How can we obtain the optimal solution numerically using a computer?
   (b) Is convex optimization feasible? (Some problems involve non-convex optimization.)

Chapter 2

Vector spaces, metrics, limits, and convergence

2.1 Vector spaces (linear space)

Definition 2.1. A **vector space** over $\mathbb{R}$ is a set V together with two operations:

1. Addition: For all $\mathbf{u}, \mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} \in V$.
2. Scalar multiplication: For all $\alpha \in \mathbb{R}$ and $\mathbf{v} \in V$, $\alpha\mathbf{v} \in V$.

These two operations satisfy the following eight properties for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\alpha, \beta \in \mathbb{R}$:

$(+1)$	$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$,	(Addition Commutativity)
$(+2)$	$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$,	(Addition Associativity)
$(+3)$	There exists $\mathbf{0} \in V$ such that $\mathbf{u} + \mathbf{0} = \mathbf{u}$,	(Zero Exists)
$(+4)$	For every $\mathbf{u} \in V$, there exists $\mathbf{u}' \in V$ such that $\mathbf{u} + \mathbf{u}' = \mathbf{0}$ $(\mathbf{u}' = -\mathbf{u})$,	(Additive Inverse Exists)
$(\cdot 1)$	$(\alpha\beta)\mathbf{u} = \alpha(\beta\mathbf{u})$,	(Multiplication Associativity)
$(\cdot 2)$	$1 \cdot \mathbf{u} = \mathbf{u}$,	(Unity)
$(\cdot 3)$	$\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$,	(Distributivity 1)
$(\cdot 4)$	$(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}$.	(Distributivity 2)

Remark 2.1.1. A vector space over the complex domain $\mathbb{C}$ can be defined in a similar way.

Example 2.1. The set of real numbers $\mathbb{R}$, with the standard addition and multiplication of real numbers, forms a vector space.

Example 2.2. The n-dimensional Euclidean space $\mathbb{R}^n$, with the following operations, forms a vector space over $\mathbb{R}$:

$$+ : \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \ \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \in \mathbb{R}^n,$$

$$\bullet : \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \text{ and } \alpha \in \mathbb{R}, \ \alpha \cdot \mathbf{x} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix} \in \mathbb{R}^n.$$

Remark 2.1.2. Many types of input data can be modeled as vectors in $\mathbb{R}^n$, such as:

1. Digital signals of length n,
2. Stock prices over n time intervals,
3. n different features or attributes of a single object.

Example 2.3. All real $m \times n$ matrices $\mathbb{R}^{m\times n}$ with

$$+: \quad \text{For all } \mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n},$$

$$\text{we have } \mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \dots & a_{mn} + b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n},$$

$$\bullet: \quad \text{For all } \mathbf{B} = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n} \text{ and } \alpha \in \mathbb{R}, \text{ we have } \alpha \cdot \mathbf{B} = \begin{pmatrix} \alpha b_{11} & \dots & \alpha b_{1n} \\ \vdots & \ddots & \vdots \\ \alpha b_{m1} & \dots & \alpha b_{mn} \end{pmatrix} \in \mathbb{R}^{m\times n}$$

is a vector space over $\mathbb{R}$.

Remark 2.1.3. This vector space is equivalent to $\mathbb{R}^{mn}$:

$$\begin{pmatrix} x_{11} & \dots & x_{1n} \\ x_{21} & \dots & x_{2n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} \xrightarrow{\text{vectorization}} \begin{pmatrix} x_{11} \\ \vdots \\ x_{1n} \\ x_{21} \\ \vdots \\ x_{2n} \\ \vdots \\ x_{mn} \end{pmatrix} \in \mathbb{R}^{mn}.$$

Remark 2.1.4. An $m \times n$ matrix can be used to represent a black-and-white digital image.

Example 2.4. All real 3-arrays of size $m \times n \times \ell$, $\mathbb{R}^{m\times n\times \ell}$, with

$+:$ For all $X = (x_{ijk})_{i,j,k}, Y = (y_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$, we have $X + Y = (x_{ijk} + y_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$,

$\bullet:$ For all $X = (x_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$ and $\alpha \in \mathbb{R}$, we have $\alpha \cdot X = (\alpha x_{ijk})_{i,j,k} \in \mathbb{R}^{m\times n\times \ell}$

is a vector space over $\mathbb{R}$.

Remark 2.1.5. Similar to matrices, this vector space is equivalent to $\mathbb{R}^{mn\ell}$.

Remark 2.1.6. Many types of data can be modeled by this vector space $\mathbb{R}^{m\times n\times \ell}$, such as:

1. Color images with 3 color channels (RGB) ($\ell = 3$),
2. Black-and-white videos with ℓ frames, each of size $m \times n$.

More complex data, such as color videos, are represented as 4-arrays. These arrays are usually collectively called **tensors**.

Example 2.5. Consider the set of all strings. We can quickly see that strings do not satisfy the commutativity property:

$$\text{'Standing'} = \text{'Stand'} + \text{'ing'} \neq \text{'ing'} + \text{'Stand'} = \text{'ingStand'}.$$

Therefore, vector spaces cannot be used to model text data.

Example 2.6. The set of all continuous functions on $[a, b]$, denoted by:

$$\mathcal{C}[a, b] = \{f : f \text{ is a continuous function on } [a, b]\},$$

with, for all $t \in [a, b]$,

$$\begin{aligned} + &: \quad \text{For all } f, g \in \mathcal{C}[a, b], \text{ we have } (f + g)(t) = f(t) + g(t) \in \mathcal{C}[a, b], \\ \bullet &: \quad \text{For all } f \in \mathcal{C}[a, b] \text{ and } \alpha \in \mathbb{R}, \text{ we have } (\alpha f)(t) = \alpha f(t) \in \mathcal{C}[a, b]. \end{aligned}$$

is a vector space over $\mathbb{R}$. It is referred to as a **function space**.

Remark 2.1.7. $\mathcal{C}[a, b]$ can be considered as a hypothesis space of a learner with one input and one output. Given a dataset of $x_i \in [a, b]$ and $y_i \in \mathbb{R}$, find the function $f \in \mathcal{C}[a, b]$ such that $f(x_i) \approx y_i$ for all i.

Example 2.7. The infinite sequences:

$$\ell_\infty = \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ \vdots \end{pmatrix} : \text{There exists } c < \infty \text{ such that } |a_i| \leq c \text{ for any } i \right\},$$

with

$$\begin{aligned} + &: \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \vdots \end{pmatrix} \in \ell_\infty, \text{ we have } \mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \\ \vdots \end{pmatrix} \in \ell_\infty, \\ \bullet &: \quad \text{For all } \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \end{pmatrix} \in \ell_\infty \text{ and } \alpha \in \mathbb{R}, \text{ we have } \alpha \cdot \mathbf{x} = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \\ \vdots \end{pmatrix} \in \ell_\infty. \end{aligned}$$

is a vector space over $\mathbb{R}$.

2.2 Metrics on vector space

We can convert some types of input data into vector space. However, how do we determine the distance between two vectors? Our goal is to define the distance in order to perform calculus.
Let V be a vector space and $\mathbf{u}, \mathbf{v} \in V$. We want to find a function such that:

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \text{dist}(\mathbf{0}, \mathbf{u} - \mathbf{v}) = \text{length of } \mathbf{u} - \mathbf{v}.$$

Therefore, to define the distance, we only need to define the length of vectors.

Definition 2.2. Let V be a vector space over $\mathbb{R}$. A **norm** on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in V$,

1. $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$.
2. $\|\alpha\mathbf{x}\| = |\alpha| \|\mathbf{x}\|$ for $\alpha \in \mathbb{R}$.
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

The ordered pair $(V, \|\cdot\|)$ is called a **normed vector space**.

Remark 2.2.1. If it is clear from the context which norm is intended, then it is common to denote the normed vector space by V.

Remark 2.2.2. $\|\mathbf{x}\|$ represents the **length** of $\mathbf{x}$.

Remark 2.2.3. The distance between $\mathbf{x}$ and $\mathbf{y}$ can now be defined as $\text{dist}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Example 2.8. If we set $V = \mathbb{R}$, the following can be norms on $\mathbb{R}$ for all $x \in \mathbb{R}$:

1. $\|x\| = |x|$,
2. $\|x\| = \frac{1}{2}|x|$,
3. $\|x\| = c|x|$ for some $c > 0$.

Remark 2.2.4. In fact, there are infinitely many norms on the same vector space.

Example 2.9. For $\mathbf{x} \in \mathbb{R}^n$, the **Manhattan norm** (1-norm or L_1-norm) defined by:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

is a norm on $\mathbb{R}^n$. The induced distance $\|\mathbf{x} - \mathbf{y}\|_1$ for $\mathbf{x}, \mathbf{y} \in V$ is called the **Manhattan distance**.

Proof.

1. For any $\mathbf{x}, \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \geq 0.$$

Moreover,

$$\begin{aligned} \|\mathbf{x}\|_1 = 0 &\iff \sum_{i=1}^{n} |x_i| = 0 \\ &\iff |x_i| = 0 \iff x_i = 0 \qquad \text{for } i = 1, \cdots, n, \\ &\iff \mathbf{x} = \mathbf{0}. \end{aligned}$$

2. For any $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,

$$\|\alpha\mathbf{x}\|_1 = \sum_{i=1}^{n} |\alpha x_i| = |\alpha| \sum_{i=1}^{n} |x_i| = |\alpha| \|\mathbf{x}\|_1.$$

3. Using the Triangle Inequality, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x} + \mathbf{y}\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} (|x_i| + |y_i|) = \|\mathbf{x}\|_1 + \|\mathbf{y}\|_1.$$

Therefore, by definition, the 1-norm is a norm on $\mathbb{R}^n$. □

Example 2.10. For $\mathbf{x} \in \mathbb{R}^n$, the Euclidean norm (2-norm or L_2-norm) defined by:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

is a norm on $\mathbb{R}^n$. The induced distance $\|\mathbf{x}-\mathbf{y}\|_2$ for $\mathbf{x}, \mathbf{y} \in V$ is called the **Euclidean distance**.

Proof.

1. For $\mathbf{x} \in \mathbb{R}^n$,
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \geq 0.$$
Moreover,
$$\begin{aligned} \|\mathbf{x}\|_2 = 0 &\iff \sum_{i=1}^{n} x_i^2 = 0 \\ &\iff x_i^2 = 0 \iff x_i = 0 \qquad \text{for } i = 1, \cdots, n, \\ &\iff \mathbf{x} = \mathbf{0}. \end{aligned}$$

2. For any $\alpha \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$,
$$\|\alpha\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} (\alpha x_i)^2} = \sqrt{\alpha^2}\sqrt{\sum_{i=1}^{n} x_i^2} = |\alpha| \|\mathbf{x}\|_2 .$$

3. This is a bit more complicated, but it is similar to the proof of the Cauchy-Schwarz inequality.
For $\mathbf{x} \in \mathbb{R}^n$, if $\mathbf{y} = \mathbf{0}$, then:
$$\|\mathbf{x}+\mathbf{y}\|_2 = \|\mathbf{x}\|_2 = \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$$
If $\mathbf{y} \neq 0$, then for any $t \in \mathbb{R}$,
$$\|\mathbf{x}+t\mathbf{y}\|_2^2 = \sum_{i=1}^{n} (x_i + ty_i)^2 = \left(\sum_{i=1}^{n} x_i^2\right) + t\left(2\sum_{i=1}^{n} x_i y_i\right) + t^2\left(\sum_{i=1}^{n} y_i^2\right).$$
By (1), $\|\mathbf{x}+t\mathbf{y}\|_2^2 \geq 0$. Therefore, since $\sum_{i=1}^{n} y_i^2 > 0$, $\|\mathbf{x}+t\mathbf{y}\|_2^2$ is a quadratic function with at most one real root. Using the discriminant, we have:
$$\begin{aligned} \Delta &\leq 0, \\ \left(2\sum_{i=1}^{n} x_i y_i\right)^2 - 4\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right) &\leq 0, \\ \left(\sum_{i=1}^{n} x_i y_i\right)^2 &\leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right), \\ \sum_{i=1}^{n} x_i y_i &\leq \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}. \end{aligned} \qquad \text{(Cauchy-Schwarz Inequality for } \mathbb{R}^n)$$
Consequently,
$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|_2^2 &= \sum_{i=1}^{n} (x_i + y_i)^2 \\ &= \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &\leq \sum_{i=1}^{n} x_i^2 + 2\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} + \sum_{i=1}^{n} y_i^2 = \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2, \\ \|\mathbf{x}+\mathbf{y}\|_2 &\leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2 . \end{aligned}$$

Therefore, by definition, the 2-norm is a norm on $\mathbb{R}^n$. □

Example 2.11. For $\mathbf{x} \in \mathbb{R}^n$, the p-norm or L_p-norm with $p \geq 1$ defined by:

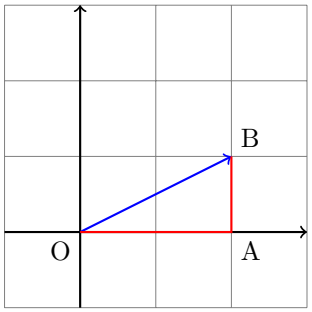
$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

is a norm on $\mathbb{R}^n$.

Example 2.12. For $\mathbf{x} \in \mathbb{R}^n$, the maximum norm ($L_\infty$-norm) defined by:

$$\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p = \max_{1 \leq i \leq n} |x_i|$$

is a norm on $\mathbb{R}^n$.



Red: Manhattan distance from O to B

$|\text{OA}| + |\text{AB}|$

Blue: Euclidean distance from O to B

$|\text{OB}|$

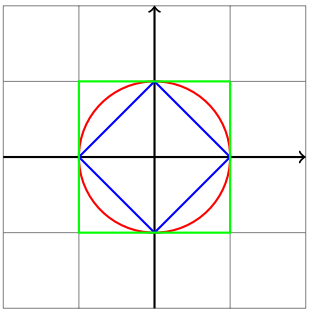
Figure 2.1: Difference between Manhattan distance and Euclidean distance

Definition 2.3. The **open unit ball** of a norm $\|\cdot\|$ is defined by:

$$B = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < 1\}.$$

Example 2.13. For p-norm, the open unit ball of $\|\cdot\|_p$ is defined by:

$$B_p = \left\{ \mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} < 1 \right\}.$$



$B_2 = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 < 1\}$

$B_1 = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 < 1\}$

$B_\infty = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_\infty < 1\}$

Figure 2.2: Unit balls of p-norm for different p

Theorem 2.4. If $0 < q \leq p < \infty$, then for any $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_q .$$

Remark 2.4.1. There are other norms on $\mathbb{R}^n$, not just p-norms.

For a set of matrices $\mathbb{R}^{m\times n}$, we can view them as $\mathbb{R}^{mn}$.

Example 2.14. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, the **vector p-norm** with $p\geq 1$ defined by:

$$\|\mathbf{A}\|_{p,\text{vec}}=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{\frac{1}{p}}$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.15. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=2$, the **Frobenius norm** (vector 2-norm) defined by:

$$\|\mathbf{A}\|_F=\|\mathbf{A}\|_{2,\text{vec}}=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}a_{ij}^2}$$

is a norm on $\mathbb{R}^{m\times n}$.

We can also view $\mathbb{R}^{m\times n}$ as a linear transformation $\mathbb{R}^n\to\mathbb{R}^m$. Given a vector $\mathbf{x}\in\mathbb{R}^n$, one can consider the function:

$$f(\mathbf{x})=\mathbf{A}\mathbf{x}\in\mathbb{R}^m$$

as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Example 2.16. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, the **matrix p-norm** with $p\geq 1$ defined by:

$$\|\mathbf{A}\|_p=\max_{\substack{\mathbf{x}\neq\mathbf{0}\\ \mathbf{x}\in\mathbb{R}^n}}\frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_p=1}}\|\mathbf{A}\mathbf{x}\|_p$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.17. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=1$, the matrix 1-norm defined by:

$$\|\mathbf{A}\|_1=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_1=1}}\|\mathbf{A}\mathbf{x}\|_1=\max_{1\leq j\leq n}\sum_{i=1}^{m}|a_{ij}|$$

is a norm on $\mathbb{R}^{m\times n}$.

Example 2.18. For $\mathbf{A}\in\mathbb{R}^{m\times n}$, when $p=2$, the matrix 2-norm can be described with the following properties:

$$\|\mathbf{A}\|_2^2=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_2=1}}\|\mathbf{A}\mathbf{x}\|_2^2=\max_{\substack{\mathbf{x}\in\mathbb{R}^n\\ \|\mathbf{x}\|_2=1}}\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}=\text{max eigenvalue of }\mathbf{A}^T\mathbf{A},$$

$$\|\mathbf{A}\|_2=\sqrt{\text{max eigenvalue of }\mathbf{A}^T\mathbf{A}}=\text{max singular value of }\mathbf{A}.$$

Therefore, the matrix 2-norm is also called the **operator norm** of $\mathbf{A}$.

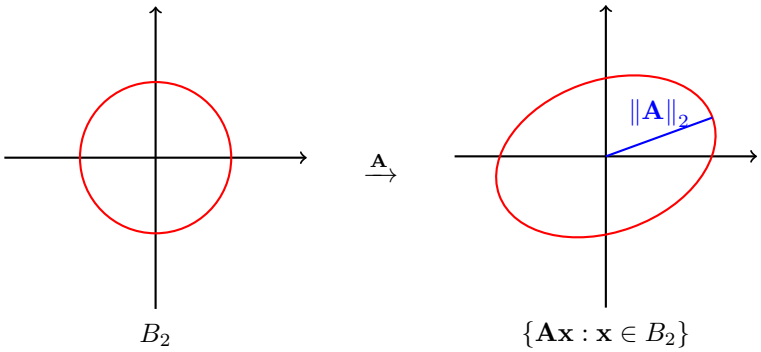


Figure 2.3: Meaning of matrix 2-norm when $n=2$ and $m=2$

In addition, the matrix p-norm can be generalized with two different p-norms.

Example 2.19. For $\mathbf{A} \in \mathbb{R}^{m\times n}$, the matrix norm induced by p-norm in $\mathbb{R}^n$ and q-norm in $\mathbb{R}^m$ defined by:

$$\|\mathbf{A}\|_{p\to q} = \max_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_p=1}} \|\mathbf{A}\mathbf{x}\|_q$$

is a norm on $\mathbb{R}^{m\times n}$. Moreover, the matrix p-norm is a special case of this norm.

Matrix norms do not need to be defined using vector norms.

Example 2.20. For $\mathbf{A} \in \mathbb{R}^{m\times n}$, the **nuclear norm** (trace norm or Ky Fan n-norm) defined by:

$$\|\mathbf{A}\|_* = \mathrm{Tr}(\sqrt{\mathbf{A}^T\mathbf{A}})$$

is a norm on $\mathbb{R}^{m\times n}$.

How about norms for continuous functions?

Example 2.21. For $f \in \mathcal{C}[a,b]$, where the set is defined as:

$$\mathcal{C}[a,b] = \{f : f \text{ is a continuous function on } [a,b]\},$$

the **maximum norm** (Chebyshev norm) defined by:

$$\|f\|_\infty = \max_{t\in[a,b]} |f(t)|$$

is a norm on $\mathcal{C}[a,b]$.

Example 2.22. For $f \in \mathcal{C}[a,b]$, the **p-norm** with $p \geq 1$ defined by:

$$\|f\|_p = \left(\int_a^b |f(t)|^p \, dt\right)^{\frac{1}{p}}$$

is a norm on $\mathcal{C}[a,b]$.

Is there a norm for the set of vectors with infinite dimensions?

Example 2.23. For $\mathbf{x} \in \ell_\infty$, where the set is defined as:

$$\ell_\infty = \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_n \\ \vdots \end{pmatrix} : \text{There exists } c < \infty \text{ such that } |a_i| \leq c \text{ for any } i \right\},$$

the **supremum norm** defined by:

$$\|\mathbf{x}\|_\infty = \sup_i |x_i|$$

is a norm on ℓ_∞.

Example 2.24. For $\mathbf{x} \in \ell_p$ with $p \geq 1$, where the set is defined as:

$$\ell_p = \{\mathbf{x} \in \ell_\infty : \|\mathbf{x}\|_p < \infty\} \subset \ell_\infty,$$

the p-norm defined by:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^\infty |x_i|^p\right)^{\frac{1}{p}}$$

is a norm on ℓ_p. However, it is important to note that it is not a norm on ℓ_∞.

Remark 2.4.2. For the same vector space, we can define infinitely many norms on it. The simplest would be by adding or subtracting different norms.

2.3 Limit and convergence on normed vector space

In this section, assume that V is a vector space over $\mathbb{R}$ with norm $\|\cdot\|$.
In data analysis, many algorithms are iterative algorithms, which generate n sequences of vectors:

$$\{\mathbf{x}_i^{(k)}\} \subset V, \qquad i = 1, \cdots, n,$$

where V is endowed with a norm $\|\cdot\|$. As the iteration continues, it is important that the output stays within the hypothesis space. If the algorithm generates an output that is outside the expected domain, then it is not considered a good solution.

Definition 2.5. Let $\mathbf{x} \in V$. We say the sequence $\{\mathbf{x}^{(k)}\} \in V$ **converges** to $\mathbf{x}$, denoted by $\mathbf{x}^{(k)} \to \mathbf{x}$, if:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\| = 0.$$

More rigorously, the sequence $\{\mathbf{x}^{(k)}\}$ **converges** to $\mathbf{x}$ if, for any $\varepsilon > 0$, there exists N such that for any $n \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}\| < \varepsilon.$$

Example 2.25. Consider $V = \mathbb{R}^n$ with $\|\cdot\|_2$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} \frac{1}{k} \\ \vdots \\ \frac{n}{k} \end{pmatrix}, \qquad \mathbf{x} = \mathbf{0}.$$

Then we have:

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \|\mathbf{x}^{(k)}\|_2 = \sqrt{\sum_{i=1}^{n} \left(\frac{i}{k}\right)^2} = \frac{1}{k}\sqrt{\sum_{i=1}^{n} i^2},$$

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \lim_{k\to\infty} \frac{1}{k}\sqrt{\sum_{i=1}^{n} i^2} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$.

Example 2.26. Consider $V = \mathcal{C}[0,1]$ with $\|\cdot\|_\infty$. Let:

$$f^{(k)}(t) = \frac{\sin(2\pi k t)}{k^2} \in \mathcal{C}[0,1].$$

Let 0 be the zero function. We can easily find that $0 \in \mathcal{C}[0,1]$. Then we have:

$$\|f^{(k)} - 0\|_\infty = \|f^{(k)}\|_\infty = \max_{t\in[0,1]} \left|\frac{\sin(2\pi k t)}{k^2}\right| = \frac{1}{k^2},$$

$$\lim_{k\to\infty} \|f^{(k)} - 0\|_\infty = \lim_{k\to\infty} \frac{1}{k^2} = 0.$$

Therefore, $f^{(k)} \to 0$.

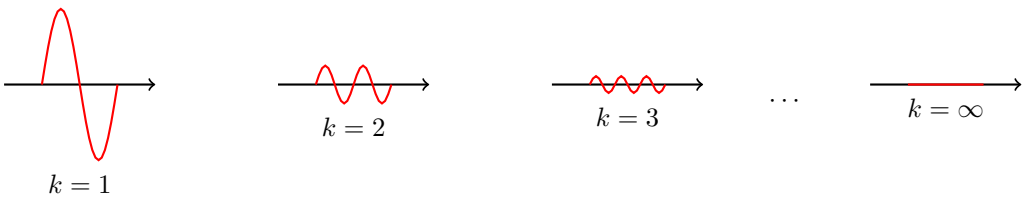


Figure 2.4: As k increases, the wave keeps shrinking in amplitude.

Remark 2.5.1. Convergence depends on the norm used.

Example 2.27. Consider $V = \ell_p$ for any p with $\|\cdot\|_p$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} \frac{1}{k} \\ \vdots \\ \frac{1}{k} \\ 0 \\ \vdots \end{pmatrix} \begin{matrix} \\ \\ k\text{-terms} \\ \\ \\ \end{matrix} = \sum_{i=1}^{k} \frac{1}{k}\mathbf{e}_i \in \ell_p, \qquad \mathbf{x} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{pmatrix} = \mathbf{0} \in \ell_p.$$

When $p = 2$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \|\mathbf{x}^{(k)}\|_2 = \sqrt{\sum_{i=1}^{k} \frac{1}{k^2}} = \frac{1}{\sqrt{k}},$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_2 = \lim_{k\to\infty} \frac{1}{\sqrt{k}} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_2$.

When $p = \infty$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \|\mathbf{x}^{(k)}\|_\infty = \frac{1}{k},$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \lim_{k\to\infty} \frac{1}{k} = 0.$$

Therefore, $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_\infty$.

When $p = 1$,

$$\|\mathbf{x}^{(k)} - \mathbf{x}\|_1 = \|\mathbf{x}^{(k)}\|_1 = \sum_{i=1}^{k} \frac{1}{k} = 1,$$
$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_1 = \lim_{k\to\infty} 1 = 1 \neq 0.$$

Therefore, $\mathbf{x}^{(k)} \not\to \mathbf{x}$ with the norm $\|\cdot\|_1$.

Remark 2.5.2. The limit may not be in the same vector space. If this happens, the normed vector space is called **incomplete**.

Example 2.28. Consider $V = \ell_1$ with $\|\cdot\|_\infty$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{k} \\ 0 \\ \vdots \end{pmatrix} = \sum_{i=1}^{k} \frac{1}{i}\mathbf{e}_i \in \ell_1, \qquad \mathbf{x} = \begin{pmatrix} 1 \\ \frac{1}{2} \\ \vdots \\ \frac{1}{k} \\ \frac{1}{k+1} \\ \vdots \end{pmatrix} = \sum_{i=1}^{\infty} \frac{1}{i}\mathbf{e}_i.$$

Then we have:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_\infty = \lim_{k\to\infty} \left\| \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{k+1} \\ \vdots \end{pmatrix} \right\|_\infty = \lim_{k\to\infty} \frac{1}{k+1} = 0.$$

However, $\sum_{i=1}^{\infty} \frac{1}{i} = \infty$, and thus $\mathbf{x} \notin \ell_1$. Therefore, we cannot say that $\mathbf{x}^{(k)} \to \mathbf{x}$ with the norm $\|\cdot\|_\infty$.

With the definition of convergence, we can define the completeness of a vector space.

Definition 2.6. The sequence $\{\mathbf{x}^{(k)}\} \subset V$ is a Cauchy sequence if, for any $\varepsilon > 0$, there exists N such that for any $n, m \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| < \varepsilon.$$

Theorem 2.7. If $\mathbf{x}^{(k)} \to \mathbf{x}$ in $(V, \|\cdot\|)$, then $\{\mathbf{x}^{(k)}\}$ is a Cauchy sequence.

Proof.
If $\mathbf{x}^{(k)} \to \mathbf{x}$, then for all $\varepsilon > 0$, there exists N such that for all $n \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}\| < \frac{\varepsilon}{2}.$$

Therefore, for all $n, m \geq N$,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| \leq \|\mathbf{x}^{(n)} - \mathbf{x}\| + \|\mathbf{x}^{(m)} - \mathbf{x}\| < \varepsilon.$$

□

Remark 2.7.1. The converse is not necessarily true.

Definition 2.8. A normed vector space $(V, \|\cdot\|)$ is **complete** if the limit of all Cauchy sequences in V is in V.

Definition 2.9. A complete normed vector space is called a **Banach space**.

Example 2.29. $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$, or $\mathbb{R}^{m\times n\times \ell}$ with any norm is a Banach space.

Example 2.30. $\mathcal{C}[a, b]$ with $\|\cdot\|_\infty$ is a Banach space.

Example 2.31. For $p \geq 1$, including $p = \infty$, $(\ell_p, \|\cdot\|_p)$ is a Banach space.

Remark 2.9.1. We can always include all the limits of the Cauchy sequences to convert an incomplete normed vector space into a complete one.

Example 2.32. $(\ell_1, \|\cdot\|_\infty)$ is an incomplete normed vector space. Its completion is ℓ_∞.

Example 2.33. For $p \geq 1$, $(\mathcal{C}[a, b], \|\cdot\|_p)$ is an incomplete normed vector space. Its completion is $L^p[a, b]$.

In practical cases, how do we check the convergence of a given sequence?

Example 2.34. From an iterative algorithm, we generate a sequence of vectors $\{\mathbf{x}^{(k)}\}$. Our goal is to check if this sequence converges. Pick a threshold $\varepsilon > 0$. We can check computationally that for large n, m,

$$\|\mathbf{x}^{(n)} - \mathbf{x}^{(m)}\| < \varepsilon.$$

Practically, to reduce computational cost, we usually only check for large n,

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| < \varepsilon.$$

2.4 Finite Dimensional Vector Spaces

In most cases, we deal with finite-dimensional vector spaces.

Remark 2.9.2. Every finite-dimensional vector space with any norm is complete. That is, any finite-dimensional vector space is Banach.

Remark 2.9.3. For a finite-dimensional vector space V, all norms are equivalent. For any two norms $\|\cdot\|_A$ and $\|\cdot\|_B$, there exist $c_1, c_2 > 0$ such that:

$$c_1 \|\mathbf{x}\|_A \leq \|\mathbf{x}\|_B \leq c_2 \|\mathbf{x}\|_A , \qquad \text{for all } \mathbf{x} \in V.$$

Theorem 2.10. The limit of the same finite-dimensional sequence under any norm is the same. That means, given two finite-dimensional normed vector spaces $(V, \|\cdot\|_A)$ and $(V, \|\cdot\|_B)$, for any sequence $\{\mathbf{x}^{(k)}\}$ and $\mathbf{x} \in V$,

$$\mathbf{x}^{(k)} \to \mathbf{x} \text{ in } \|\cdot\|_A \iff \mathbf{x}^{(k)} \to \mathbf{x} \text{ in } \|\cdot\|_B .$$

Proof.
Since $\mathbf{x}^{(k)} \to \mathbf{x}$ in $\|\cdot\|_A$,

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A = 0.$$

Therefore, there exist $c_1, c_2 > 0$ such that:

$$c_1 \|\mathbf{x}^{(k)} - \mathbf{x}\|_A \leq \|\mathbf{x}^{(k)} - \mathbf{x}\|_B \leq c_2 \|\mathbf{x}^{(k)} - \mathbf{x}\|_A.$$

Taking $k \to \infty$, we have:

$$0 \leq c_1 \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A \leq \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_B \leq c_2 \lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_A = 0.$$

By the Squeeze Theorem, we find that:

$$\lim_{k\to\infty} \|\mathbf{x}^{(k)} - \mathbf{x}\|_B = 0.$$

To conclude, $\mathbf{x}^{(k)} \to \mathbf{x}$ in $\|\cdot\|_B$. The proof for the converse is similar. □

Example 2.35. Consider $V = \mathbb{R}^n$ with $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$.

1. $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\, \|\mathbf{x}\|_2 .$$

2. $\|\cdot\|_2$ and $\|\cdot\|_\infty$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\, \|\mathbf{x}\|_\infty .$$

3. $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are equivalent because for $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \leq n\, \|\mathbf{x}\|_\infty .$$

Remark 2.10.1. Based on the theorem, the convergence speed depends on the norms used.

Example 2.36. Consider $V = \mathbb{R}^2$. Let:

$$\mathbf{x}^{(k)} = \frac{1}{k}\begin{pmatrix}\cos\left(\frac{(2k-1)\pi}{4}\right)\\ \sin\left(\frac{(2k-1)\pi}{4}\right)\end{pmatrix} \in \mathbb{R}^2.$$

We can easily find that $\mathbf{x}^{(k)} \to \mathbf{0}$ with the norms $\|\cdot\|_1$ and $\|\cdot\|_2$. However,

$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_1 = \frac{\sqrt{2}}{k}, \qquad \|\mathbf{x}^{(k)} - \mathbf{0}\|_2 = \frac{1}{k}.$$

To achieve ε-precision:

$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_1 < \varepsilon \iff \frac{\sqrt{2}}{k_1} < \varepsilon \implies k_1 > \frac{\sqrt{2}}{\varepsilon},$$
$$\|\mathbf{x}^{(k)} - \mathbf{0}\|_2 < \varepsilon \iff \frac{1}{k_2} < \varepsilon \implies k_2 > \frac{1}{\varepsilon}.$$

The norm $\|\cdot\|_1$ takes about $\sqrt{2}$ times as many iterations as the norm $\|\cdot\|_2$ to check convergence using ε as the threshold.

Remark 2.10.2. In the case of infinite-dimensional vector spaces, not all norms are equivalent.

Example 2.37. Consider $V = \ell_1$. Let:

$$\mathbf{x}^{(k)} = \begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\end{pmatrix} \begin{matrix}\\ \\ k\text{-terms}\\ \\ \\ \end{matrix} = \sum_{i=1}^{k} \mathbf{e}_i \in \ell_1.$$

Consider the norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$. We find that:

$$\|\mathbf{x}^{(k)}\|_\infty = 1, \qquad \|\mathbf{x}^{(k)}\|_1 = k, \qquad \lim_{k\to\infty} \frac{\|\mathbf{x}^{(k)}\|_1}{\|\mathbf{x}^{(k)}\|_\infty} = \lim_{k\to\infty} k = \infty.$$

Therefore, the two norms are not equivalent.

Read Appendix A (Clustering, K-means, K-medians) to see the case study for Chapter 2.

Chapter 3

Inner products, Hilbert Spaces

3.1 Inner product

How do we describe whether two vectors are correlated? We cannot use norms to describe this because they are scaling-sensitive. As such, we define the inner product to quantify the relationship between two vectors.

Definition 3.1. Let V be a vector space over $\mathbb{R}$. An **inner product** over $\mathbb{R}$ is a binary operator $\langle \cdot, \cdot \rangle : (V, V) \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$:

1. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}$.
2. $\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle$ for all $\alpha, \beta \in \mathbb{R}$.
3. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$.

Remark 3.1.1. Property (2) and (3) are equivalent to the following: for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle \mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle + \beta \langle \mathbf{x}, \mathbf{z} \rangle .$$

Remark 3.1.2. Inner products can also be defined over $\mathbb{C}$, but property (3) would change to:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}.$$

Example 3.1. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the **dot product** for $\mathbb{R}^n$ defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i = \mathbf{x}^T \mathbf{y}$$

is the standard inner product in $\mathbb{R}^n$.

Example 3.2. For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the "weighted" inner product defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}$$

is an inner product on $\mathbb{R}^n$. The standard inner product is a special case of this inner product where $\mathbf{A} = \mathbf{I}$. However, if $\mathbf{A}$ is a symmetric positive semi-definite matrix, then the weighted inner product is not a true inner product. Instead, it is a **semi-inner-product**.

Proof.

1. For $\mathbf{x} \in \mathbb{R}^n$:
$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0.$$
Moreover:
$$\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = 0 \iff \mathbf{x}^T \mathbf{A} \mathbf{x} = 0 \iff \mathbf{x} = \mathbf{0}.$$
2. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$:
$$\langle \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} \rangle_{\mathbf{A}} = (\alpha\mathbf{x} + \beta\mathbf{y})^T \mathbf{A} \mathbf{z} = \alpha \mathbf{x}^T \mathbf{A} \mathbf{z} + \beta \mathbf{y}^T \mathbf{A} \mathbf{z} = \alpha \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{A}} + \beta \langle \mathbf{y}, \mathbf{z} \rangle_{\mathbf{A}} .$$
3. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:
$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y} = (\mathbf{x}^T \mathbf{A} \mathbf{y})^T = \mathbf{y}^T \mathbf{A}^T \mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}} .$$

□

Example 3.3. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, the inner product defined by:

$$\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij} = \mathrm{Tr}(\mathbf{A}^T\mathbf{B}) = \mathrm{Tr}(\mathbf{B}^T\mathbf{A}) = \mathrm{Tr}(\mathbf{A}\mathbf{B}^T) = \mathrm{Tr}(\mathbf{B}\mathbf{A}^T)$$

is the standard inner product in $\mathbb{R}^{m\times n}$.

Example 3.4. For $\mathbf{x}, \mathbf{y} \in \ell_2$, the inner product defined by:

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^{\infty} x_i y_i$$

is the standard inner product in ℓ_2.

Example 3.5. For $f, g \in \mathcal{C}[a, b]$, the inner product defined by:

$$\langle f, g\rangle = \int_a^b f(t)g(t)\,dt$$

is the standard inner product in $\mathcal{C}[a, b]$.

3.2 Properties of inner products

Let V be a vector space over $\mathbb{R}$ with an inner product $\langle\cdot,\cdot\rangle$.
Using the definition of inner products, we can discuss some properties of inner products.

Theorem 3.2. For any $\mathbf{x} \in V$, we have:

$$\langle \mathbf{x}, \mathbf{0}\rangle = \langle \mathbf{0}, \mathbf{x}\rangle = 0.$$

Proof.

$$\langle \mathbf{x}, \mathbf{0}\rangle = \langle \mathbf{x}, \mathbf{x} - \mathbf{x}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle - \langle \mathbf{x}, \mathbf{x}\rangle = 0.$$

□

In the previous chapter, we proved the Cauchy-Schwarz Inequality in $\mathbb{R}^n$ when verifying whether the 2-norm in $\mathbb{R}^n$ is a norm. We can generalize this inequality to all inner products.

Theorem 3.3. (**Cauchy-Schwarz Inequality**) For all $\mathbf{x}, \mathbf{y} \in V$:

$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 \le \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

Equality holds if and only if $\mathbf{y} = \alpha\mathbf{x}$ for some $\alpha \in \mathbb{R}$.

Proof.
For $\mathbf{x} \in V$, if $\mathbf{y} = \mathbf{0}$, then:

$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 = |\langle \mathbf{x}, \mathbf{0}\rangle|^2 = 0 = \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

If $\mathbf{y} \neq \mathbf{0}$, then for any $\lambda \in \mathbb{R}$:

$$0 \le \langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + \lambda \langle \mathbf{x}, \mathbf{y}\rangle + \lambda \langle \mathbf{y}, \mathbf{x}\rangle + \lambda^2 \langle \mathbf{y}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + \lambda(2\langle \mathbf{x}, \mathbf{y}\rangle) + \lambda^2 \langle \mathbf{y}, \mathbf{y}\rangle .$$

Since $\langle \mathbf{y}, \mathbf{y}\rangle > 0$, $\langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle$ is a quadratic function with at most one real root. Using the discriminant, we have:

$$\Delta = (2\langle \mathbf{x}, \mathbf{y}\rangle)^2 - 4\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle \le 0,$$
$$|\langle \mathbf{x}, \mathbf{y}\rangle|^2 \le \langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle .$$

For the equality case, notice that $\Delta = 0$ if and only if:

$$\langle \mathbf{x} + \lambda\mathbf{y}, \mathbf{x} + \lambda\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + 2\lambda\langle \mathbf{x}, \mathbf{y}\rangle + \lambda^2\langle \mathbf{y}, \mathbf{y}\rangle = \left(\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} + \lambda\sqrt{\langle \mathbf{y}, \mathbf{y}\rangle}\right)^2 - 2\lambda\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle \langle \mathbf{y}, \mathbf{y}\rangle} + 2\lambda\langle \mathbf{x}, \mathbf{y}\rangle = 0.$$

Solving this equation gives $\lambda = -\sqrt{\frac{\langle \mathbf{x}, \mathbf{x}\rangle}{\langle \mathbf{y}, \mathbf{y}\rangle}}$, and we find $\mathbf{y} = \sqrt{\frac{\langle \mathbf{y}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}}\mathbf{x}$ by definition.
Substituting $\mathbf{y} = \alpha\mathbf{x}$ for any $\alpha \in \mathbb{R}$ shows that $\Delta = 0$ if and only if $\mathbf{y} = \alpha\mathbf{x}$. □

With the Cauchy-Schwarz Inequality, we can construct a norm using an inner product.

Theorem 3.4. The **norm induced by the inner product** is a norm defined by, for any $\mathbf{x} \in V$:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

Proof.

1. For any $\mathbf{x} \in V$:
$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \geq 0.$$
Moreover:
$$\|\mathbf{x}\| = 0 \iff \langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}.$$

2. For any $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$:
$$\|\alpha \mathbf{x}\| = \sqrt{\langle \alpha \mathbf{x}, \alpha \mathbf{x} \rangle} = \sqrt{\alpha^2} \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |\alpha| \, \|\mathbf{x}\| .$$

3. For any $\mathbf{x}, \mathbf{y} \in V$:
$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 |\langle \mathbf{x}, \mathbf{y} \rangle| && (c \leq |c|) \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2 \|\mathbf{x}\| \, \|\mathbf{y}\| = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2, && \text{(Cauchy-Schwarz Inequality)} \\ \|\mathbf{x} + \mathbf{y}\| &\leq \|\mathbf{x}\| + \|\mathbf{y}\| . \end{aligned}$$

□

Remark 3.4.1. Inner product spaces are a subset of normed vector spaces. Not all normed vector spaces can define an inner product.

Remark 3.4.2. The Cauchy-Schwarz Inequality can be rewritten as, for $\mathbf{x}, \mathbf{y} \in V$:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \, \|\mathbf{y}\| .$$

Example 3.6. Consider $V = \mathbb{R}^n$ with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The induced norm is:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2} = \|\mathbf{x}\|_2 , \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

Remark 3.4.3. For $V = \mathbb{R}^n$, among all p-norms, only the 2-norm can be induced by an inner product.

Example 3.7. For a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, consider $V = \mathbb{R}^n$ with the weighted inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The induced norm is:

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j}.$$

If $\mathbf{A}$ is symmetric positive semi-definite, then the induced norm is not a true norm. Instead, it is a semi-norm (discussed in Appendix C).

Example 3.8. Consider $V = \mathbb{R}^{m\times n}$ with the standard inner product:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}, \qquad \text{for } \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}.$$

The induced norm is:

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \|\mathbf{A}\|_F = \|\mathbf{A}\|_{2,\text{vec}}, \qquad \text{for } \mathbf{A} \in \mathbb{R}^{m\times n}.$$

Example 3.9. Consider $V = \ell_2$ with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \ell_2.$$

The induced norm is:

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{\infty} x_i^2} = \|\mathbf{x}\|_2, \qquad \text{for } \mathbf{x} \in \ell_2.$$

Example 3.10. Consider $V = \mathcal{C}[a, b]$ with the standard inner product:

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt, \qquad \text{for } f, g \in \mathcal{C}[a, b].$$

The induced norm is:

$$\|f\| = \sqrt{\int_a^b (f(t))^2\, dt} = \|f\|_2, \qquad \text{for } f \in \mathcal{C}[a, b].$$

What conditions are required to define an inner product based on the normed vector space?

Theorem 3.5. Let $(V, \|\cdot\|)$ be a normed vector space over $\mathbb{R}$. The norm $\|\cdot\|$ is induced by an inner product if and only if the parallelogram law holds. This means that for all $\mathbf{x}, \mathbf{y} \in V$,

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

Proof.

$\Longrightarrow$ Suppose that the norm is induced by an inner product $\langle\cdot,\cdot\rangle$. This means for any $\mathbf{x}, \mathbf{y} \in V$:

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \langle \mathbf{x}+\mathbf{y}, \mathbf{x}+\mathbf{y}\rangle = \|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 ,\\ \|\mathbf{x}-\mathbf{y}\|^2 &= \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y}\rangle = \|\mathbf{x}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle + \|\mathbf{y}\|^2 .\end{aligned}$$

Therefore,

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

$\Longleftarrow$ Suppose that the parallelogram law holds. We may define a binary operator as:

$$\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2), \qquad \text{for } \mathbf{x}, \mathbf{y} \in V.$$

We check whether this is an inner product.

1. For any $\mathbf{x} \in V$,
$$\langle \mathbf{x}, \mathbf{x}\rangle = \frac{1}{4}(\|2\mathbf{x}\|^2 + \|\mathbf{0}\|^2) = \|\mathbf{x}\|^2 \geq 0.$$
Moreover,
$$\langle \mathbf{x}, \mathbf{x}\rangle = \|\mathbf{x}\|^2 = 0 \Longleftrightarrow \mathbf{x} = \mathbf{0}.$$

2. Since proving homogeneity extending to $\mathbb{R}$ is out of scope, we will only prove additivity here. For any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, by the parallelogram law,
$$\begin{aligned}\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 ,\\ &= 2\|\mathbf{y}\|^2 + 2\|\mathbf{x}+\mathbf{z}\|^2 - \|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 ,\\ \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 &= 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}-\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 ,\\ &= 2\|\mathbf{y}\|^2 + 2\|\mathbf{x}-\mathbf{z}\|^2 - \|-\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 .\end{aligned}$$
Combining the formulas, we have:
$$\begin{aligned}\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}+\mathbf{z}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 ,\\ \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}+\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 ,\\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|-\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2 - \frac{1}{2}\|\mathbf{x}-\mathbf{y}-\mathbf{z}\|^2 .\end{aligned}$$
Therefore,
$$\langle \mathbf{x}+\mathbf{y}, \mathbf{z}\rangle = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{x}+\mathbf{y}-\mathbf{z}\|^2) = \frac{1}{4}(\|\mathbf{x}+\mathbf{z}\|^2 - \|\mathbf{x}-\mathbf{z}\|^2 + \|\mathbf{y}+\mathbf{z}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2) = \langle \mathbf{x}, \mathbf{z}\rangle + \langle \mathbf{y}, \mathbf{z}\rangle .$$

3. For any $\mathbf{x}, \mathbf{y} \in V$,
$$\langle \mathbf{y}, \mathbf{x}\rangle = \frac{1}{4}(\|\mathbf{y}+\mathbf{x}\|^2 - \|\mathbf{y}-\mathbf{x}\|^2) = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2) = \langle \mathbf{x}, \mathbf{y}\rangle .$$

Therefore, since additionally $\sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} = \|\mathbf{x}\|$, the binary operator is an inner product that induces the norm.

□

Similar to normed vector spaces, we can define the completeness of a vector space with an inner product.

Definition 3.6. A complete inner product space is called a **Hilbert space**.

Example 3.11. $\mathbb{R}^n$ with the standard inner product or the weighted inner product for any symmetric positive definite $\mathbf{A} \in \mathbb{R}^{n\times n}$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}, \qquad \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^T \mathbf{A} \mathbf{y}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

are Hilbert spaces.

Example 3.12. $\mathbb{R}^{m\times n}$ with the standard inner product:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{Tr}(\mathbf{A}^T \mathbf{B}), \qquad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n},$$

is a Hilbert space.

Example 3.13. ℓ_2 with the standard inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \mathbf{x}, \mathbf{y} \in \ell_2,$$

is a Hilbert space.

Example 3.14. $\mathcal{C}[a, b]$ with the standard inner product:

$$\langle f, g \rangle = \int_a^b f(t) g(t)\, dt, \qquad f, g \in \mathcal{C}[a, b],$$

is not a Hilbert space. Its completion is $L^2(a, b)$.

3.3 Orthogonality

By the Cauchy-Schwarz Inequality, for all non-zero $\mathbf{x}, \mathbf{y} \in V$:

$$-\|\mathbf{x}\|\|\mathbf{y}\| \leq \langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|\|\mathbf{y}\| \iff -1 \leq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} \leq 1.$$

Considering when the Cauchy-Schwarz Inequality achieves equality:

1. If $\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} = 1$, then $\mathbf{y} = \alpha\mathbf{x}$ with $\alpha > 0$. (If $\alpha \leq 0$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{x} \rangle = \alpha \|\mathbf{x}\|^2 \leq 0$.)

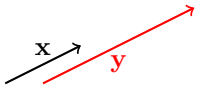


Figure 3.1: $\mathbf{y} = 2\mathbf{x}$

2. If $\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} = -1$, then $\mathbf{y} = \alpha\mathbf{x}$ with $\alpha < 0$. (If $\alpha \geq 0$, then $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{x} \rangle = \alpha \|\mathbf{x}\|^2 \geq 0$.)

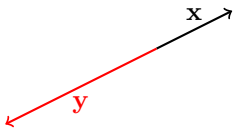


Figure 3.2: $\mathbf{y} = -2\mathbf{x}$

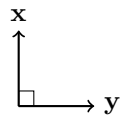
Definition 3.7. The **angle** between nonzero $\mathbf{x}, \mathbf{y} \in V$ is defined by:

$$\angle(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right).$$

With angles defined, we can define orthogonality.

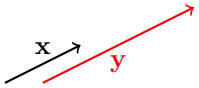
Definition 3.8. For $\mathbf{x}, \mathbf{y} \in V$:

1. If $\left|\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right| = 1$, then $\mathbf{x}$ and $\mathbf{y}$ are the **most correlated**.
2. If $\left|\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|}\right| = 0$ ($\langle \mathbf{x}, \mathbf{y} \rangle = 0$), then $\mathbf{x}$ and $\mathbf{y}$ are the **least correlated**. We say $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal**.

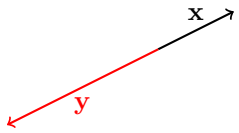


$\angle(\mathbf{x}, \mathbf{y}) = \frac{\pi}{2}$

(a) The least correlated



$\angle(\mathbf{x}, \mathbf{y}) = 0$



$\angle(\mathbf{x}, \mathbf{y}) = \pi$

(b) The most correlated

Based on orthogonality, we have the following theorem.

Theorem 3.9. (**Pythagorean Theorem**) For $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if:

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

Proof.
If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Therefore:

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

If $\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$, then:

$$0 = \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle - \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle = 2\langle \mathbf{x}, \mathbf{y} \rangle.$$

Therefore, $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, and thus $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. □

Read Appendix B (Kernel K-means/Kernel Trick) and C (Metric Learning) to see the case studies for Chapter 3.

Chapter 4

Linear Functions and Differentiation

In Chapter 2, we observed that the norm does not preserve vector addition but does preserve scalar multiplication. In Chapter 3, by fixing one of the vectors, we demonstrated linearity. Here, we aim to investigate the behavior of linear functions in vector spaces.

4.1 Linear Functions

Definition 4.1. Let V be a vector space over $\mathbb{R}$. A function $f : V \to \mathbb{R}$ is **linear** if, for all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$,

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

Example 4.1. The mean of a vector (not to be confused with mean vectors): for all $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a linear function.

Proof.
For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i + \beta y_i) = \frac{\alpha}{n}\sum_{i=1}^{n} x_i + \frac{\beta}{n}\sum_{i=1}^{n} y_i = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

□

Example 4.2. The maximum entry of a vector: for all $\mathbf{x} \in \mathbb{R}^n$,

$$f(\mathbf{x}) = \max_{1 \le i \le n} x_i$$

is not a linear function.

Proof.
Assume that $\mathbf{x} = \begin{pmatrix}1\\0\end{pmatrix}$, $\mathbf{y} = \begin{pmatrix}0\\1\end{pmatrix}$, $\alpha = 1$, $\beta = 1$. We have:

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = f\left(\begin{pmatrix}1\\1\end{pmatrix}\right) = 1, \qquad \alpha f(\mathbf{x}) + \beta f(\mathbf{y}) = f\left(\begin{pmatrix}1\\0\end{pmatrix}\right) + f\left(\begin{pmatrix}0\\1\end{pmatrix}\right) = 2 \neq f(\alpha\mathbf{x} + \beta\mathbf{y}),$$

which violates the definition of a linear function. □

Example 4.3. Let V be a normed vector space with an inner product $\langle \cdot, \cdot \rangle$ and let $\mathbf{a} \in V$. The function $f : V \to \mathbb{R}$ defined by:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in V$$

is a linear function.

Proof.
For all $\mathbf{x}, \mathbf{y} \in V$ and $\alpha, \beta \in \mathbb{R}$,

$$f(\alpha\mathbf{x} + \beta\mathbf{y}) = \langle \mathbf{a}, \alpha\mathbf{x} + \beta\mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}).$$

□

Example 4.4. A functional $F : \mathcal{C}[-1,1] \to \mathbb{R}$ defined by:

$$F(f) = f(0) \qquad \text{for } f \in \mathcal{C}[-1,1]$$

is a linear function.

Proof.
For all $f, g \in \mathcal{C}[-1,1]$ and $\alpha, \beta \in \mathbb{R}$,

$$F(\alpha f + \beta g) = (\alpha f + \beta g)(0) = \alpha f(0) + \beta g(0) = \alpha F(f) + \beta F(g).$$

□

Example 4.5. A functional $F : \mathcal{C}[a,b] \to \mathbb{R}$ defined by:

$$F(f) = \int_a^b f(t)\,dt \qquad \text{for } f \in \mathcal{C}[a,b]$$

is a linear function.

Proof.
For all $f, g \in \mathcal{C}[a,b]$ and $\alpha, \beta \in \mathbb{R}$,

$$F(\alpha f + \beta g) = \int_a^b (\alpha f + \beta g)(t)\,dt = \int_a^b (\alpha f(t) + \beta g(t))\,dt = \alpha \int_a^b f(t)\,dt + \beta \int_a^b g(t)\,dt = \alpha F(f) + \beta F(g).$$

□

Theorem 4.2. For any vector space V, any norm function $\|\cdot\|$ on V is not a linear function.

Proof.
By the absolute homogeneity property of the norm, for all $\mathbf{x} \in V$,

$$\|-\mathbf{x}\| = \|\mathbf{x}\|.$$

Assume that the norm function is linear. Then:

$$\|-\mathbf{x}\| = \|-\mathbf{x} + 0\mathbf{x}\| = -\|\mathbf{x}\| + 0\|\mathbf{x}\| = -\|\mathbf{x}\|,$$

which results in a contradiction. Therefore, any norm function is not a linear function. □

The following properties can be easily derived from the definition.

Theorem 4.3. A linear function f has the following properties:

1. Homogeneity: For all $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$,
$$f(\alpha\mathbf{x}) = \alpha f(\mathbf{x}).$$
2. Additivity: For any $\mathbf{x}, \mathbf{y} \in V$,
$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}).$$

Proof.

1. Setting $\beta = 0$, for any $\mathbf{x} \in V$ and $\alpha \in \mathbb{R}$,
$$f(\alpha\mathbf{x}) = f(\alpha\mathbf{x} + 0\mathbf{y}) = \alpha f(\mathbf{x}) + 0 f(\mathbf{y}) = \alpha f(\mathbf{x}), \qquad \text{for } \mathbf{y} \in V.$$
2. Setting $\alpha = \beta = 1$, for any $\mathbf{x}, \mathbf{y} \in V$,
$$f(\mathbf{x} + \mathbf{y}) = f(1\mathbf{x} + 1\mathbf{y}) = 1f(\mathbf{x}) + 1f(\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}).$$

□

Remark 4.3.1. By induction, for $\mathbf{x}_1, \cdots, \mathbf{x}_k \in V$ and $\alpha_1, \cdots, \alpha_k \in \mathbb{R}$,

$$f(\alpha_1\mathbf{x}_1 + \cdots + \alpha_k\mathbf{x}_k) = \alpha_1 f(\mathbf{x}_1) + \cdots + \alpha_k f(\mathbf{x}_k).$$

Let H be a Hilbert space. From Example 4.3, we have shown that for all $\mathbf{a} \in H$, the function:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in H$$

is a linear function. Is it true that for all linear functions $f : H \to \mathbb{R}$, there exists a fixed vector $\mathbf{a} \in H$ such that the function can be written in inner product form? The answer is yes!

Theorem 4.4. (**Riesz Representation Theorem**) Let H be a Hilbert space with an inner product $\langle \cdot, \cdot \rangle$. The function $f : H \to \mathbb{R}$ is linear and bounded if and only if:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle, \qquad \text{for } \mathbf{x} \in H$$

for some unique $\mathbf{a} \in H$, called the **Riesz representation** of f.

Proof (When $H = \mathbb{R}^n$).
For all $\mathbf{x} \in \mathbb{R}^n$, we have:

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n.$$

Therefore, we have:

$$f(\mathbf{x}) = f(x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n) = x_1 f(\mathbf{e}_1) + \cdots + x_n f(\mathbf{e}_n) = \left\langle \begin{pmatrix} f(\mathbf{e}_1) \\ \vdots \\ f(\mathbf{e}_n) \end{pmatrix}, \mathbf{x} \right\rangle \Longrightarrow \mathbf{a} = \begin{pmatrix} f(\mathbf{e}_1) \\ \vdots \\ f(\mathbf{e}_n) \end{pmatrix}.$$

Suppose that $\mathbf{a}$ is not unique. This means there exist $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle = \langle \mathbf{b}, \mathbf{x} \rangle \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

If we choose $\mathbf{x} = \mathbf{e}_i$ for $i = 1, \cdots, n$,

$$\begin{aligned} f(\mathbf{e}_i) = \langle \mathbf{a}, \mathbf{e}_i \rangle = \langle \mathbf{b}, \mathbf{e}_i \rangle \Longrightarrow a_i = b_i, \qquad & \text{for } i = 1, \cdots, n, \\ \Longrightarrow \mathbf{a} = \mathbf{b}. & \end{aligned}$$

This results in a contradiction. Therefore, $\mathbf{a}$ is unique. □

Remark 4.4.1. The "boundedness" implied in this theorem does not mean the codomain of the function is a bounded set. Given a function $f : X \to Y$ that maps between two normed vector spaces, if X has a norm $\|\cdot\|_X$ and Y has a norm $\|\cdot\|_Y$, then there exists some $M > 0$ such that:

$$\|f(\mathbf{x})\|_Y \leq M \|\mathbf{x}\|_X, \qquad \text{for } \mathbf{x} \in X.$$

In this theorem, if H induces a norm $\|\cdot\|_H$, then there exists some $M > 0$ such that:

$$|f(\mathbf{x})| \leq M \|\mathbf{x}\|_H, \qquad \text{for } \mathbf{x} \in H.$$

Remark 4.4.2. The smallest value M is often called the **operator norm**. Recall Example 2.18. It can be generalized to:

$$M = \|f\|_{op} = \sup\{\|f(\mathbf{x})\|_Y : \mathbf{x} \in X \text{ with } \|\mathbf{x}\|_X \leq 1\},$$

which depends on the choice of norms.

Remark 4.4.3. Any linear function defined on a finite-dimensional normed vector space is always bounded.

Example 4.6. The mean of a vector $\mathbf{x} \in \mathbb{R}^n$ can be represented by:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i = \left\langle \frac{1}{n} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{x} \right\rangle.$$

Example 4.7. The trace of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ can be represented by:

$$\mathrm{Tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = \langle \mathbf{I}, \mathbf{A} \rangle.$$

Remark 4.4.4. In infinite-dimensional Hilbert spaces, there exist linear but unbounded functions.

Example 4.8. We consider $L^2(-1,1)$, which is the completion of $\mathcal{C}[-1,1]$ under:

$$\langle f,g\rangle = \int_{-1}^{1} f(t)g(t)\,dt, \qquad \|f\|_2 = \sqrt{\langle f,f\rangle}.$$

Let $F(f) = f(0)$ for all $f \in L^2(-1,1)$. Consider that:

$$f(t) = \begin{cases} 1, & t \neq 0, \\ +\infty, & t = 0 \end{cases}, \qquad \text{for } t \in (-1,1).$$

Therefore, $F(f) = f(0) = +\infty$. There is, in fact, no inner product representation for $F(f) = f(0)$.

Example 4.9. We consider $G : L^2(-1,1) \to \mathbb{R}$ defined by:

$$G(f) = \int_{-1}^{1} f(t)\,dt, \qquad \text{for } f \in L^2(-1,1).$$

For any $f, g \in L^2(-1,1)$ and $\alpha, \beta \in \mathbb{R}$,

$$G(\alpha f + \beta g) = \int_{-1}^{1} (\alpha f + \beta g)(t)\,dt = \alpha \int_{-1}^{1} f(t)\,dt + \beta \int_{-1}^{1} g(t)\,dt = \alpha G(f) + \beta G(g).$$

Therefore, we can find that G is linear. We can also see that for any $f \in L^2(-1,1)$,

$$G(f) = \int_{-1}^{1} f(t)\,dt = \int_{-1}^{1} 1 \cdot f(t)\,dt = \langle 1, f\rangle \leq \|f\|_2 \sqrt{\int_{-1}^{1} 1^2\,dt} = \sqrt{2}\,\|f\|_2 .$$

Therefore, G is also bounded. In fact, we have found that:

$$G(f) = \langle 1, f\rangle .$$

Example 4.10. We consider ℓ_2, which is the completion of the space of all finite sequences under:

$$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{\infty} x_i y_i, \qquad \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle}.$$

Let $H : \ell_2 \to \mathbb{R}$ be defined by:

$$H(\mathbf{x}) = x_1, \qquad \text{for } \mathbf{x} \in \ell_2.$$

For any $\mathbf{x}, \mathbf{y} \in \ell_2$ and $\alpha, \beta \in \mathbb{R}$,

$$H(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha x_1 + \beta y_1 = \alpha H(\mathbf{x}) + \beta H(\mathbf{y}).$$

Therefore, we can find that H is linear. We can also see that for any $\mathbf{x} \in \ell_2$,

$$H(\mathbf{x}) = x_1 \leq \sqrt{\sum_{i=1}^{\infty} x_i^2} = \|\mathbf{x}\|_2 .$$

Therefore, H is also bounded. In fact, we have found that:

$$H(\mathbf{x}) = \left\langle \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \mathbf{x} \right\rangle .$$

Additionally, we have a function that is similar to a linear function but shifted in the output space.

Definition 4.5. Let V be a vector space over $\mathbb{R}$. A function $f : V \to \mathbb{R}$ is **affine** if it can be written as:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad \text{for } \mathbf{x} \in V,$$

where $g : V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$.

We can also derive the following properties from the definition.

Theorem 4.6. An affine function f has the following properties:

1. For any $\alpha, \beta \in \mathbb{R}$ such that $\alpha + \beta = 1$,

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

2. If H is a Hilbert space and f is bounded, then f is affine if and only if:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b,$$

for some $\mathbf{a} \in H$ and $b \in \mathbb{R}$.

Proof.

1. There exists a linear function g such that:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad \text{for } \mathbf{x} \in V,$$

where $b \in \mathbb{R}$. If $\alpha + \beta = 1$,

$$\begin{aligned} f(\alpha\mathbf{x} + \beta\mathbf{y}) &= g(\alpha\mathbf{x} + \beta\mathbf{y}) + b, \\ &= \alpha g(\mathbf{x}) + \beta g(\mathbf{y}) + (\alpha + \beta)b, && (\alpha + \beta = 1) \\ &= \alpha(g(\mathbf{x}) + b) + \beta(g(\mathbf{y}) + b), \\ &= \alpha f(\mathbf{x}) + \beta f(\mathbf{y}). \end{aligned}$$

2. $\implies$ If f is affine, then there exists a linear function g such that:

$$f(\mathbf{x}) = g(\mathbf{x}) + b, \qquad g(\mathbf{x}) = f(\mathbf{x}) - b, \qquad \text{for } \mathbf{x} \in V,$$

where $b \in \mathbb{R}$. Since f is bounded, there exists some $M_f > 0$ such that:

$$|f(\mathbf{x})| \leq M_f \|\mathbf{x}\|_H, \qquad b = |f(\mathbf{0})| \leq M_f \|\mathbf{0}\|_H = 0.$$

By the Riesz Representation Theorem, there exists some $\mathbf{a} \in H$ such that:

$$g(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle, \qquad \text{for } \mathbf{x} \in H.$$

Therefore,

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b, \qquad \text{for } \mathbf{x} \in H.$$

$\impliedby$ If there exists some $\mathbf{a} \in H$ and $b \in \mathbb{R}$ such that:

$$f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + b, \qquad \text{for } \mathbf{x} \in H,$$

then, by Example 4.3, $\langle \mathbf{a}, \mathbf{x} \rangle$ is a linear function. Therefore, f is affine.

□

4.2 Hyperplane

Definition 4.7. Let V be a vector space over $\mathbb{R}$. A subset $U \subset V$ is a **(linear) subspace** of V if, for any $\mathbf{u}, \mathbf{v} \in U$ and $\alpha, \beta \in \mathbb{R}$,

$$\alpha\mathbf{u} + \beta\mathbf{v} \in U.$$

Definition 4.8. Let V be a vector space over $\mathbb{R}$. A set W is an **affine subspace** if:

$$W = \mathbf{a} + U = \{\mathbf{a} + \mathbf{u} : \mathbf{u} \in U\},$$

where $\mathbf{a} \in V$ and U is a linear subspace of V.

Let H be a Hilbert space and $\mathbf{a} \in H$. We denote:

$$S_{\mathbf{a},b} = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\} \subset H.$$

Consider $b = 0$. For all $\mathbf{x}, \mathbf{y} \in S_{\mathbf{a},0}$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle \mathbf{a}, \alpha\mathbf{x} + \beta\mathbf{y} \rangle = \alpha \langle \mathbf{a}, \mathbf{x} \rangle + \beta \langle \mathbf{a}, \mathbf{y} \rangle = 0.$$

This means that $\alpha\mathbf{x} + \beta\mathbf{y} \in S_{\mathbf{a},0}$. Therefore, $S_{\mathbf{a},0}$ is a linear subspace of H.
For a general b, let $\mathbf{x}_0 \in S_{\mathbf{a},b}$. Then we have:

$$\langle \mathbf{a}, \mathbf{x}_0 \rangle = b.$$

For all $\mathbf{x} \in S_{\mathbf{a},b}$,

$$\begin{aligned} \langle \mathbf{a}, \mathbf{x} - \mathbf{x}_0 \rangle = \langle \mathbf{a}, \mathbf{x} \rangle - \langle \mathbf{a}, \mathbf{x}_0 \rangle = b - b = 0 &\implies \mathbf{x} - \mathbf{x}_0 \in S_{\mathbf{a},0}, \\ &\implies \mathbf{x} \in \mathbf{x}_0 + S_{\mathbf{a},0}, \\ &\implies S_{\mathbf{a},b} \subset \mathbf{x}_0 + S_{\mathbf{a},0}. \end{aligned}$$

For all $\mathbf{x} \in S_{\mathbf{a},0}$,

$$\begin{aligned} \langle \mathbf{a}, \mathbf{x} + \mathbf{x}_0 \rangle = \langle \mathbf{a}, \mathbf{x} \rangle + \langle \mathbf{a}, \mathbf{x}_0 \rangle = 0 + b = b &\implies \mathbf{x} + \mathbf{x}_0 \in S_{\mathbf{a},b}, \\ &\implies S_{\mathbf{a},0} + \mathbf{x}_0 \subset S_{\mathbf{a},b}. \end{aligned}$$

Therefore, we have $S_{\mathbf{a},b} = \mathbf{x}_0 + S_{\mathbf{a},0}$, and thus $S_{\mathbf{a},b}$ is an affine subspace. We can define such a set as a hyperplane.

Definition 4.9. A **hyperplane** S in the Hilbert space H is defined by:

$$S = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\} \subset H,$$

where $\mathbf{a} \in H$ and $b \in \mathbb{R}$.

Remark 4.9.1. If H has a dimension of n, then the hyperplane S has a dimension of $n - 1$ or a codimension of 1.

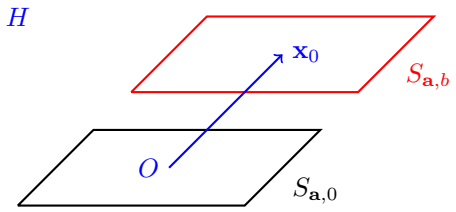


Figure 4.1: Hyperplanes

For any vectors that do not lie on the hyperplane, we can project those vectors onto the hyperplane.

Definition 4.10. Consider a hyperplane $S \subset H$. For any $\mathbf{y} \in H$, the vector on S that is closest to $\mathbf{y}$ is called the **projection** of $\mathbf{y}$ onto S, denoted by $P_S\mathbf{y}$. Equivalently:

$$P_S\mathbf{y} = \operatorname*{argmin}_{\mathbf{x}\in S} \|\mathbf{y} - \mathbf{x}\| .$$

How do we find the explicit form of such an equation?

Theorem 4.11. Given a hyperplane $S \subset H$, the vector $\mathbf{z} \in H$ is a solution of:

$$\min_{\mathbf{x}\in S} \|\mathbf{y} - \mathbf{x}\|$$

if and only if $\mathbf{z} \in S$ and $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$.

Proof.

$\implies$ Assume that $\mathbf{z}$ is a solution of the minimization equation. Then $\mathbf{z} \in S$. For all $\mathbf{x} \in S$ and $t \in \mathbb{R}$,

$$\begin{aligned}\langle \mathbf{a}, \mathbf{z} + t(\mathbf{x} - \mathbf{z})\rangle &= \langle \mathbf{a}, \mathbf{z}\rangle + t(\langle \mathbf{a}, \mathbf{x}\rangle - \langle \mathbf{a}, \mathbf{z}\rangle),\\ &= b + t(b - b),\\ &= b.\end{aligned}$$

Therefore, $\mathbf{z} + t(\mathbf{x} - \mathbf{z}) \in S$. We have:

$$\begin{aligned}\|\mathbf{z} - \mathbf{y}\|^2 \le \|\mathbf{z} + t(\mathbf{x} - \mathbf{z}) - \mathbf{y}\|^2 &= \|\mathbf{z} - \mathbf{y}\|^2 + 2t\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle + t^2 \|\mathbf{x} - \mathbf{z}\|^2 ,\\ -t^2 \|\mathbf{x} - \mathbf{z}\|^2 &\le 2t\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle .\end{aligned}$$

Assume that $t > 0$. As $t \to 0^+$,

$$\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle \ge -\lim_{t\to 0^+} \frac{t}{2} \|\mathbf{x} - \mathbf{z}\|^2 = 0.$$

Assume that $t < 0$. As $t \to 0^-$,

$$\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle \le -\lim_{t\to 0^-} \frac{t}{2} \|\mathbf{x} - \mathbf{z}\|^2 = 0.$$

Therefore, $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$.

$\impliedby$ Assume that $\mathbf{z} \in S$ and $\langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z}\rangle = 0$ for all $\mathbf{x} \in S$. This means that:

$$\begin{aligned}\|\mathbf{x} - \mathbf{y}\|^2 &= \|(\mathbf{x} - \mathbf{z}) + (\mathbf{z} - \mathbf{y})\|^2 ,\\ &= \|\mathbf{x} - \mathbf{z}\|^2 + 2\langle \mathbf{x} - \mathbf{z}, \mathbf{z} - \mathbf{y}\rangle + \|\mathbf{z} - \mathbf{y}\|^2 ,\\ &= \|\mathbf{x} - \mathbf{z}\|^2 + \|\mathbf{z} - \mathbf{y}\|^2 ,\\ &\ge \|\mathbf{z} - \mathbf{y}\|^2 .\end{aligned}$$

Therefore, we can find that:

$$\mathbf{z} = \operatorname*{argmin}_{\mathbf{x}\in S} \|\mathbf{x} - \mathbf{y}\|^2 .$$

□

Using the last theorem, we can obtain the explicit form of the projection.

Theorem 4.12. Let H be a Hilbert space and S be the hyperplane defined by:

$$S = \{\mathbf{x} \in H : \langle \mathbf{a}, \mathbf{x} \rangle = b\},$$

where $\mathbf{a} \in H$ and $b \in \mathbb{R}$. Given $\mathbf{y} \in H$, we have a unique solution:

$$P_S \mathbf{y} = \operatorname*{argmin}_{\mathbf{x} \in S} \|\mathbf{y} - \mathbf{x}\| = \mathbf{y} - \left(\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \right) \mathbf{a}.$$

Proof.
Let $\mathbf{z} = \mathbf{y} - \left(\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \right) \mathbf{a}$.

$$\begin{aligned} \langle \mathbf{a}, \mathbf{z} \rangle &= \langle \mathbf{a}, \mathbf{y} \rangle - \frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \langle \mathbf{a}, \mathbf{a} \rangle, \\ &= \langle \mathbf{a}, \mathbf{y} \rangle - (\langle \mathbf{a}, \mathbf{y} \rangle - b), \\ &= b. \end{aligned}$$

Therefore, $\mathbf{z} \in S$. For any $\mathbf{x} \in S$,

$$\begin{aligned} \langle \mathbf{z} - \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle &= \left\langle -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} \mathbf{a}, \mathbf{x} - \mathbf{z} \right\rangle, \\ &= -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} (\langle \mathbf{a}, \mathbf{x} \rangle - \langle \mathbf{a}, \mathbf{z} \rangle), \\ &= -\frac{\langle \mathbf{a}, \mathbf{y} \rangle - b}{\|\mathbf{a}\|^2} (b - b), \\ &= 0. \end{aligned}$$

Therefore, by Theorem 4.11, $\mathbf{z}$ is a solution to:

$$\min_{\mathbf{x} \in S} \|\mathbf{y} - \mathbf{x}\|.$$

We prove that this solution is unique.
Suppose that it has two distinct solutions $\mathbf{z}_1$ and $\mathbf{z}_2$. Then it implies that $\mathbf{z}_1, \mathbf{z}_2 \in S$. By Theorem 4.11,

$$\langle \mathbf{z}_1 - \mathbf{y}, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0, \qquad \langle \mathbf{z}_2 - \mathbf{y}, \mathbf{z}_1 - \mathbf{z}_2 \rangle = \langle \mathbf{y} - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0.$$

Adding the two identities, we have:

$$\begin{aligned} \langle \mathbf{z}_1 - \mathbf{y}, \mathbf{z}_2 - \mathbf{z}_1 \rangle + \langle \mathbf{y} - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = \langle \mathbf{z}_1 - \mathbf{z}_2, \mathbf{z}_2 - \mathbf{z}_1 \rangle = 0 &\iff - \|\mathbf{z}_1 - \mathbf{z}_2\|^2 = 0, \\ &\iff \mathbf{z}_1 = \mathbf{z}_2. \end{aligned}$$

This results in a contradiction. Therefore, the solution $\mathbf{z}$ is unique. □

Appendix A

Clustering, K-means, K-medians

This case study assumes that you have already read Chapter 2.
Suppose we are given N vectors in $\mathbb{R}^n$:

$$\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^n.$$

We want to group them into K different clusters.

Remark A.0.1. $\mathbb{R}^n$ is used for simplicity. In fact, it can be replaced by any vector space.

Before performing clustering on any vector space, we must formulate the problem mathematically.

1. Representation: Starting with N vectors $\{\mathbf{x}_i\}_{i=1}^N$ and K clusters $\{G_j\}_{j=1}^K$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the vectors to be grouped,
   (b) $c_i \in \{1, \cdots, K\}$: the cluster to which $\mathbf{x}_i$ belongs,
   (c) $G_j = \{i : c_i = j\}$: the clusters, which are sets of indices representing the vectors in the group,
   (d) $\mathbf{z}_j \in \mathbb{R}^n$: the representative vector in G_j, not necessarily one of the vectors in $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$.
2. Evaluation: What problem do we want to solve? The vectors in each cluster should be close to each other.
   (a) The distance between the vectors in a cluster and the corresponding representative vector should be minimized. Therefore, we define an optimization function:

   $$d_j = \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

   Our objective for this cluster is to minimize this optimization function.
   (b) Altogether, we get the overall optimization function:

   $$d = \sum_{j=1}^K d_j.$$

   Then, we solve:

   $$\min_{\substack{G_1, \cdots, G_K \\ \mathbf{z}_1, \cdots, \mathbf{z}_K}} d \iff \min_{\substack{G_1, \cdots, G_K \\ \mathbf{z}_1, \cdots, \mathbf{z}_K}} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

3. Optimization: We now have two sets of unknowns $\{G_1, \cdots, G_K\}$ and $\{\mathbf{z}_1, \cdots, \mathbf{z}_K\}$. However, both influence each other. How do we tackle this issue? We use alternating minimization.

Step 0: Initialize $\mathbf{z}_1, \cdots, \mathbf{z}_K$.

Step 1: Fix $\mathbf{z}_1, \cdots, \mathbf{z}_K$ and solve the function with respect to $G_1, \cdots, G_K$:

$$\min_{G_1, \cdots, G_K} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

Step 2: Fix $G_1, \cdots, G_K$ and solve the function with respect to $\mathbf{z}_1, \cdots, \mathbf{z}_K$:

$$\min_{\mathbf{z}_1, \cdots, \mathbf{z}_K} \sum_{j=1}^K \sum_{i \in G_j} \|\mathbf{x}_i - \mathbf{z}_j\|^2.$$

Repeat Steps 1 and 2 until convergence is achieved.

How do we solve the functions in the alternating minimization algorithm? To obtain the clusters, we solve the following function:

$$\begin{aligned}\min_{G_1,\cdots,G_K}\sum_{j=1}^{K}\sum_{i\in G_j}\|\mathbf{x}_i-\mathbf{z}_j\|^2 &\iff \min_{c_1,\cdots,c_N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{z}_{c_i}\|^2\\ &\iff \min_{c_i\in\{1,\cdots,K\}}\|\mathbf{x}_i-\mathbf{z}_{c_i}\|^2\\ &\iff c_i=\operatorname*{argmin}_{j\in\{1,\cdots,K\}}\|\mathbf{x}_i-\mathbf{z}_j\|, \qquad \text{for } i=1,\cdots,N.\end{aligned}$$

In simpler terms, finding clusters that minimize the distance between the vectors and the representatives is the same as assigning each vector to the cluster that minimizes the distance to its representative.
Therefore, $\mathbf{x}_i$ is assigned to the cluster whose representative is closest to $\mathbf{x}_i$. The new G_j is then created by:

$$G_j=\{i:c_i=j\}.$$

To obtain the representative vectors, we solve the following function:

$$\min_{\mathbf{z}_1,\cdots,\mathbf{z}_K}\sum_{j=1}^{K}\sum_{i\in G_j}\|\mathbf{x}_i-\mathbf{z}_j\|^2\iff\min_{\mathbf{z}_j}\sum_{i\in G_j}\|\mathbf{x}_i-\mathbf{z}_j\|^2, \qquad \text{for } j=1,\cdots,K.$$

We only need to consider each cluster separately to find the corresponding representative vector.
At this point, we haven't defined which norms we use to construct the clusters. In $\mathbb{R}^n$, one of the most commonly used norms for clustering is the 2-norm. For $j=1,\cdots,K$:

$$\begin{aligned}\min_{\mathbf{z}_j}\sum_{i\in G_j}\|\mathbf{x}_i-\mathbf{z}_j\|_2^2 &\iff \min_{z_{j1},\cdots,z_{jn}}\sum_{i\in G_j}\sum_{k=1}^{n}(x_{ik}-z_{jk})^2\\ &\iff 2\sum_{i\in G_j}(z_{jk}-x_{ik})=0\iff z_{jk}=\frac{1}{|G_j|}\sum_{i\in G_j}x_{ik}, \qquad \text{for } k=1,\cdots,n,\\ &\iff \mathbf{z}_j=\frac{1}{|G_j|}\sum_{i\in G_j}\mathbf{x}_i.\end{aligned}$$

This is called the K-means algorithm.

Definition A.1. The **K-means algorithm** is a method that assigns N vectors into K clusters, where each vector belongs to the cluster with the nearest mean. The steps of the algorithm are as follows:

0: Initialize $\mathbf{z}_1,\cdots,\mathbf{z}_K$.

1: Fix $\mathbf{z}_1,\cdots,\mathbf{z}_K$. For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Euclidean distance:

$$c_i=\operatorname*{argmin}_{j\in\{1,\cdots,K\}}\|\mathbf{x}_i-\mathbf{z}_j\|_2^2, \qquad \text{for } i=1,\cdots,N, \qquad G_j=\{i:c_i=j\}, \qquad \text{for } j=1,\cdots,K.$$

2: For each cluster G_j, calculate the new $\mathbf{z}_j$ as the mean of vectors in G_j:

$$\mathbf{z}_j=\frac{1}{|G_j|}\sum_{i\in G_j}\mathbf{x}_i, \qquad \text{for } j=1,\cdots,K.$$

Repeat until convergence is achieved.

What happens if we switch from the 2-norm to the 1-norm? Derivations are omitted, but we find that it turns into the K-medians algorithm.

Definition A.2. The **K-medians algorithm** is a method that assigns N vectors into K clusters, where each vector belongs to the cluster with the nearest median. The steps of the algorithm are as follows:

0: Initialize $\mathbf{z}_1, \cdots, \mathbf{z}_K$.

1: Fix $\mathbf{z}_1, \cdots, \mathbf{z}_K$. For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Manhattan distance:

$$c_i = \underset{j \in \{1, \cdots, K\}}{\operatorname{argmin}} \|\mathbf{x}_i - \mathbf{z}_j\|_1^2, \qquad \text{for } i = 1, \cdots, N, \qquad G_j = \{i : c_i = j\}, \qquad \text{for } j = 1, \cdots, K.$$

2: For each cluster G_j, calculate the new $\mathbf{z}_j$ as the median of vectors in G_j:

$$\mathbf{z}_j = \text{median}\{\mathbf{x}_i : i \in G_j\}, \qquad \text{for } j = 1, \cdots, K.$$

Repeat until convergence is achieved.

Remark A.2.1. The K-means algorithm is more sensitive to outliers, while the K-medians algorithm is more robust to outliers.

Appendix B

Kernel K-means/Kernel Trick

This case study assumes that you have already read Chapter 3 and Appendix A.
In the regular K-means algorithm, we assume that the vectors can be separated linearly. However, it fails when the boundary is curved. How do we modify the K-means algorithm to deal with curved data? We can transform the data to a new domain and then apply the K-means algorithm in that transformed domain.

1. Representation: Starting with N vectors $\{\mathbf{x}_i\}_{i=1}^N$ and K clusters $\{G_j\}_{j=1}^K$, we define the following variables:
   (a) $\mathbf{x}_i \in \mathbb{R}^n$: the vectors to be grouped.
   (b) $c_i \in \{1, \cdots, K\}$: the cluster to which $\mathbf{x}_i$ belongs.
   (c) $G_j = \{i : c_i = j\}$: the clusters, which are sets of indices representing the vectors in the group.
   (d) H: the feature space that contains the transformed vectors.
   (e) $\phi : \mathbb{R}^n \to H$: the feature map that transforms the vectors.
   (f) $\mathbf{z}_j \in H$: the representative vector in G_j, not necessarily one of the vectors in $\{\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)\}$.
2. Evaluation: With the feature map, our objective becomes finding:
$$\min_{\substack{G_1,\cdots,G_K \\ \mathbf{z}_1,\cdots,\mathbf{z}_K}} \sum_{j=1}^{K} \sum_{i \in G_j} \|\phi(\mathbf{x}_i) - \mathbf{z}_j\|^2$$
   In this algorithm, we also need to find a good ϕ that can transform the data well. If we assume that the norm we use has an inner product, our goal is to ensure that:
   (a) If $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ are close ($\|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|_2$ is small), then $\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2$ is small.
   (b) If $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$ are far apart ($\|\mathbf{x}_{i_1} - \mathbf{x}_{i_2}\|_2$ is large), then $\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2$ is large.
3. Optimization: Since ϕ depends on the shape of $\mathbf{x}_1, \cdots, \mathbf{x}_N$, it is not easy to find a good ϕ and H explicitly.
   Therefore, we can utilize the **Kernel trick** – define ϕ and H implicitly.
   Moreover, if we only need to know $G_1, \cdots, G_K$, we can eliminate $\mathbf{z}_1, \cdots, \mathbf{z}_K$ in our algorithm.

The K-means algorithm can be modified as follows:

0: Initialize $G_1, \cdots, G_K$.

1: For each $\mathbf{x}_i$, assign it to the cluster whose representative is closest in Euclidean distance:
$$c_i = \operatorname*{argmin}_{j \in \{1,\cdots,K\}} \left\| \phi(\mathbf{x}_i) - \frac{1}{|G_j|} \sum_{k \in G_j} \phi(\mathbf{x}_k) \right\|_2^2, \qquad \text{for } i = 1, \cdots, N, \quad G_j = \{i : c_i = j\}, \qquad \text{for } j = 1, \cdots, K.$$
   Repeat until convergence is achieved.

We can rearrange the equation in the algorithm:

$$\begin{aligned}\left\|\phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i)\right\|_2^2 &= \left\langle \phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i), \phi(\mathbf{x}_i) - \frac{1}{|G_j|}\sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle \\ &= \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_i)\rangle - \frac{2}{|G_j|}\left\langle \phi(\mathbf{x}_i), \sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle + \frac{1}{|G_j|^2}\left\langle \sum_{i\in G_j}\phi(\mathbf{x}_i), \sum_{i\in G_j}\phi(\mathbf{x}_i)\right\rangle \\ &= \langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_i)\rangle - \frac{2}{|G_j|}\sum_{k\in G_j}\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_k)\rangle + \frac{1}{|G_j|^2}\sum_{k\in G_j}\sum_{\ell\in G_j}\langle\phi(\mathbf{x}_k),\phi(\mathbf{x}_\ell)\rangle\,.\end{aligned}$$

We can see that all the inner products are usually in the form of:

$$\langle\phi(\cdot),\phi(\cdot)\rangle\,.$$

By using the Kernel trick, we can define ϕ and H implicitly by defining the kernel function.

Definition B.1. For some ϕ, the **kernel function** is a binary operator $\kappa : (\mathbb{R}^n, \mathbb{R}^n) \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\kappa(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle\,.$$

We can then form a generalized K-means algorithm.

Definition B.2. The **Kernel K-means algorithm** is a method that assigns N vectors into K clusters, where each vector is transformed and classified into the cluster with the nearest mean. The steps of the algorithm are as follows:

0: Initialize $G_1, \cdots, G_K$ and define a kernel function κ.

1: For each $\mathbf{x}_i$, assign it to the cluster whose mean is the closest in Euclidean distance after transformation.

$$\begin{aligned} c_i &= \operatorname*{argmin}_{j\in\{1,\cdots,K\}}\left(\kappa(\mathbf{x}_i,\mathbf{x}_i) - \frac{2}{|G_j|}\sum_{k\in G_j}\kappa(\mathbf{x}_i,\mathbf{x}_k) + \frac{1}{|G_j|^2}\sum_{k\in G_j}\sum_{\ell\in G_j}\kappa(\mathbf{x}_k,\mathbf{x}_\ell)\right), && \text{for } i = 1,\cdots,N \\ G_j &= \{i : c_i = j\}, && \text{for } j = 1,\cdots,K.\end{aligned}$$

Repeat until convergence is achieved.

Now the problem switches to determining which kernel function we can choose. We have some necessary conditions.

Definition B.3. The kernel function κ is a **symmetric kernel** if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{y},\mathbf{x}).$$

With a set of vectors $\mathbf{y}_1, \cdots, \mathbf{y}_m \in \mathbb{R}^n$, we can define a matrix by:

$$\boldsymbol{\kappa} = [\kappa(\mathbf{y}_i,\mathbf{y}_j)]_{i,j}.$$

Assume that we have a new vector that is the linear combination of the m transformed vectors. For any $\mathbf{z} \in \mathbb{R}^m$,

$$0 \le \left\langle \sum_{i=1}^m z_i\phi(\mathbf{y}_i), \sum_{i=1}^m z_i\phi(\mathbf{y}_i)\right\rangle = \sum_{i=1}^m\sum_{j=1}^m z_i z_j\langle\phi(\mathbf{y}_i),\phi(\mathbf{y}_j)\rangle = \sum_{i=1}^m\sum_{j=1}^m z_i z_j\kappa(\mathbf{y}_i,\mathbf{y}_j) = \mathbf{z}^T\boldsymbol{\kappa}\mathbf{z}.$$

Definition B.4. The kernel function κ is symmetric positive semi-definite (SPSD) if:

1. $\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{y},\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
2. For any $m > 0$ and $\mathbf{y}_1, \cdots, \mathbf{y}_m \in \mathbb{R}^n$, the matrix:

$$\boldsymbol{\kappa} = [\kappa(\mathbf{y}_i,\mathbf{y}_j)]_{i,j}$$

is symmetric positive semi-definite.

Therefore, by the following theorem, we have a way to determine which mapping can be used as a kernel function.

Theorem B.5. (**Mercer's Theorem**) If the kernel function κ is continuous, symmetric, and positive semi-definite, then there exists a Hilbert space H and a mapping such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle .$$

Example B.1. The traditional kernel, which involves no transformation, is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}, \qquad \phi(\mathbf{x}) = \mathbf{x}, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Applying this kernel function gives the normal K-means algorithm.

Remark B.5.1. Most of the time, finding the feature map ϕ is extremely difficult.

Example B.2. The **polynomial kernel** for $\alpha \in \mathbb{Z}$ and $c \in \mathbb{R}$ is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^\alpha, \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

If $\alpha = 2$, then the feature map is an extremely large vector defined by:

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ \vdots \\ x_n^2 \\ \sqrt{2} x_1 x_2 \\ \vdots \\ \sqrt{2} x_{n-1} x_n \\ \sqrt{2} c x_1 \\ \vdots \\ \sqrt{2} c x_n \\ c \end{pmatrix}, \qquad \text{for } \mathbf{x} \in \mathbb{R}^n.$$

A more general term would require the multinomial theorem.

Example B.3. The **Gaussian kernel** for $\sigma > 0$ is defined by:

$$\kappa(\mathbf{x}, \mathbf{y}) = e^{-\frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{y}\|_2^2} = \exp\left(-\frac{1}{\sigma^2} \|\mathbf{x} - \mathbf{y}\|_2^2 \right), \qquad \text{for } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Obtaining its feature map would require using the Taylor expansion.

Example B.4. Consider that we use the Gaussian kernel to perform the Kernel K-means algorithm. For the same vectors:

$$\kappa(\mathbf{x}_i, \mathbf{x}_i) = \exp\left(-\frac{1}{\sigma^2} \|\mathbf{x}_i - \mathbf{x}_i\|_2^2 \right) = e^0 = 1, \qquad \text{for } i = 1, \cdots, N.$$

For different vectors, we can normalize the distances between two vectors to lie between 0 and 1 via the transformation:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) \begin{cases} \approx 1, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is small} \\ \approx 0, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is large.} \end{cases}$$

Does the kernel fulfill the goal for what we aimed for with the feature map?

$$\begin{aligned} \|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2^2 &= \|\phi(\mathbf{x}_{i_1})\|_2^2 - 2 \langle \phi(\mathbf{x}_{i_1}), \phi(\mathbf{x}_{i_2}) \rangle + \|\phi(\mathbf{x}_{i_2})\|_2^2 \\ &= \kappa(\mathbf{x}_{i_1}, \mathbf{x}_{i_1}) - 2\kappa(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) + \kappa(\mathbf{x}_{i_2}, \mathbf{x}_{i_2}). \end{aligned}$$

Therefore, the distance in transformed data is given by:

$$\|\phi(\mathbf{x}_{i_1}) - \phi(\mathbf{x}_{i_2})\|_2^2 \begin{cases} \approx 0, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is small} \\ \approx 2, & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\|_2 \text{ is large.} \end{cases}$$

Appendix C

Metric Learning

This case study assumes that you have already read Chapter 3.
Suppose we are given N vectors in $\mathbb{R}^n$:

$$\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathbb{R}^n,$$

and a set of vector pairs S and D such that:

$$(\mathbf{x}_i, \mathbf{x}_j) \in \begin{cases} S, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are similar,} \\ D, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are dissimilar (different).} \end{cases}$$

How do we find a metric $\|\cdot\|$ such that:

$$\begin{aligned} &\|\mathbf{x}_i - \mathbf{x}_j\| \text{ is small,} && \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in S, \\ &\|\mathbf{x}_i - \mathbf{x}_j\| \text{ is large,} && \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in D. \end{aligned}$$

1. Representation: How do we represent the norm? We can try using the p-norms with $p \geq 1$:

   $$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

   However, the norm set is too small for us to explore. We can use the more generalized norm induced by the weighted inner product with a symmetric positive definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$:

   $$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}.$$

   The set of all SPD matrices is large enough, though it is not closed. Its closure is the set of all SPSD matrices. However, for any SPSD matrices $\mathbf{A}$ that are not SPD matrices:

   $$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}} = 0, \qquad \mathbf{x}^T \mathbf{A} \mathbf{x} \not\Rightarrow \mathbf{x} = \mathbf{0}.$$

   This violates the positive-definiteness property of a norm. It is still a semi-norm.

   Definition C.1. Let V be a vector space. A **semi-norm** on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that for $\mathbf{x}, \mathbf{y} \in V$:

   1. $\|\mathbf{x}\| \geq 0$,
   2. $\|\alpha \mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$ for $\alpha \in \mathbb{R}$,
   3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

   Since we only want to identify similarity, non-negativity is sufficient for our task.

2. Evaluation: Which SPSD matrices are the best?

   Distance should be small for pairs in S, while distance should be large for pairs in D. Since the distance can only approach 0, but extends to ∞, we can start large and minimize the distance for the pairs in S.

   $$\min_{\mathbf{A} \in \mathbb{R}^{n \times n} \text{ is SPSD}} \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \text{ such that } \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in D} \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}}^2 \geq 1.$$

3. Optimization: We do not perform it here.