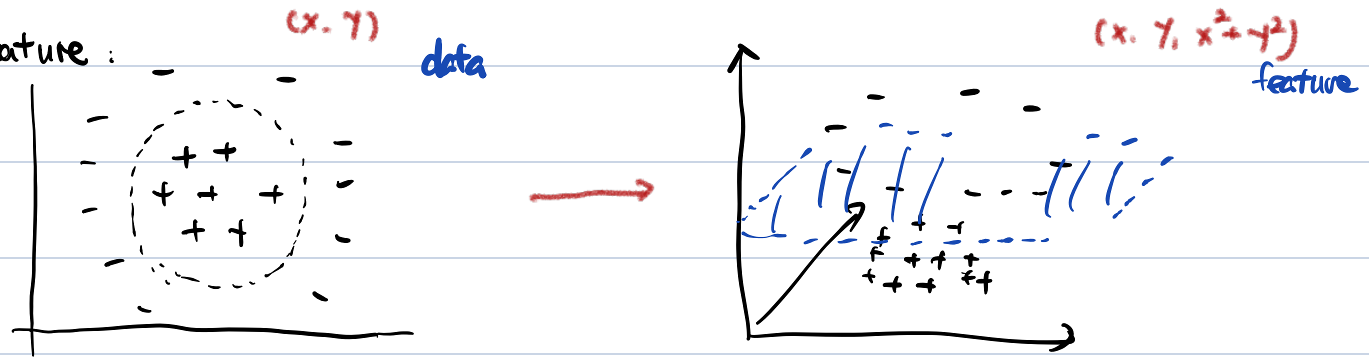
Reproducing Kernel Hilbert Space

§ Feature :



linear Tokyo − Japan + U.S = D.C.

① Construct My feature by hand.

② Save all my feature. (If ∞ feature, the computation is hard)

− Primal: size # data. Dual: size # feature.

§ Revisiting Ridge Regression From Dual Side.

$$\min_w \|Xw - y\|_2^2 + \lambda \|w\|^2.$$

⇒ introduce $z = Xw$ "Operator Splitting"

$$\min_w \|z - y\|_2^2 + \lambda \|w\|_2^2$$

$$\text{s.t.} \quad z = Xw$$

$$L(w, z, \alpha) = \|z - y\|^2 + \lambda\|w\|_2^2 + \alpha^T(z - Xw)$$

− $\nabla_w L = 0$ $\underbrace{2}_{\text{Constant}}\lambda w = X^T\alpha$

$$\Rightarrow w = X^T\alpha \quad (\Delta)$$

put (Δ) back to the objective. $\min \|XX^T\alpha - y\|_2^2 + \lambda\|X^T\alpha\|^2$

$$\Rightarrow \alpha = (XX^T + \lambda I)^{-1} y$$

$$\min_w \|Xw - y\|_2^2 + \lambda \|w\|^2$$

<u>Dual</u>: $w = X^T \alpha$. where. $\alpha = (\underbrace{XX^T}_{\#data} + \lambda I)^{-1} y$

#data ·

<u>Primal</u>: $w = (\underbrace{X^T X}_{\#feature} + \lambda I)^{-1} X^T y$

#feature

§ Revisit of Polynomial Ression,

$$Y = \alpha_1 X^2 + \alpha_2 X + \alpha_3 = \underbrace{(X^2, X, 1)}_{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$$

This is my new data

⇒ we can use linear regression on data $(X^2, X, 1)$ to do poly regress

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = (X^T X + \lambda I)^{-1} X^T y$$ primal

$$= X^T \underline{(XX^T + \lambda I)}^{-1} y$$ dual

$$\begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & & \\ x_n^2 & x_n & 1 \end{pmatrix} \begin{pmatrix} x_1^2 & x_2^2 & \cdots & x_n^2 \\ x_1 & x_2 & & x_n \\ 1 & 1 & & 1 \end{pmatrix}$$

$$(XX^T)_{i,j} = \langle (x_i^2, x_i, 1), (x_j^2, x_j, 1) \rangle$$

"inner product of feature"

"How similar are the two features"

Kernel
$$\begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & \\ \vdots & & & \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{pmatrix}$$

how the data x_i are similar to x_j

Idea. We can't compute the $w = X^T \underline{(XX^T + \lambda I)^{-1} y}$

I still needs to know the feature.

This can be compute

if I have a test data (x_{test})

$$\langle w, x_{test} \rangle = x_{test} X^T (XX^T + \lambda I)^{-1} y$$

$$(k(x_{test}, x_1), k(x_{test}, x_2) \cdots k(x_{test}, x_n))$$

- privacy: $x_1 \cdots x_n$ are used in testing
- $(XX^T + \lambda I)^{-1}$ is costly in large data sets.
- It is hard for online learning.

---

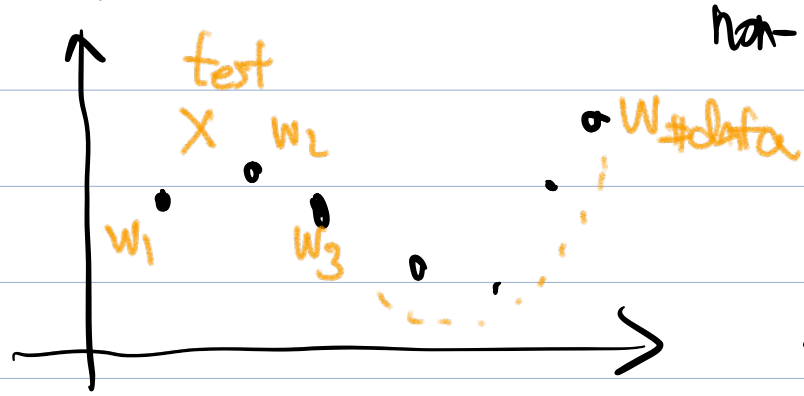
linear Behavior of RKHS

$$\langle w, x_{test} \rangle = \underbrace{x_{test} X^T}_{\text{a } \mathbb{R}^{\# data} \text{ vector}} \underbrace{(XX^T + \lambda I)^{-1}}_{\text{only depend on } X, \text{ but not depend on } y} y$$

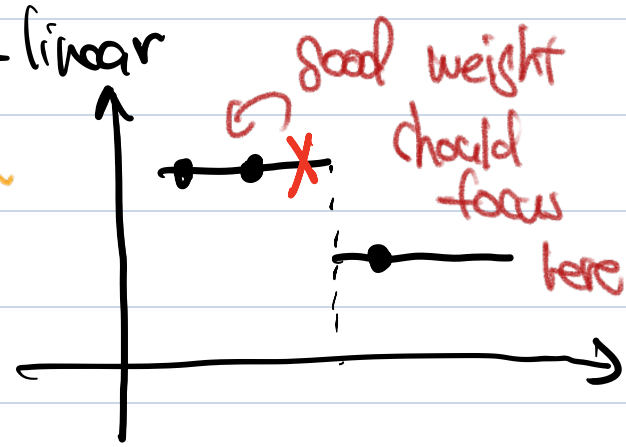
is a reweighting of y

Remark.

linear



non-linear



non-parametric Estimator (?)

" Ridge Regression"   " Lasso "

① Why NN is better than Kernel

↳ adapt to jump (one of the answer)

---

① $W = X^T \alpha = \alpha_1 X_1 + \cdots + \alpha_n X_n$

what does data here mean.

$f(x) = \langle w, x \rangle$

X (data) actually means the mapping from function → function value.

Reproducing: $f \in H$, there always exists a (bounded) linear mapping K_x, such that $\langle f, K_x \rangle = f(x)$

$f \in L_\infty$, "RKHS is not a large space".

$\frac{2s}{d+2s}$

↳ RKHS in $\mathbb{R}^d$, always have $d/2$ smooth f

Hilbert Space: linear subspace with inner product.

$k(x, y) = \langle k_x, k_y \rangle$

$k = XX^T \Rightarrow k$ should be positive definite ✓

<u>Question</u>. can $k(x, y)$ be arbitrary?

<u>Example</u>.

① $k(x, y) = x^T y$

② k_1 is a kernel, k_2 is a kernel. then, $k_1 + k_2$ is a kernel

$\phi_1(x)$ is a feature, $\phi_2(x)$ is a feature. $[\phi_1(x), \phi_2(x)]$ is also feature

③ $k_1 \cdot k_2$ is also a kernel. $\phi_1 \cdot \phi_2$ is a feature

④ $\exp(k(x, y))$ is a kernel.

by ②, ③, $k(x, y)$ is a kernel, then

$f(k(x, y))$ is a kernel, if f is a polynomial with positive coef.

$$\exp(x) = \lim_{i \to \infty} \left(1 + x + \cdots + \frac{x^i}{i!}\right)$$

⑤ $k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right)$ is a kernel.

$$= \exp\left(-\frac{\|x\|^2}{\sigma^2}\right) \exp\left(-\frac{\|y\|^2}{\sigma^2}\right) \exp\left(\frac{2x \cdot y}{\sigma^2}\right)$$

$\Downarrow$

rank-1 $\leftarrow$ $k(u, v) = g(u) g(v)$ is a kernel

⑥ Translation-invariant kernel.

$k(u, v) = f(u - v)$.

$$f(x) = \sum_{n=0}^{\infty} a_n \cos(nx)$$

$$f(u-v) = \sum_{n=0}^{\infty} a_n \cos\big(\underbrace{n(u-v)}\big) \quad - (\Delta)$$

rank 1

rank -1

Eigen decomposition

$\sin(nu)\sin(nv) + \cos(nu)\cos(nv)$

⇓

$\{\sin(nu) : u \geq 1\} \cup \{\cos(nu) : u \geq 1\}$ is the feature of a translation invariant kernel.

[Random Fourier feature] random sample sin and cos to construct feature map, is a good approximation to the translation-invariant kernel.

Remark. (Δ) is actually eigen decomposition.

$A u_i = \lambda_i u_i$

- If A is symmetric. $A = \sum_{i=1}^{n} \underbrace{\lambda\, u\, u^T}_{\text{rank-1 matrix}}$

↓

K #data

#data

$u(x_1)$ #data. $u(x_1)$ → $u(x_n)$

$$\underbrace{\boxed{K}}_{\#data} \#data = \sum_{i=1}^{n} \lambda_i \begin{bmatrix} u(x_1) \\ \vdots \\ u(x_n) \end{bmatrix} \#data$$

⇓

$$\begin{bmatrix} u(x_1)u(x_1) & u(x_1)u(x_2) & \cdots \\ u(x_2)u(x_1) & u(x_2)u(x_2) & \cdots \\ \vdots & & \\ u(x_n)u(x_1) & u(x_n)u(x_1) & \cdots \end{bmatrix}$$

Mercer's theorem.

$$k(x, y) = \sum_{i=1}^{n} \lambda_i \phi(x_i) \cdot \phi(x_j)$$

$\Downarrow$

is the eigens of kernel matrix

$\Rightarrow$ $k(x, y)\, \phi(y) = \lambda \phi$

$\Downarrow$

$$\int k(x, x')\, \phi(x') = \lambda \phi(x')$$