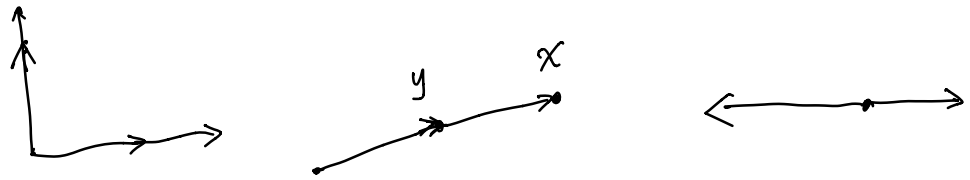
Ch. 3. Inner product, Hilbert space

§ 3.1. Inner Product

Norms give only metrics, which are "scaling" sensitive.

However, in many applications, "scaling" does not 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 "angle" of two vectors, which is "scaling" insensitive.



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 $\forall x \in V$, $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$

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, $\langle x, \alpha y_1 + \beta y_2 \rangle = \alpha \langle x, y_1 \rangle + \beta \langle x, y_2 \rangle$ $\forall \alpha, \beta \in \mathbb{R}$, $x, y_1, y_2 \in V$.
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

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

$(\equiv x^Ty)$

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 $x^TAx > 0 \ \forall\ x \neq 0$)

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

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

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

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

③ $\langle x, y\rangle_A = x^TAy = (x^TAy)^T = y^TA^Tx = y^TAx = \langle y, x\rangle_A$.

For the same vector space, we can define infinitely many inner products.

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

← Define

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

$$= \text{trace}(A^TB)$$

$$= \text{trace}(B^TA)$$

This is an inner product on $\mathbb{R}^{m\times n}$.

Example 4: For two infinite sequences

$a = (a_i)_{i=1,2,3,\cdots}$

$b = (b_i)_{i=1,2,3,\cdots}$

Then $\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$ defines an inner product.

Example 5: In $\mathbf{C}[a,b]$, we can define an inner product as

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

§ 3.2. Properties of inner product, orthogonality

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 $x = \alpha y$ or $y = \alpha x$ for some $\alpha \in \mathbb{R}$.

proof • if $y = 0$. Then obviously

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

• It remains to prove the inequality with $y \neq 0$.

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

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

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

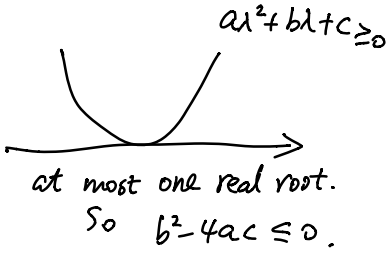
Thus, $\lambda^2 \langle y, y \rangle + 2\lambda \langle x, y \rangle + \langle x, x \rangle \geq 0$. $\forall \lambda \in \mathbb{R}$.

$\equiv f(\lambda)$

Since $y \neq 0$, $\langle y, y \rangle > 0$.

So, $f(\lambda)$ a quadratic function of λ that takes non-negative values only.

There is at most one root of the quadratic function.



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

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

• Next when the equality holds true, i.e., $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$,

  • if $y = 0$, then obviously $y = \alpha x$, where $\alpha = 0$.

  • if $y \neq 0$, there is exacty one real root of $f(\lambda)$.

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

$\Rightarrow \langle x + \lambda y, x + \lambda y \rangle = 0 \Rightarrow x + \lambda y = 0 \Rightarrow x = \alpha y$ with $\alpha = -\lambda$.

Finally, we show that if $x = \alpha y$ or $y = \alpha y$, the $\langle x, y \rangle^2 = \langle x, x \rangle \langle y, y \rangle$.

By direct calculation:

$x=\alpha y \Rightarrow \quad \langle x,y\rangle^2 = \langle \alpha y, y\rangle^2 = \alpha^2 \langle y,y\rangle = \langle \alpha y, \alpha y\rangle \langle y,y\rangle$

$y \Rightarrow \alpha x \Rightarrow \quad \langle x,y\rangle^2 = \langle x, \alpha x\rangle^2 = \alpha^2 \langle x,x\rangle = \langle \alpha x, x\rangle \langle \alpha x, \alpha x\rangle$ ⊠

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

$\|x\| = \langle x,x\rangle^{1/2}$ 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 $\|x+y\|_2^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle$

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

$\leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|$ (Note that Cacuchy-Schwartz becomes $|\langle x,y\rangle| \leq \|x\|\|y\|$)

$= (\|x\| + \|y\|)^2$ ⊠

We call the norm $\|x\| = \langle x,x\rangle^{1/2}$ the norm induced by inner product $\langle \cdot, \cdot \rangle$.
This will be considered as a default norm in an inner product space.

Cauchy-Schwartz is restated as:

$$|\langle x,y\rangle| \leq \|x\| \, \|y\|$$

Example 1: $\mathbb{R}^n$ with inner product $\langle x,y\rangle = x^T y$.

The induced norm is

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

Example 2: $\mathbb{R}^n$ with inner product $\langle x,y\rangle_A = x^T A y$, where A is SPD.

The induced norm is

$$\|x\|_A = \langle x,x\rangle_A^{1/2} = (x^T A x)^{1/2} = \left(\sum_{i,j} a_{ij} x_i x_j\right)^{1/2}$$

Example 3: The p-norm in $\mathbb{R}^n$,

$\|x\|_p$, $p \neq 2$, are not induced by inner products.

Example 4: $\mathbb{R}^{m\times n}$ with inner product

$$\langle A, B\rangle = \sum_{i,j} a_{ij} b_{ij}$$

The induced norm is

$$\|A\| = \langle A, A\rangle^{1/2} = \left(\sum_{i,j} a_{ij}^2\right)^{1/2} \equiv \|A\|_F - \text{the Frobenius norm}.$$

"Angles" in inner product spaces.

The equality "=" is attained if and only if x and y are aligned exactly, which should have the least angle. 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 angles 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

$$\cos \angle(x,y) = \frac{\langle x, y\rangle}{\|x\| \|y\|}$$

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

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.

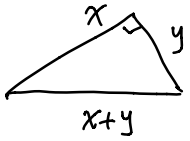
- When the angle of x and y is $\frac{\pi}{2}$, we call they are orthogonal, denoted by $x \perp y$, i.e.,

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

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



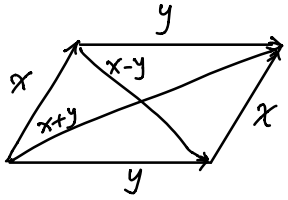
proof. $\|x+y\|^2 = \langle x+y, x+y\rangle$

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

$= \|x\|^2 + \|y\|^2$ ⊠

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

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



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

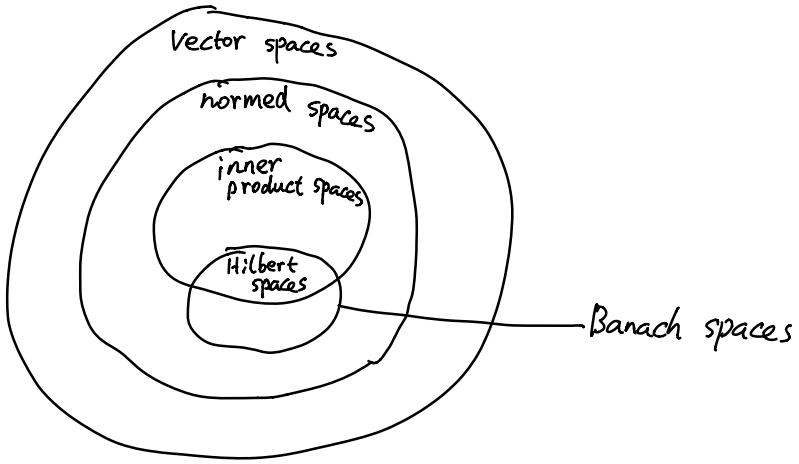
$= \langle x+y, x+y\rangle + \langle x-y, x-y\rangle$

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

$= 2\|x\|^2 + 2\|y\|^2$ ⊠

Hilbert space:

A Hilbert space is a Banach space in which an inner product is equipped and the norm is induced by the 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

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

Example 3: $\mathbb{R}^{m \times n}$ with inner product $\langle A, B \rangle = \text{trace}(A^T B)$ is a Hilbert space.

Example 4: $l_2 = \{ a \mid a \text{ is an infinite sequence and } \|a\|_2 < +\infty \}$ with inner product $\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$ is a Hilbert space.

Example 5: $C[a,b]$ with inner product

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

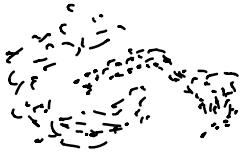
is NOT a Hilbert space, because it is not complete.

To complete $C[a,b]$ under the norm $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$,
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.

§ 3.3 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

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

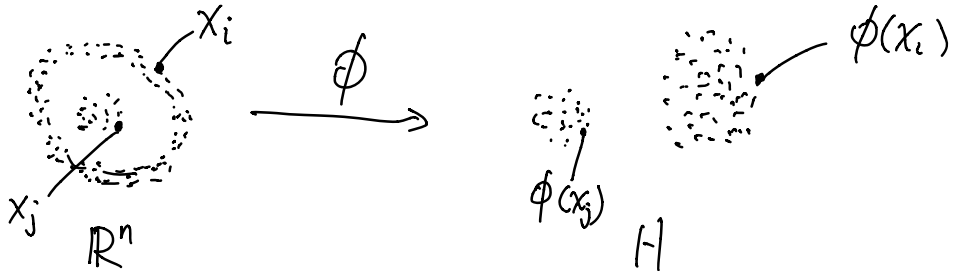
② Then $G_j = \{i \mid c_i = j\}$, for $j = 1, 2, \cdots, k$.

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.

① $C_i = \arg\min_{j \in \{1,2,\cdots,k\}} \left\| \phi(x_i) - \frac{1}{|G_j|} \sum_{l \in G_j} \phi(x_l) \right\|^2, \quad i = 1, 2, \cdots, N.$

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

- 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

Only inner products in the feature space are involved.

Therefore,

An explicit expression of ϕ 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,

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

$$= \sum_{i=1}^m \sum_{j=1}^m c_i c_j \langle \phi(y_i), \phi(y_j) \rangle \quad \text{(By inner product property.)}$$

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

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

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

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

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.

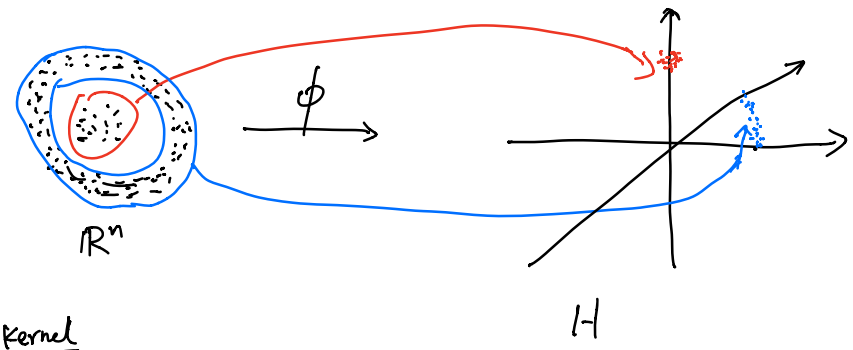
Example:

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.

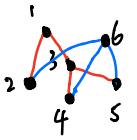
§ 3.4. Case Study: Metric Learning

- Given a set of vectors $x_1, x_2, \cdots, x_N \in \mathbb{R}^n$, and given information that certain pairs of them are similar / dissimilar

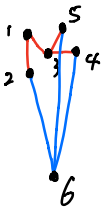
  S: $(x_i, x_j) \in S$ if x_i and x_j are similar.

  D: $(x_i, x_j) \in D$ if x_i and x_j are dissimilar.

  Can we find a distance metric such that "similar" points are close to each other and "dissimilar" points are far away from each other ?



re-organize them by a "metric"



— similar

— dissimilar

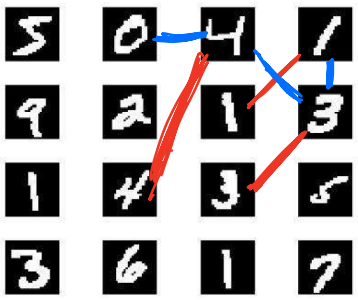
blank unknown (e.g., (x_2, x_3), (x_3, x_6))

So, the similarities are measuared by the distance. Then we can infer that, e.g., (x_2, x_3) are similar (x_3, x_6) are dissimilar.

- Potential applications

  — "Supervised" clustering.

  - Given a data set with partial clustering information



— Similar (i.e., in the same cluster)

— dissimilar (i.e., in different clusters)

  - Metric learning (i.e., find a distance)
  - cluster the points under the learned distance metric.

Representation:

There are many different forms of norms, even in $\mathbb{R}^n$.

Candidate forms of norms

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

- This class of norms is too small, because there is only one parameter (p) in the definition of these norms.

— norms induced by the A-inner product

$$\|x\|_A = \left(\langle x, x\rangle_A\right)^{1/2} = \left(x^T A x\right)^{1/2},$$

where $A \in \mathbb{R}^{n\times n}$ is an SPD matrix.

- This class of norms are parametrized by an $n \times n$ matrix A that contains enough parameters (n^2 parameters)
- We use this class of norms as candidate norms for our distance metric learning.
- Then

find a distance metric $\Longleftrightarrow$ Find an SPD matrix $A \in \mathbb{R}^{n\times n}$ and the distance metric is $\|\cdot\|_A$

Remarks:

① The SPD-ness of A is usually relaxed to A is symmetric positive semi-definite (SPSD). Correspondingly, the resulting norm will NOT satisfy $\|x\|_A = 0 \Leftrightarrow x = 0$, and other conditions in the definition of norms are satisfied; thus, $\|\cdot\|_A$ with SPSD A is NOT exactly a norm, but a pseudo-norm.

② Given $x, y \in \mathbb{R}^n$, the disance induced by $\|\cdot\|_A$

$$\|x-y\|_A = \left((x-y)^T A (x-y)\right)^{1/2}$$

is also known as Mahalanobis distance.

Evaluation: Which A is the best for metric learning?

For $(x_i, x_j) \in S$, their distance is small, i.e,

$$\sum_{(x_i,x_j)\in S} \|x_i - x_j\|_A^2 \quad \text{is small}$$

For $(x_i, x_j) \in D$, their distance is large, i.e,

$$\sum_{(x_i,x_j)\in D} \|x_i - x_j\|_A^2 \quad \text{is large}$$

So, we find the best $A \in \mathbb{R}^{n\times n}$ is the sense of

$$\min_{\substack{A\in\mathbb{R}^{n\times n}\\ \text{is SPSD}}} \frac{\sum_{(x_i,x_j)\in S} \|x_i - x_j\|_A^2}{\sum_{(x_i,x_j)\in D} \|x_i - x_j\|_A^2}$$

Or, equivalently, we solve

$$\begin{cases} \min\limits_{\substack{A\in\mathbb{R}^{n\times n}\\ \text{is SPSD}}} \sum\limits_{(x_i,x_j)\in S} \|x_i - x_j\|_A^2 \\ \text{s.t.} \sum\limits_{(x_i,x_j)\in D} \|x_i - x_j\|_A^2 \geq 1 \end{cases}$$

Optimization:

The optimization is introduced later.