Final Review - ch3

Sunday, 6 December 2020 10:23 AM

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 $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

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

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

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

Example: R^n inner product(最傳統的x^Ty)

Exmaple 2:$\langle x,y \rangle_A$

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

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

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

Example 3: $\langle A,B \rangle$ for A,B in $R^{m \times n}$

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

- Define

$$\langle A, B \rangle \stackrel{def}{=} \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}$$

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

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

Other examples:

$\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$. Where a,b = infinite sequences

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

3.2 Properties of inner product, orthogonality

Cauchy-Schwartz Ineqality:

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

Pf8 : proof of cauchy schwartz inequality

Properties:

if $x=\alpha y$ or $y=\alpha y$, the $\langle x,y\rangle^2 = \langle x,x\rangle\langle y,y\rangle$.

Pf9 : conditions for <x,y>^2 = <x,x><y,y>

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

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

Pf10: sqrt(<x,x>) defines a norm

$\boxed{\begin{array}{l}\text{We call the norm } \|x\| \equiv \langle x, x\rangle^{1/2} \\ \qquad \text{the norm induced by inner product } \langle \cdot, \cdot \rangle. \\ \text{This will be considered as a default norm in an inner product space.}\end{array}}$

Examples:

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

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

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

$\langle A, B\rangle = \sum_{i,j} a_{ij} b_{ij}$ | $\|A\| = \langle A, A\rangle^{1/2} = \left(\sum_{i,j} a_{ij}^2\right)^{1/2} \equiv \|A\|_F$

We define cos to quantitize the closeness to exact alignment

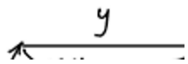
$$\cos \angle(x,y) = \frac{\langle x, y\rangle}{\|x\| \|y\|}$$

Orthonality:

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

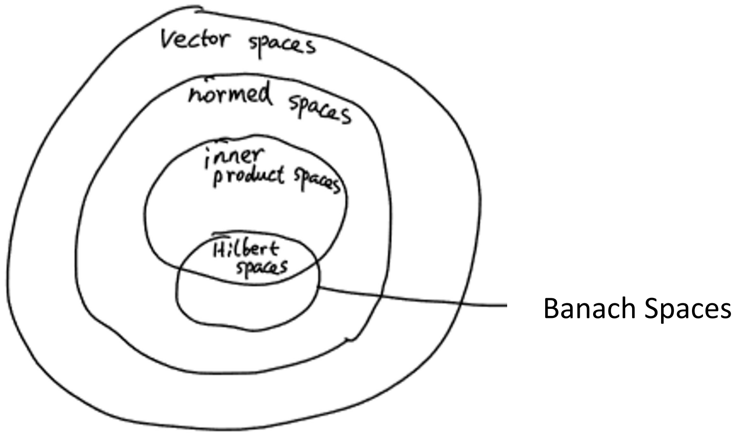
Pythagoras' theorem: Let x, y be two vectors in an inner product space.
If $x \perp y$, (if & only if) then $\|x+y\|^2 = \|x\|^2 + \|y\|^2$

Parallelogram law: $\forall\, x, y \in H$ — an inner product space,
$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$$



A Hilbert space is a Banach space(complete norm space) in which

1. an inner product is equipped
2. the norm is induced by the inner product



Examples for hilbert space:

1. R^n with $\langle x,y\rangle = x^T y$ is hilbert space
2. R^n with $\langle x,y\rangle_A = x^T A y$, A is SPD matrix, is hilbert space, norm = $||x||_A = (x^T A x)^{1/2}$
3. $R^{m \times n}$ with $\langle A,B\rangle = \text{trace}(A^T B)$ is hilbert space
4. L_2 = {a|a is an infinite sequence and $||a||_2 < +\text{inf}$} with $\langle a,b\rangle$ = sum_i=(1 to inf) $(a_i b_i)$ is hilbert space(Pf 11: l2 space is a hilbert space)

Counter Examples:

5. C[a,b] with inner product $\langle f,g\rangle = \int_a^b f(x)g(x)\,dx$
   Is not hilbert space, because only inf norm in c[a,b] is complete, and so sqrt(<f,g>) is not complete as well

Section 3.3: Case Study: Kernel Trick, Kernel k-means

將x 變成 $\phi(x)$，將其 map 至 hilbert space.

Algorithm:

Kernel k-means algorithm

- Choose a kernel function $K(\cdot,\cdot)$

- 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

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

- go back and repeat

Pf12: 改寫k-mean做<phi(x_i), phi(x_j)>版本,map佢去feature space

Kernel function的必要條件

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

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 $\mathcal{H}$.

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

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

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

Section 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

$$S: \quad (x_i, x_j) \in S \text{ if } x_i \text{ and } x_j \text{ are similar.}$$

$$D: \quad (x_i, x_j) \in D \text{ if } x_i \text{ and } x_j \text{ are dissimilar.}$$

給定一些vector,要用一個metric衡量方法讓相似的放同一個set, 不相似的放另一個set

這裏教的是<x,y>$_A$,因爲佢parameter多(有n^2)

- Then

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

通常SPD都會relax到SPSD, 雖然x = 0的時候||x||$_A$未必是0，所以SPSD A不是真的norm

example

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.

運算化簡後得到objective function

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