Ch2 - Vector Space

2020年9月21日　上午 10:46

Ch2_Vec...

Ch. 2. Vector spaces, metric and convergence

§ 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., $\alpha \cdot x$ or simply αx, where $\alpha \in \mathbb{R}, x \in V$)

  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 $\alpha, \beta \in \mathbb{R}, \quad \alpha(\beta x) = (\alpha\beta)x$.

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

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

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. Stock price in base 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 $\mathbb{R}^{mn}$. A long vector
- An $m \times n$ matrix can be used to represent a black-white digital image of $m \times n$ pixels.

All real 3-array(tensor) of size mxnxl is a vector space

Example 4: All 3-arrays of size $m \times n \times l$ is a vector space, if "$+$" and "$\cdot$" is defined by Three dimensional matrix

$+$: $\forall X=[x_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{l}$ and $Y=[y_{ijk}]_{i=1}^{n}{}_{j=1}^{m}{}_{k=1}^{l}$, $1<=i<=n, 1<=j<=m, 1<=k<=l$

$$X+Y = [x_{ijk}+y_{ijk}]_{i=1,}^{n}{}_{j=1,}^{m}{}_{k=1}^{l}$$

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

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

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

$\forall X, Y \in \mathbb{R}^{m \times n}$, $X+Y = \begin{bmatrix} x_{11}+y_{11} & x_{12}+y_{12} & \cdots \\ x_{21}+y_{21} & \cdots & \cdots \\ \vdots & & \\ x_{m1}+y_{m1} & \cdots & \cdots \end{bmatrix}$

$\forall \alpha \in \mathbb{R}$

$\forall X \in \mathbb{R}^{m \times n}$, $\alpha X = \begin{bmatrix} \alpha x_{11} & \cdots & \alpha x_{1n} \\ \vdots & & \vdots \end{bmatrix}$

$\forall X \in \mathbb{R}^{m\times n}$, $\alpha X = \begin{bmatrix} \alpha x_{11} & \cdots & \alpha x_{1n} \\ \vdots & & \vdots \\ \alpha x_{m1} & \cdots & \alpha x_{mn} \end{bmatrix}$

Example 5: Consider the set of all strings.

Define the addition by, e.g.,

'I ' + 'am ' = 'I am '

and some scalar multiplication.

Not commutive!

Then it doesn't form a vector space.

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

  How to "vectorize" the texts is a fundamental research topic in text data analysis and natural language processing.

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 "·" 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.,

In unsupervised learning, 從hypothesis space 穩適合嘅function

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

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

Example 7: The infinite sequence

$$l_\infty = \left\{ \begin{pmatrix} a_1 \\ a_2 \\ \vdots \end{pmatrix} \,\middle|\, \exists \text{ a finite } C \text{ s.t } |a_i| \le C \ \forall i \right\}$$

with: "+": $(a+b)_i = a_i + b_i$ for all $a, b \in l_\infty$ for all i

"·": $(\alpha a)_i = \alpha a_i$ for all $a \in l_\infty$, for all i, $\alpha \in \mathbb{R}$, 綁 $\exists$ a finite C s.t. $|\alpha a_i| \le C$

It forms a vector space.

因為+完同乘完constant之後都會被bounded,所以成立

- This vector space can be used to model, e.g., time series with a very long time and/or a very fine time resolution.

A very long time signal or very very fine resolution

§1.2. Metric in vector 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,

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

(distance should be shift invariant.) $\quad = \|x-y\|$ length of $x-y$.

Shift invariant = 搬x同y去其他地方都唔應該影響呢個distance

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 non negative, i.e.,

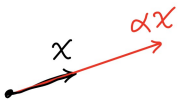
$$\|x\| \geq 0 \quad \forall x \in V.$$

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

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

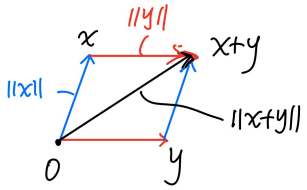
② The length of a multiple of a vector should be the multiple of the length of the vector. i.e.,

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



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 $\mathbb{R}$.

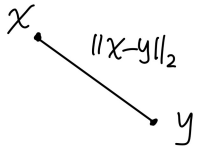
(Can you find other norms on $\mathbb{R}$?) 2 * absolute就係

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

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

★ 2-norm: (Euclidean norm)

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

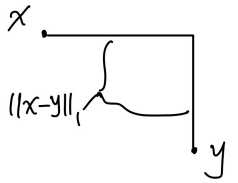


The induced distance

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

★ 1-norm:

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



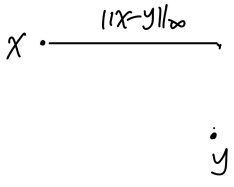
The induced distance

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

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

★ ∞-norm:

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



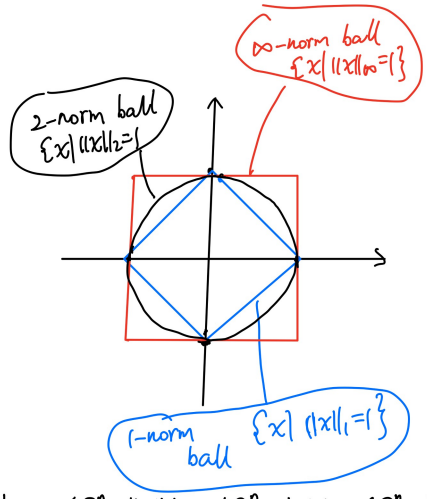
The induced distance is

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

* 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 $x = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

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

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

Example 3: $\mathbb{R}^{m\times n}$ is a vector space

- Since $\mathbb{R}^{m\times n}$ can be viewed as $\mathbb{R}^{mn}$, we can define vector p-norms for matrices

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

- $p=1$: sum of absolute values of entries of A.
- $p=2$: sum of squares and then take square root (also known as Frobenius norm)

$$\|A\|_F = \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |a_{ij}|^2\right)^{1/2}$$

— $p=\infty$: max entry of A in magnitude

— Calculate the Frobenius norm of $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \in \mathbb{R}^{2\times 2}$

$$\|A\|_F = \left(\sum_{j=1}^{2}\sum_{i=1}^{2} |a_{ij}|^2\right)^{1/2} = \left(1^2+2^2+3^2+4^2\right)^{1/2} = \sqrt{30}$$

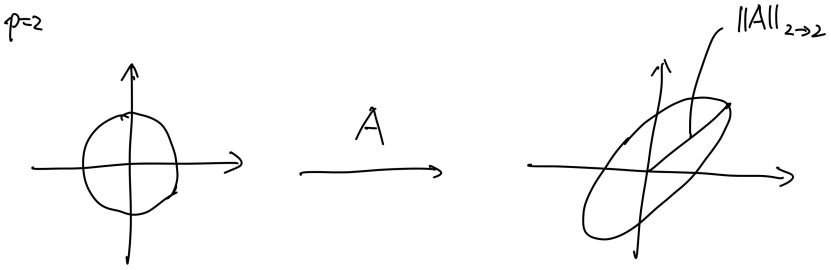
- $A \in \mathbb{R}^{m\times n}$ is also a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$

Define p-norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively.

We can define matrix p-norm of a matrix by

$$\|A\|_{p\to p} = \sup_{\substack{x\in\mathbb{R}^n \\ x\neq 0}} \frac{\|Ax\|_p}{\|x\|_p} \quad \text{(Check it is a norm!)}$$

$$= \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|_p=1}} \|Ax\|_p$$



- Find the vector 2-norm (Frobenius norm) and the matrix 2-norm of $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

所以如果A是一個Data Matrix
咁$\|A\|_F$ makes more sense
如果係linear transformation
咁$\|A\|_2$ makes more sense

Frobenius norm: $\|A\|_F = \left(1^2+0^2+0^2+1^2\right)^{1/2} = \sqrt{2}$

matrix 2-norm: $\|A\|_{2\to 2} = \sup_{\substack{x\in\mathbb{R}^2 \\ x\neq 0}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{\substack{x\in\mathbb{R}^2 \\ x\neq 0}} \frac{\|x\|_2}{\|x\|_2} = 1$

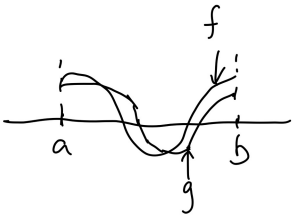
Example 4: $C[a,b]$ is a vector space over $\mathbb{R}$.

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

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

Example 5: • ℓ_∞ is a vector space over $\mathbb{R}$

$\forall\ a \in \ell_\infty$, define $\|a\|_\infty = \sup_i |a_i|$

It is a norm on ℓ_∞, called ℓ_∞-norm.

• Similarly, for any infinite sequence $a = (a_i)_{i\in\mathbb{Z}}$

we define $\|a\|_p = \left(\sum_{i=1}^{\infty} |a_i|^p\right)^{1/p}$, where $p \geq 1$.

Consider the space $\ell_p = \{ a \mid \|a\|_p < +\infty \} \subset \ell_\infty$

We can show that ℓ_p is a vector space
and $\|\cdot\|_p$ is a norm on ℓ_p, call ℓ_p-norm.

• Example: $a = \begin{pmatrix} 1 \\ 1/2 \\ 1/3 \\ \vdots \end{pmatrix}$

Then $\|a\|_\infty = \sup_i |a_i| = 1$ 呢個係從sin(x) taylor series推的

$$\|a\|_2 = \left(\sum_{i=1}^{\infty} \left(\frac{1}{i}\right)^2\right)^{1/2} = \left(\frac{\pi^2}{6}\right)^{1/2} = \frac{\pi}{\sqrt{6}}$$

$$\|a\|_1 = \left(\sum_{i=1}^{\infty} \frac{1}{i}\right) = +\infty$$

Therefore, $a \in l_\infty$, $a \in l_2$, but $a \notin l_1$

Remarks: 1. For the same vector space, we can define infinitely many different norms.

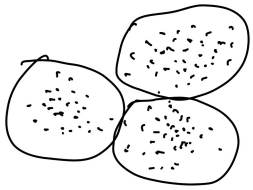
2. A common technique in machine learning to find the solution is to optimize some norm of the unknown. Different norms lead to very different solutions.

§ 1.2 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.
  - Recommandation 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 applicat

Mathematical formulation:

c係每個point屬於咩group
G係每個group有咩point

- Representation:

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

Let $G_j \subset \{1, 2, \cdots, N\}$ be the data points of group j. $j = 1, 2, \cdots, k$.

Then $\boxed{G_j = \{ i \mid C_i = j \}}$ and $\boxed{C_i = j \text{ for all } i \in G_j}$

Then, to find the grouping, we only need to find C_i, $i = 1, \cdots, k$ or G_j, $j = 1, \cdots, N$.

Let $C_i \in \{1, 2, \cdots, k\}$ be the

Then, group j, denoted by

We assign each group a

the representative vectors are not necessarily one of given vectors.

Assign representative
姐係可能每group assign中間個個點做representative vector咁，Ci,Gj,Zi都係representation黎

- Evaluation:

白話：同一group嘅vector應該close啲

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.

J_j 係total closeness for group j，但我地通常唔想用呢舊野因為用2-norm有可能會令呢舊野undifferentiable(not smooth),所以我地take square

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

姐係咁多group一齊都要最細

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

min擺底嘅野代表呢啲野係adjustable嘅可以當係given(= parameter)

- 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

fix咗一個然後計第二個

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

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

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

$$\Updownarrow$$

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

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.,

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

In other words,

> x_i is assigned to the group whose representative vector is the closest to x_i.

For (2): It is rewritten as

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

Obviously, it is equivalent to minimize each term independently,

Gj係disjoint嘅，淨係出現一次

$$(1) \Leftrightarrow \min_{c_1,\cdots,c_N} \sum_{j=1}^{K} \sum_{i \in G_j} \|x_i - z_j\|_2^2$$

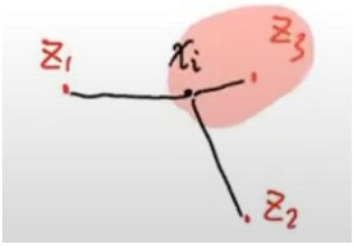
$$\Leftrightarrow \min_{c_1,\cdots,c_N} \left(\|x_1 - z_{c_1}\|_2^2 + \|x_2 - z_{c_2}\|_2^2 + \cdots + \|x_N - z_{c_N}\|_2^2 \right)$$

(其實呢到就係kmean個定義黎z_{c_1}係指x_1呢點所屬嘅group嘅representative,
每個點同佢group相減得出total distance

$$\Leftrightarrow \min_{c_i \in \{1,\dots,k\}} \|x_i - z_{c_i}\|_2^2 \quad , \quad 1 \le i \le N$$

$$\Leftrightarrow \quad c_i = \arg\min_{j=1,\dots k} \|x_i - z_j\|_2^2$$

另一種寫法，咁樣loop唔洗compare all point,只係compare all z_j



例子如左圖： x_i 穩佢同邊個z_j最近即可

Then we set $G_j = \{ j \mid C_i = j \}$, $j = 1,\dots,k$

Step 2 的推導:

係交換summation嘅次序，將j=1tok sub落去

$$(2) \Leftrightarrow \min_{z_1,\dots,z_K} \sum_{j=1}^{k} \sum_{i \in G_j} \|x_i - z_j\|_2^2$$

$$\Leftrightarrow \min_{z_1,\dots,z_K} \left(\sum_{i \in G_1} \|x_i - z_1\|_2^2 + \sum_{i \in G_2} \|x_i - z_2\|_2^2 + \cdots + \sum_{i \in G_K} \|x_i - z_K\|_2^2 \right)$$

$$\Leftrightarrow \min_{z_j} \sum_{i \in G_j} \|x_i - z_j\|_2^2 , \qquad j = 1, \dots, K$$

係觀察到z1..zk都係獨立事件
可以抽出黎sum

例子：

呢三個應該要同一group，
要穩到 z_j s.t. $\|x_i - z_j\|_2^2$ is minimized

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} are l-th component of x_i
z_{jl} z_j respectively

For l-th ... Summation are independent ...

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)

Each term is this summation are independent again.

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

↑ 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,

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

In other words,

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

↑ centroid of △.

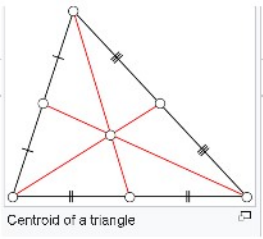
Altogether, we get the following clustering algorithm.

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

↑ Centroid =



Centroid of a triangle

from $x_1, x_2, \cdots, x_N$ randomly

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

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

and define

人說:

人話:

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

x_i is assigned to the group whose representative vector is the closest to x_i.

and define

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

C_i=j such that z_j is the closet representative to x_i

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

z_j is the mean of all vectors in G_j

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

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

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