Ch. 2. Normed Spaces / Inner product spaces

§ 2.1. Vector spaces

- Definition: A vector space over $\mathbb{R}$ (the real domain) is a set V together with two functions:

  vector addition: $+ : V \times V \to V$ (i.e., $x+y$, where $x, y \in V$)

  scalar multiplication: $\cdot : \mathbb{R} \times V \to V$ (i.e., $\alpha \cdot x$, where $\alpha \in \mathbb{R}, x \in V$)
  or simply αx

  that satisfying the following.

  ① Associativity of addition: $x+(y+z) = (x+y)+z \quad \forall x, y, z \in V$

  ② Commutativity of addition: $x+y = y+x \quad \forall x, y \in V$.

  ③ Zero vector: $\exists$ an element, denoted by 0, in V, s.t.
  $$x+0 = 0+x = x \quad \forall x \in V.$$

  ④ Negative vector: $\forall x \in V$, $\exists$ an element, denoted by $-x$, in V, s.t.
  $$x+(-x) = (-x)+x = 0$$

  ⑤ $\forall x \in V, \quad 1x = x$.

  ⑥ $\forall x \in V$ and $\alpha, \beta \in \mathbb{R}$, $\alpha(\beta x) = (\alpha\beta)x$.

  ⑦ $\forall x \in V$ and $\alpha, \beta \in \mathbb{R}$, $(\alpha+\beta)x = \alpha x + \beta x$

  ⑧ $\forall x, y \in V$ and $\alpha \in \mathbb{R}$, $\alpha(x+y) = \alpha x + \alpha y$ ☒

Remark:
- We can define vector space over $\mathbb{C}$ (the complex domain) similarly.
- We will assume vector space over $\mathbb{R}$ for default. Vector space over $\mathbb{C}$ is used very rarely.

Example 1: $\mathbb{R}$ is a vector space, with "$+$" the standard addition of real numbers and "$\cdot$" the standard multiplication of real numbers.

Example 2: $\mathbb{R}^n$ is a vector space, with "$+$" and "$\cdot$" defined by:

addition: $\forall \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$, $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1+y_1 \\ x_2+y_2 \\ \vdots \\ x_n+y_n \end{bmatrix}$

scalar multiplication: $\forall \alpha \in \mathbb{R}$ and $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, $\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}$.

Many input data can be modeled by vectors in $\mathbb{R}^n$.

- Digital sound signals of length n.
- Time series of length n.
- n different attributes/features of a single thing or object.

Example 3: All real $m \times n$ matrices is a vector space, with standard matrix addition and standard scalar multiplication.

- This vector space is the same as $\mathbb{R}^{mn}$.
- An $m \times n$ matrix can be used to represent a black-white digital image of $m \times n$ pixels.

Example 4: All 3-arrays of size $m \times n \times \ell$ is a vector space, if "+" and "·" is defined by

$+$: $\forall x = [x_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{\ell}$ and $y = [y_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{\ell}$,

$$x + y = [x_{ijk} + y_{ijk}]_{i=1,}^{n}{}_{j=1,}^{m}{}_{k=1}^{\ell}$$

$\cdot$: $\forall x = [x_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{\ell}$ and $\alpha \in \mathbb{R}$,

$$\alpha x = [\alpha x_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{\ell}$$

- This vector space is the same as $\mathbb{R}^{mn\ell}$.
- An $n \times m \times 3$ 3-array can be used to model a color digital image, where x_{ijk} means (i,j)-th pixel in channel k, and Channel 1, 2, 3 means Red, Green, Blue channels of the image.
- An $n \times m \times \ell$ 3-array can be used to model a black-white video, where x_{ijk} means the (i,j)-th pixel at k-th frame.

Example 5: Consider the set of all strings.

Define the addition by, e.g.,

$$'I' + 'am' = 'I\ am'$$

and some scalar multiplication.

Then it doesn't form a vector space.

- Therefore, we cannot use vector space to model text data in this naive way.

Example 6: The function space $C[a,b] = \{ f \mid f \text{ is continous on } [a,b] \}$ is a vector space if we define "$+$" and "$\cdot$" by:

$+ : \forall f, g \in C[a,b], \quad (f+g)(t) = f(t) + g(t), \ \forall t \in [a,b].$

$\cdot : \forall f \in C[a,b], \alpha \in \mathbb{R}, \quad (\alpha f)(t) = \alpha f(t).$

- $C[a,b]$ is referred to as a function space, since any vector in the vector space is a function.
- $C[a,b]$ could be the hypothesis space of a learner with one input and one output, i.e.,

$$x_i \to \boxed{?} \to y_i, \quad \text{with } x_i \in [a,b] \text{ and } y_i \in \mathbb{R}.$$

Learn a $f \in C[a,b]$ s.t. $f(x_i) \approx y_i$ for all i.

§ 1.2. Normed spaces and Banach Spaces

In order to do calculus on vector spaces, we need to define 'distance, closeness between vectors.

Let V be a vector space. Let $x, y \in V$. Then,

$$\text{distance}(x, y) \underset{\uparrow}{=} \text{distance}(x-y, y-y) = \text{distance}(x-y, 0)$$

distance should be shift invariant.

$$\| \quad \text{length of } x-y.$$

Therefore, to define a distance, we only need to define a length for each vector in V.

Let $x \in V$. Let $\|x\|$ be its length, called norm, which should satisfy

① a length should be nonnegative, i.e..

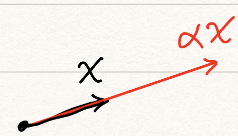
$$\|x\| \geq 0 \quad \forall x \in V.$$

Moreover, only the zero vector has a zero length, i.e.,

$$\|x\| = 0 \iff x = 0.$$

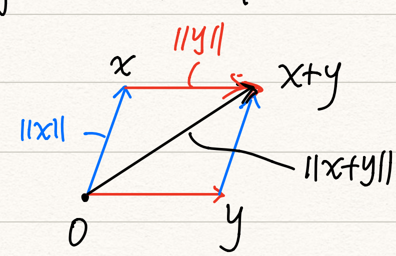
② The length of a multiple of a vector should be the multiple of the length of the vector. i.e.,

$$\forall \alpha \in \mathbb{R}, \quad \|\alpha x\| = |\alpha| \, \|x\|$$



③ Triangular inequality: the length of the direct path is the smallest

$$\|x+y\| \leq \|x\| + \|y\|$$



Definition: Let V be a vector space over $\mathbb{R}$. A norm on V is a function $\|\cdot\| : V \to \mathbb{R}$ such that:

① $\|x\| \geq 0, \forall x \in V$ and $\|x\| = 0 \Leftrightarrow x = 0$.

② $\|\alpha x\| = |\alpha| \, \|x\|, \quad \forall x \in V$ and $\alpha \in \mathbb{R}$.

③ $\|x+y\| \leq \|x\| + \|y\|, \quad \forall x, y \in V$.

Example 1: $\mathbb{R}$ is a vector space over $\mathbb{R}$.

Let $\|x\| = |x| \quad \forall x \in \mathbb{R}$. Then it is a norm on $\mathbb{R}$.

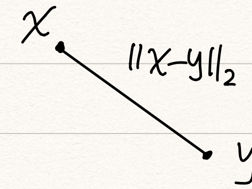
(Can you find other norms on $\mathbb{R}$?)

Example 2: $\mathbb{R}^n$ is a vector space over $\mathbb{R}$.

There are many norms on $\mathbb{R}^n$.

☆ 2-norm: (Euclidean norm)

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$$

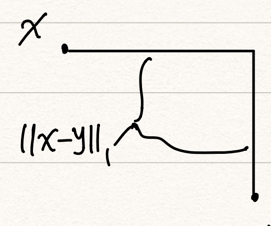


The induced distance

$$\|x-y\|_2 = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}$$ is the Euclidean distance

☆ 1-norm:

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



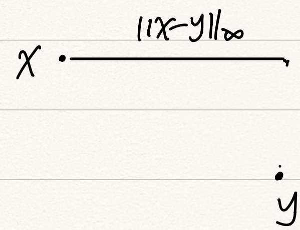
The induced distance

$$\|x-y\|_1 = \sum_{i=1}^{n} |x_i - y_i|$$

is known as Mahanttan distance (You can walk only horizontally and vertically)

☆ ∞-norm:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$



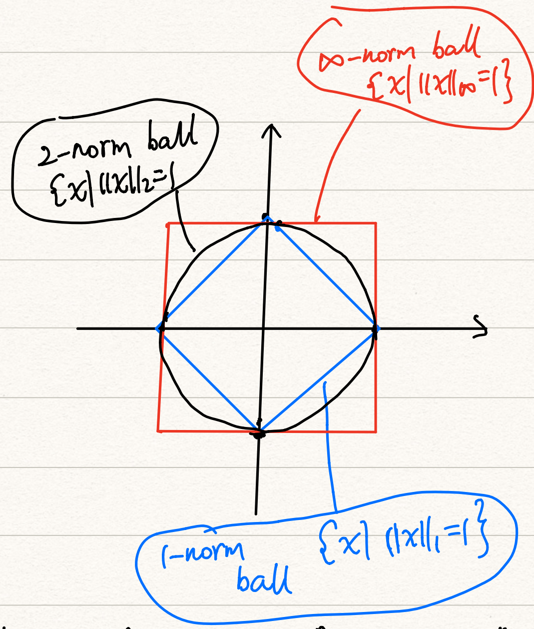
The induced distance is

$$\|x-y\|_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$$

✩ p-norm $(p \geq 1)$

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

- Comparison of unit balls.



- Note that $(\mathbb{R}^n, \|\cdot\|_1)$, $(\mathbb{R}^n, \|\cdot\|_2)$, $(\mathbb{R}^n, \|\cdot\|_\infty)$, ⋯ are all different normed spaces. So, for a given vector space, we can obtain various normed space by choosing different norms.

- Calculate the norms of $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

  $\|x\|_2 = (3^2 + 4^2)^{1/2} = 5$ $\quad \|x\|_1 = |3| + |4| = 7$

  $\|x\|_\infty = \max\{3, 4\} = 4$.

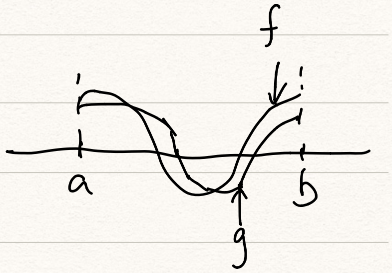
Example 3: $C[a,b]$ is a vector space over $\mathbb{R}$.

The 0 vector in $C[a,b]$ is the function that takes value 0 on $[a,b]$.

To measure how large a function f is,

we can use the following norms

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



Then distance of two functions $f, g \in C[a,b]$

is $\|f - g\|_\infty = \sup_{x \in [a,b]} |f(x) - g(x)|$.

Some other norms of $C[a,b]$ can be

$$\|f\|_1 = \int_a^b |f(x)|\,dx$$
$$\|f\|_2 = \left(\int_a^b |f(x)|^2\,dx\right)^{1/2}$$
$$\|f\|_p = \left(\int_a^b |f(x)|^p\,dx\right)^{1/p}$$

To define calculus, we need first define convergent sequence.

Let V be a normed vector space.

Let $\{x_k\}_{k\in\mathbb{N}}$ be a sequence in V, (i.e., $x_k \in V$ $\forall k = 1, 2, 3, \cdots$).

Let $x \in V$. We say $\{x_k\}_{k\in\mathbb{N}}$ converges to x, denoted by $x_k \to x$, if

$$\lim_{k\to\infty} \|x_k - x\| = 0$$

Example 1: Consider $\mathbb{R}^n$ with $\|\cdot\|_2$.

Let $x_k = \begin{pmatrix} 1/k \\ 2/k \\ \vdots \\ n/k \end{pmatrix} \in \mathbb{R}^n$ $\forall k$ and $x = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \equiv 0$

Then, $$\|x_k - x\|_2 = \left\| \begin{pmatrix} 1/k \\ 2/k \\ \vdots \\ n/k \end{pmatrix} \right\|_2 = \left(\left(\tfrac{1}{k}\right)^2 + \left(\tfrac{2}{k}\right)^2 + \cdots + \left(\tfrac{n}{k}\right)^2\right)^{1/2}$$
$$= \frac{1}{k}\left(1^2 + 2^2 + \cdots + n^2\right)^{1/2} = \frac{F(n)}{k}$$ a function of n.

$$\lim_{k\to\infty} \|x_k - x\|_2 = \lim_{k\to\infty} \frac{F(n)}{k} = F(n)\cdot\left(\lim_{k\to\infty} \frac{1}{k}\right) = 0.$$

Therefore, $x_k \to x$ as $k \to \infty$.

Example 2: Consider $C[0,1]$ with $\|\cdot\|_\infty$

Let $f_k(t) = \sin(2\pi k t)/k^2$

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

$$\lim_{k\to\infty} \|f_k - 0\|_\infty = \lim_{k\to\infty} \frac{1}{k^2} = 0$$

So, $f_k \to 0$ as $k \to \infty$.

Unfortunately, not all normed space is not "closed" under limit operation.

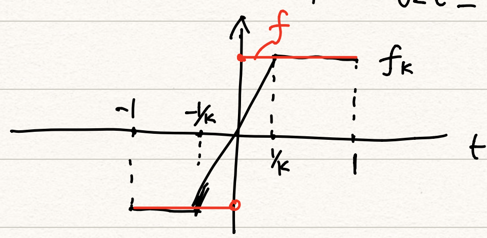
Example: Consider $C[-1,1]$ with $\|\cdot\|_1$ (Recall $\|f\|_1 = \int_{-1}^{1} |f(t)|\,dt$)

Define f_k as

$$f_k(t) = \begin{cases} -1 & -1 \le t \le -\frac{1}{k} \\ kt & \\ 1 & \frac{1}{k} \le t \le 1 \end{cases}$$

and f as

$$f(t) = \begin{cases} -1 & -1 \le t < 0 \\ 1 & 0 \le t \le 1 \end{cases}$$



Then $\|f_k - f\|_1 = \int_{-1}^{1} |f_k(t) - f(t)|\,dt = \frac{1}{k}$

and $\lim_{k\to\infty} \|f_k - f\|_1 = \lim_{k\to\infty} \frac{1}{k} = 0.$

So $f_k \to f$ as $k \to \infty$

$f_k \in C[-1,1]$, but $f \notin C[-1,1]$.

Therefore, $(C[-1,1], \|\cdot\|_1)$ is not complete.

We call a complete normed space a Banach space.

Examples of Banach spaces:

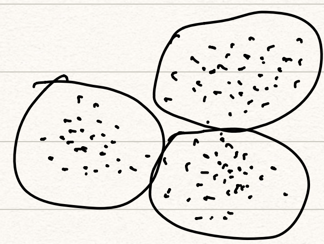
- $\mathbb{R}^n$ with any norm
- $C[a,b]$ with $\|\cdot\|_\infty$

§ 1.3 Case study: Clustering, K-means, K-medians

Clustering

Suppose we are given N vectors $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$

The goal of clustering is to group or partition the vectors into k groups or clusters, with the vectors in each group close to each other.



- We use $\mathbb{R}^n$ because it is simple yet able to model a variety of data sets (e.g., singals, images, videos, attributes of things)
- Actually, the methods can be extended to any normed spaces.
- Applications:
  - Topic discovery. Suppose the N vectors are word histograms with N documents respectively, i.e., the j-th component in x_i is the counts of the j-th word in document i.

    A clustering algorithm patition the documents into k groups, which typically can be intepreted as groups of documents with the same topics, genre, or author.
  - Patient Clustering. If $\{x_i\}_{i=1}^N$ are feature vectors associated with N patients admitted to a hospital, a clustering algorithm clusters the patient into k groups of similar patients.
  - Recommendation system. A group of N people respond to ratings of n movies. A clustering algorithm can be used to cluster the people into k groups, each with similar taste.

Then we can recommend new movies liked by someone to people in the same group as him/her.

— Many other applications.

Mathematical formulation:

- Representation:

Let $C_i \in \{1, 2, \cdots, k\}$ be the group that x_i belongs to. $i = 1, 2, \cdots, N$.

Then, group j, denoted by G_j, is $G_j = \{ i \mid C_i = j \}$. $j = 1, 2, \cdots, k$.

We assign each group a representative vector, denoted by $z_1, z_2, \cdots, z_K$.

The representative vectors are not necessarily one of given vectors.

- Evaluation:

First of all, within one specific group j G_j, all vectors should be close to the representative vector z_j. More precisely, let

$$J_j = \sum_{i \in G_j} \| x_i - z_j \|_2^2$$

Then, J_j should be small.

Secondly, consider all groups, since each J_j is small,

$$J = J_1 + J_2 + \cdots + J_K$$

should be small.

Altogether, we solve the following

$$\min_{\substack{G_1, \cdots, G_K \\ z_1, \cdots, z_K}} J \iff \min_{\substack{G_1, \cdots, G_K \\ z_1, \cdots, z_K}} \sum_{j=1}^{K} J_j \iff \min_{\substack{G_1, \cdots, G_K \\ z_1, \cdots, z_K}} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \| x_i - z_j \|_2^2 \right)$$

- Optimization.

We may use an alternating minimization to solve the minimization.

Step 1: Fix the representatives $z_1, \cdots, z_K$, find the best partitions $G_1, \cdots, G_K$, i.e., solve

$$\min_{G_1, \cdots, G_K} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \| x_i - z_j \|_2^2 \right) \quad \cdots\cdots (1)$$

Step 2: Fix the groups $G_1, \cdots, G_K$, find the best representatives $z_1, \cdots, z_K$, i.e., solve

$$\min_{z_1,\cdots,z_K} \sum_{j=1}^{K}\left(\sum_{i\in G_j} \|x_i - z_j\|_2^2\right) \quad - - - - \quad (2)$$

The two steps are repeated until convergence.

Let's find the solutions of the sub-problems (1) and (2) respectively.

For (1):

finding the partition $G_1, \cdots, G_K$ is equivalent to finding $c_1, c_2, \cdots, c_N$. So (1) becomes

$$\min_{c_1,c_2,\cdots,c_N}\left(\underbrace{\|x_1 - z_{c_1}\|_2^2}_{\substack{\text{depends on}\\ c_1 \text{ only}}} + \underbrace{\|x_2 - z_{c_2}\|_2^2}_{\substack{\text{depends on}\\ c_2 \text{ only}}} + \cdots + \underbrace{\|x_N - z_{c_N}\|^2}_{\substack{\text{depends}\\ \text{on } c_N \text{ only}}}\right)$$

$$\Updownarrow$$

$$\min_{c_i} \|x_i - z_{c_i}\|_2^2 \qquad i = 1, 2, \cdots, N.$$

Since $c_i \in \{1, 2, \cdots, k\}$, to get c_i, we only need to compare

$$\|x_i - z_1\|_2^2, \ \|x_i - z_2\|_2^2, \ \cdots, \ \|x_i - z_k\|_2^2$$

and choose the minimum from it. i.e.,

$$c_i = \arg\min_{j\in\{1,\cdots,k\}} \|x_i - z_j\|_2^2, \quad i = 1, 2, \cdots, k.$$

In other words,

$$\boxed{\text{x_i is assigned to the group whose representative vector is the closest to x_i.}}$$

For (2): It is rewritten as

$$\min_{z_1,\cdots z_K}\left(\underbrace{\sum_{i\in G_1}\|x_i - z_1\|_2^2}_{\substack{\text{depends on } z_1\\ \text{only}}} + \underbrace{\sum_{i\in G_2}\|x_i - z_2\|_2^2}_{\substack{\text{depends on}\\ z_2 \text{ only}}} + \cdots + \underbrace{\sum_{i\in G_K}\|x_i - z_k\|_2^2}_{\substack{\text{depends on}\\ z_K \text{ only}}}\right)$$

Obviously, it is equivalent to minimize each term independently, i.e., solve K independent problems.

$$\min_{z_j} \left(\sum_{i \in G_j} \|x_i - z_j\|_2^2 \right), \qquad j = 1, 2, \cdots, K.$$

Note that

$$\sum_{i \in G_j} \|x_i - z_j\|_2^2 = \sum_{i \in G_j} \sum_{l=1}^{n} (x_{il} - z_{jl})^2$$

$$= \sum_{l=1}^{n} \left(\sum_{i \in G_j} (x_{il} - z_{jl})^2 \right)$$

(x_{il}, z_{jl} are l-th component of x_i, z_j respectively)

Eeah term is this summation are independent again.

Thus, $\min_{z_j} \left(\sum_{i \in G_j} \|x_i - z_j\|_2^2 \right) \iff \boxed{\min_{z_{jl}} \sum_{i \in G_j} (x_{il} - z_{jl})^2}, \quad l = 1, 2, \cdots, n.$

one variable minimization.

Taking derivative w.r.t. z_{jl} and setting it to 0, we obtain that the solution z_{jl} satisfies

$$2 \sum_{i \in G_j} (z_{jl} - x_{il}) = 0$$

$$\Rightarrow \quad z_{jl} = \left(\sum_{i \in G_j} x_{il} \right) / |G_j| \qquad (|G_j| \text{ is the number of elements in } G_j)$$

$$l = 1, 2, \cdots, n.$$

In vector form,

$$\begin{pmatrix} z_{j1} \\ z_{j2} \\ \vdots \\ z_{jn} \end{pmatrix} = \frac{1}{|G_j|} \cdot \sum_{i \in G_j} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix} \iff z_j = \frac{1}{|G_j|} \left(\sum_{i \in G_j} x_i \right)$$

In other words,

$\boxed{z_j \text{ is the mean of all vectors in } G_j.}$

Altogether, we get the following clustering algorithm.

Input: $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$.

Output: $c_1, c_2, \cdots, c_N$ and z_j, $j = 1, \cdots, K$.

Initialization: Initialize $z_1, z_2, \cdots, z_K$ by choosing K vectors from $x_1, x_2, \cdots, x_N$ randomly.

Step 1: Given $z_1, z_2, \cdots, z_K$, compute

$$C_i = \arg\min_{j \in \{1,2,\cdots,K\}} \|x_i - z_j\|_2^2, \quad i = 1, 2, \cdots, N.$$

and define

$$G_j = \{ i \mid C_i = j \}, \quad j = 1, 2, \cdots, K.$$

step 2: Given $G_1, G_2, \cdots, G_K$, compute

$$z_j = \frac{1}{|G_j|} \left(\sum_{i \in G_j} x_i \right)$$

Go back to step 1.

This algorithm is know as "K-means" algorithm, because it computes K means of vectors at step 2.

K-medians Algorithm

In K-means, the Euclidean norm is used. We can replace it by 1-norm. We solve

$$\min_{\substack{G_1, \cdots, G_K \\ z_1, \cdots, z_K}} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \|x_i - z_j\|_1 \right)$$

The numerical solver is

Step 1: Fix $z_1, \cdots, z_K$, solve

$$\min_{G_1, \cdots, G_K} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \|x_i - z_j\|_1 \right).$$

Similar to the discussion in K-means, the solution is

$$C_i = \arg\min_{j \in \{1,2,\cdots,K\}} \|x_i - z_j\|_1, \quad i = 1, 2, \cdots, N.$$

and $G_j = \{ i \mid C_i = j \}$.

Step 2: Fix $G_1, G_2, \cdots, G_K$, solve

$$\min_{z_1, \cdots, z_K} \sum_{j=1}^{K} \left(\sum_{i \in G_j} \|x_i - z_j\|_1 \right)$$

Similar to the discussion in K-means, it is decomposed into K sub problems

$$\min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_1 \qquad j = 1, 2, \cdots, K.$$

It is well known (Galileo) that the solution is

$$z_j = \underset{i \in G_j}{\mathrm{median}} (x_i).$$

where $\underset{i \in G_j}{\mathrm{median}} (x_i)$ takes component-wise median.

This algorithm is called "k-median" algorithm.

§1.4 Inner product / Hilbert space

Norms give only metrics, i.e, measuring the distance of two vectors.

However, in many applications, the "angel" of two vectors matter.

- For example: two images x, y showing the same scene with different lights.

  For simplicity, we may assume, say, $y = \frac{1}{2}x$ (the first image is with 100% light, and the second 50%)

  Then $\|x-y\| = \frac{1}{2}\|x\|$ — not small.

  but x, y are from the same scene.

- We use inner product for "angel" of two vectors

Definition: A function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ is called an inner product on the vector space V over $\mathbb{R}$ if

① $\forall x \in V, \quad \langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0 \Leftrightarrow x = 0$

② $\langle \alpha x_1 + \beta x_2, y\rangle = \alpha\langle x_1, y\rangle + \beta\langle x_2, y\rangle \quad \forall \alpha, \beta \in \mathbb{R}, \ x_1, x_2, y \in V.$

③ $\langle x, y\rangle = \langle y, x\rangle$

Remark: 1. By ② and ③, $\langle x, \alpha y_1 + \beta y_2\rangle = \alpha\langle x, y_1\rangle + \beta\langle x, y_2\rangle \quad \forall \alpha, \beta \in \mathbb{R}, \ x, y_1, y_2 \in V.$

Therefore, $\langle\cdot,\cdot\rangle$ is a bi-linear function, i.e., it is linear with respect to one of the variable with the other fixed.

2. For inner product of vector spaces on $\mathbb{C}$, we only need to change ③ to

③' $\langle x, y\rangle = \overline{\langle y, x\rangle}$, where $\overline{\cdot}$ stands for complex conjugate.

Example 1: $\mathbb{R}^n$ is a vector space. We can define an inner product as

$$\langle x, y\rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n, \text{ where } x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \text{ and } y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

$(\equiv x^T y)$

Example 2: Another inner product in $\mathbb{R}^n$ is as follows.

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite (spd)

(Recall spd means: $A^T = A$ and $x^T A x > 0 \ \forall x \neq 0$)

Then $\langle x, y \rangle_A = x^T A y$ defines an inner product in $\mathbb{R}^n$, because

① $\langle x, x \rangle_A = x^T A x \geq 0$ and $\langle x, x \rangle_A = 0 \Leftrightarrow x^T A x = 0 \Leftrightarrow x = 0$.

② $\langle \alpha x_1 + \beta x_2, y \rangle_A = (\alpha x_1 + \beta x_2)^T A y = \alpha x_1^T A y + \beta x_2^T A y$

$= \alpha \langle x_1, y \rangle_A + \beta \langle x_2, y \rangle_A$.

③ $\langle x, y \rangle_A = x^T A y = (x^T A y)^T = y^T A^T x = y^T A x = \langle y, x \rangle_A$.

Example 3: In $\mathbf{C}[a,b]$, we can define an inner product as

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx, \quad \forall f, g \in C[a,b].$$

Cauchy-Schwartz Inequality:

> If $\langle \cdot, \cdot \rangle$ is an inner product on V, then, for any $x, y \in V$,
> $$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$
> The equality holds true if and only if $x = \alpha y$ for some $\alpha \in \mathbb{R}$.

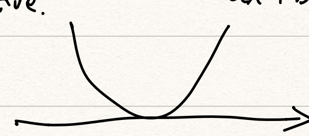
Proof. Let $\lambda \in \mathbb{R}$ be an arbitrary number

$$0 \leq \langle x + \lambda y, x + \lambda y \rangle = \langle x, x \rangle + \lambda \langle y, x \rangle + \lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle$$
$$= \langle x, x \rangle + 2\lambda \langle x, y \rangle + \lambda^2 \langle y, y \rangle$$

Thus, $\lambda^2 \langle y, y \rangle + 2\lambda \langle x, y \rangle + \langle x, x \rangle \geq 0$. $\quad \forall \lambda \in \mathbb{R}$.

The left is a quadratic function of λ and always non-negative.

There is at most one root of the quadratic function.



$a\lambda^2 + b\lambda + c \geq 0$

at most one root, So $b^2 - 4ac \leq 0$.

So, $(2\langle x, y \rangle)^2 - 4\langle y, y \rangle \langle x, x \rangle \leq 0$

$\Rightarrow \langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle$

Finally, when $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$, there is a root, i.e.,

$\exists$ a unique $\lambda \in \mathbb{R}$, $\lambda^2 \langle y, y \rangle + 2\lambda \langle x, y \rangle + \langle x, x \rangle = 0$

$$\Updownarrow$$

$\exists$ a unique $\lambda \in \mathbb{R}$, $\langle x+\lambda y, x+\lambda y \rangle = 0$

$$\Updownarrow$$

$\exists$ a unique $\lambda \in \mathbb{R}$, $x+\lambda y = 0$

$$\Updownarrow$$

$\exists$ a unique $\lambda \in \mathbb{R}$, $x = -\lambda y$ $\boxtimes$

With the Cauchy-Schwartz inegality, we can show that:

$\boxed{\|x\| = \langle x, x\rangle^{1/2} \text{ defines a norm.}}$ — Called "norm induced by the inner product".

Proof. ① $\|x\| = (\langle x,x\rangle)^{1/2} \geq 0$ and $\|x\| = (\langle x,x\rangle)^{1/2} = 0 \Leftrightarrow x = 0$.

② $\|\alpha x\| = \langle \alpha x, \alpha x\rangle^{1/2} = (\alpha^2 \langle x,x\rangle)^{1/2} = |\alpha| \, \|x\|$

③ $\|x+y\|_2^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle$

$$\begin{aligned} &= \|x\|^2 + \|y\|^2 + 2\langle x,y\rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2\|x\|\,\|y\| \\ &= (\|x\|+\|y\|)^2 \end{aligned}$$

(Note that Cacuchy-Schwarts becomes $|\langle x,y\rangle| \leq \|x\|\,\|y\|$)

$\boxtimes$

Cauchy-Schwartz is restated as:

$$\boxed{|\langle x,y\rangle| \leq \|x\|\,\|y\|}$$

"Angels" in inner product spaces.

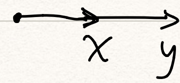
The equality "=" is attained if and only if x and y are aligned exactly, which should have the least angel. Therefore, we use the ratio of the two sides of Cauchy-Schwartz

$$\frac{\langle x,y\rangle}{\|x\|\,\|y\|}$$

to quantitize the closeness to exact alignment. Hence to define the angels between x and y.

- When $x = \alpha y$ with $\alpha > 0$,

$$\frac{\langle x, y \rangle}{\|x\| \, \|y\|} = 1$$



Since x and y are in the same direction,
It is naturally to define $\angle(x, y) = 0$

- when $x = \alpha y$ with $\alpha < 0$

$$\frac{\langle x, y \rangle}{\|x\| \, \|y\|} = -1$$

Since x and y are in the opposite direction,
it is naturally to define $\angle(x, y) = \pi$.

- We define $\cos \angle(x, y) = \dfrac{\langle x, y \rangle}{\|x\| \, \|y\|}$

or, equivalently, $\angle(x, y) = \arccos \dfrac{\langle x, y \rangle}{\|x\| \, \|y\|}$

This definition Coincides with the above two cases and the vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ equipped with the standard inner product.

Orthogonality: Let V be an inner product space.

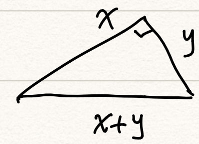
- When the angel of x and y is $\frac{\pi}{2}$, we call they are orthogonal, denoted by $x \perp y$, i.e.,

$$\boxed{x \perp y \quad \text{if} \quad \langle x, y \rangle = 0}$$

- when $x \perp y$, they are least relevant.

- Pythagoras' theorem: Let x, y be two vectors in an inner product space.

If $x \perp y$, then $\|x+y\|^2 = \|x\|^2 + \|y\|^2$



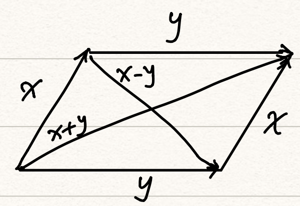
proof. $\|x+y\|^2 = \langle x+y, x+y \rangle$

$= \langle x,x \rangle + \langle x,y \rangle + \langle y,x \rangle + \langle y,y \rangle$

$= \|x\|^2 + \|y\|^2$ #

- Parallelogram law: $\forall x, y \in H$ — an inner product space,

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$



proof. $\|x+y\|^2 + \|x-y\|^2$

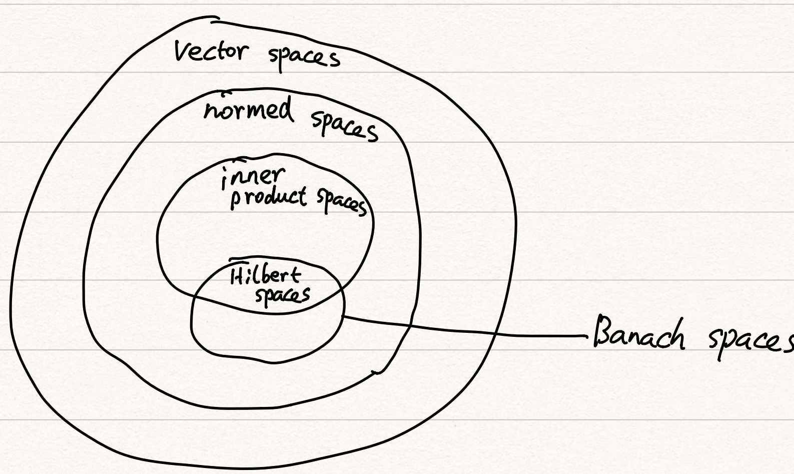
$= \langle x+y, x+y \rangle + \langle x-y, x-y \rangle$

$= \langle x,x \rangle + \langle y,x \rangle + \langle x,y \rangle + \langle y,y \rangle + \langle x,x \rangle - \langle y,x \rangle - \langle x,y \rangle + \langle y,y \rangle$

$= 2\|x\|^2 + 2\|y\|^2$ #

Hilbert space:

A Hilbert Space is a Banach space in which the norm is induced by an inner product.



Example 1: $\mathbb{R}^n$ with inner product

$$\langle x, y \rangle = x^T y$$

is a Hilbert space.

Example 2: $\mathbb{R}^n$ with inner product

$$\langle x, y \rangle_A = x^T A y, \quad \text{where } A \text{ is an spd matrix}$$

is a Hilbert space. The norm on this space is

$$\|x\|_A = (x^T A x)^{\frac{1}{2}}.$$

Example 3: $C[a,b]$ with inner product

$$\langle f, g \rangle = \int_a^b f(x) g(x) dx$$

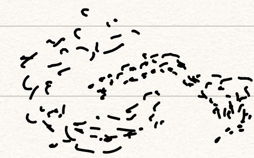
is NOT a Hilbert space, because it is not complete. (Roughly, "the" limit of a convergent sequence may not be in $C[a,b]$)

To complete $C[a,b]$ under the norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{\frac{1}{2}}$, we need to extend the Riemmanian integral to the so-called Lebesgue integral, and the resulting Hilbert space is $L^2(a,b)$.

In the following, we will consider calculus on Hilbert/Banach spaces.

§ 1.5 Case Study: Kernel trick , Kernel k-means

The k-means will not work for the following examples



Recall in a clustering, we want to group $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$ into K groups.

The k-means algorithm works like:

Initialize $z_1, z_2, \cdots, z_K$

step 1: Given $z_1, z_2, \cdots, z_K$, update the groups $G_1, \cdots, G_K$ by

① for each x_i, assign c_i, the group that x_i belongs to, by

$$c_i = \arg\min_{j \in \{1,\cdots,K\}} \|x_i - z_j\|_2^2$$

② Then $G_j = \{i \mid c_i = j\}$, for $j = 1, 2, \cdots, K$.

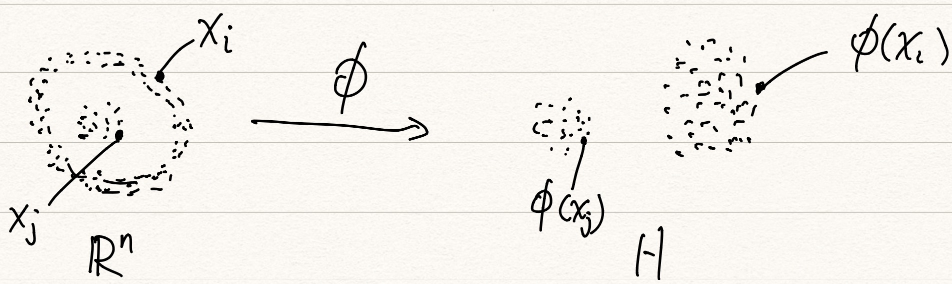
step 2: Given $G_1, \cdots, G_K$, update their representives by

$$z_j = \frac{1}{|G_j|}\left(\sum_{i \in G_j} x_i\right), \quad \text{for } j = 1, 2, \cdots, K.$$

To modify k-means to those "curved" data sets in $\mathbb{R}^n$

we use a transform to "un-curve" the data sets in a Hilbert space.

Let $\phi: \mathbb{R}^n \to H$



$\phi(x_i)$ is called the feature of x_i.
ϕ is called the feature map.
H is called the feature space.

Then we apply K-means to $\phi(x_1), \phi(x_2), \cdots, \phi(x_N)$ in H.
Let $z_1, \cdots, z_K$ be the representative vectors in H.

step 1: Given $z_1, \cdots, z_K$,

$$c_i = \arg\min_{j \in \{1,2,\cdots,K\}} \|\phi(x_i) - z_j\|^2, \quad i=1,2,\cdots,N.$$

and

$$G_j = \{i \mid c_i = j\}, \quad j=1,2,\cdots,K.$$

step 2: Given $G_1, \cdots, G_K$,

$$z_j = \frac{1}{|G_j|}\left(\sum_{i \in G_j} \phi(x_i)\right)$$

Repeat.

However, finding the feature map ϕ is not easy, because ϕ depends on the shape of $x_1, x_2, \cdots, x_N$, which generally is very complicated.

The good news is that:

There is no need to know ϕ explicitly in the K-means algorithm.

This is seen as in below:

- First of all, since we care only the groups of $x_1, \cdots, x_N$, we only need to know $G_1, \cdots, G_K$. The representatives $z_1, \cdots, z_K$ are only intemediate.
  Therefore, we can eliminate $z_1, \cdots, z_K$ in the K-means algorithm.

$$\boxed{\begin{array}{ll} ① & C_i = \arg\min\limits_{j\in\{1,2,\cdots,k\}} \left\| \phi(x_i) - \frac{1}{|G_j|}\sum\limits_{l\in G_j}\phi(x_l)\right\|^2, \quad i=1,2,\cdots,N. \\ ② & G_j \leftarrow \{i \mid C_i = j\}, \qquad j=1,2,\cdots,k. \end{array}}$$

- Now, only ① involves the feature mapping ϕ. Since H is a Hilbert space, we can expand the norm in ① by

$$\left\|\phi(x_i) - \frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l)\right\|^2$$

$$= \left\langle \phi(x_i) - \frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l),\ \phi(x_i) - \frac{1}{|G_j|}\sum_{l\in G_j}\phi(x_l)\right\rangle$$

$$= \langle \phi(x_i), \phi(x_i)\rangle - \frac{2}{|G_j|}\sum_{l\in G_j}\langle \phi(x_i), \phi(x_l)\rangle$$

$$+ \frac{1}{|G_j|^2}\sum_{l_1\in G_j}\sum_{l_2\in G_j}\langle \phi(x_{l_1}), \phi(x_{l_2})\rangle$$

We see that

$\boxed{\text{Only inner products in the feature space are involved.}}$

Therefore,

$\boxed{\text{An explicit expression of } \phi \text{ is NOT necessary.}}$

Kernel trick:

> Instead of defining $\phi(x)$ explicitly, we define a kernel function $K(x,y)$, which satisfies $K(x,y) = \langle \phi(x), \phi(y)\rangle$.
> The kernel function $K(x,y)$ can be seen as an explicit quantification of similarity of x and y.

Not all function $K(x,y)$ satisfying $K(x,y) = \langle \phi(x), \phi(y) \rangle$ for some feature map ϕ. (Example: $K(x,y) = -1$ is not good because it violates with $\langle \phi(x), \phi(x) \rangle \geq 0$)

Which function $K(x,y)$ can be an inner product $\langle \phi(x), \phi(y) \rangle$ for some feature mapping ϕ?

- First of all,

$$K(x,y) = \langle \phi(x), \phi(y) \rangle \overset{\text{inner product property}}{=} \langle \phi(y), \phi(x) \rangle = K(y,x).$$

(We say $K(\cdot,\cdot)$ is symmetric if $K(x,y) = K(y,x)$ for all $x, y \in \mathbb{R}^n$).

- Secondly, let $y_1, y_2, \cdots, y_m$ be m vectors in $\mathbb{R}^n$, then,

for any $c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix} \in \mathbb{R}^m$,

$$\left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{i=1}^m c_i \phi(y_i) \right\rangle \geq 0 \quad \text{(By inner product property)}$$

On the other hand,

$$\left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{i=1}^m c_i \phi(y_i) \right\rangle = \left\langle \sum_{i=1}^m c_i \phi(y_i), \sum_{j=1}^m c_j \phi(y_j) \right\rangle$$

By inner product property.

$$= \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \phi(y_i), \phi(y_j) \rangle$$

$$= \sum_{i=1}^m \sum_{j=1}^m c_i c_j K(y_i, y_j)$$

$$= c^T \begin{bmatrix} K(y_1,y_1) & K(y_1,y_2) & \cdots & K(y_1,y_m) \\ K(y_2,y_1) & K(y_2,y_2) & \cdots & K(y_2,y_m) \\ \vdots & \vdots & \vdots & \vdots \\ K(y_m,y_1) & K(y_m,y_2) & \cdots & K(y_m,y_m) \end{bmatrix} c \geq 0. \quad \forall c \in \mathbb{R}^m.$$

In other words, the matrix is symmetric positive semi-definite.

We say a function $K(\cdot,\cdot): \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is symmetric positive semi-definite if:

i) $K(x,y) = K(y,x) \quad \forall x, y \in \mathbb{R}^n$

ii) For any m and any vectors $y_1, y_2, \cdots, y_m \in \mathbb{R}^n$, the matrix

$$\begin{bmatrix} K(y_1,y_1) & K(y_1,y_2) & \cdots & K(y_1,y_m) \\ K(y_2,y_1) & K(y_2,y_2) & \cdots & K(y_2,y_m) \\ \vdots & \vdots & \vdots & \vdots \\ K(y_m,y_1) & K(y_m,y_2) & \cdots & K(y_m,y_m) \end{bmatrix}$$

is symmetric positive semi-definite.

Mercer's theorem tells us that: If a kernel function $K(\cdot,\cdot)$ is symmetric positive semi-definite, then there exists a feature map ϕ such that $K(x,y) = \langle \phi(x), \phi(y) \rangle$

Some popular kernels:

① $K(x,y) = x^T y$ ($\phi(x) = x$. No transform)

② $K(x,y) = (x^T y)^\alpha$ polynomial kernels

③ $K(x,y) = e^{-\frac{\|x-y\|_2^2}{\sigma^2}}$ Gaussian kernel

Kernel k-means algorithm

- choose a kernel function $K(\cdot,\cdot)$
- Initialize $G_1, G_2, \cdots, G_K$ by, e.g., one step of k-means.
- Set

$$c_i = \arg\min_{j\in\{1,2,\cdots,K\}} \left(K(x_i,x_i) - \frac{2}{|G_j|}\sum_{\ell\in G_j} K(x_i,x_\ell) + \frac{1}{|G_j|^2}\sum_{\ell_1\in G_j}\sum_{\ell_2\in G_j} K(x_{\ell_1},x_{\ell_2}) \right)$$

for $i=1,2,\cdots,N$.

- update $G_1, G_2, \cdots, G_K$ by

$$G_j = \{i \mid c_i = j\}, \text{ for } j=1,2,\cdots,K.$$

- go back and repeat

The kernel k-means works for some datasets for which k-means fail.

Example:

If we use Gaussian kernel $K(x,y) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$

then

- $K(x_i, x_i) = e^{-\frac{\|x_i - x_i\|_2^2}{\sigma^2}} = 1$

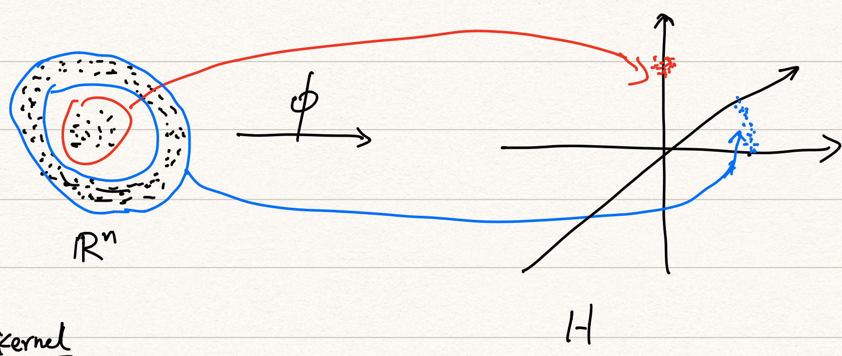
  — so all $\phi(x_1), \ldots, \phi(x_N)$ are on unit sphere in H.

- $K(x_i, x_j) \begin{cases} \approx 0 & \text{if } \|x_i - x_j\|_2 \text{ is large} \\ \approx 1 & \text{if } \|x_i - x_j\|_2 \text{ is small.} \end{cases}$

  — so $\phi(x_i)$, $\phi(x_j)$ are orthogonal in H if $\|x_i - x_j\|_2$ large.

  — $\phi(x_i) \approx \phi(x_j)$ in H if $\|x_i - x_j\|$ small.

Therefore,



Thus, kernel K-means works for this data set.