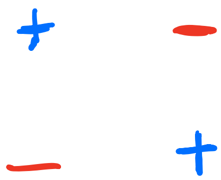
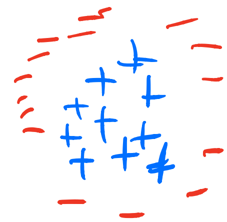


XOR



linear predictor $w^T x$

$w^T \phi(x)$

$\uparrow$ feature expansion.

Feature Expansion.

① $x \in \mathbb{R}^d$ "Quadratic Expansion"

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \dots, \sqrt{2}x_d, x_1^2, \dots, x_d^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3 \dots \dots, \sqrt{2}x_{d-1}x_{d-2})$$

$$\phi(x)^T \phi(x') = (1 + x^T x')^2$$

better than $O(d^2)$? $O(d)$ linear time.

② "Products of all subsets"

$$\phi(x) = \left(\prod_{i \in S} x_i \right)_{S \subseteq [d]}$$

$$(x_1, x_2 \dots x_d, x_1x_2 \dots x_{d-1}x_d \dots \dots, x_1 \cdots x_d) \in \mathbb{R}^{2^d}$$

$$\phi(x)^T \phi(x) = \underbrace{\prod_{i=1}^{d} (1+x_i x_i')}_{O(d) \text{ time}}$$

How hard?

$O(2^d)$

(3)

$$\phi(x)^T \phi(x') = \underbrace{\exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)}_{O(d) \text{ linear}}$$

"Gaussian Kernel"

$\phi(x) \in \mathbb{R}^\infty$

In $x \in \mathbb{R}$, $\phi(x) = \exp\left(\frac{-x^2}{2\sigma^2}\right)\left(1, \frac{x}{\sigma}, \frac{1}{2!}\left(\frac{x}{\sigma}\right)^2, \cdots\right)$

Kernel

A kernel function $K: \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$ is a symmetric function such that for any $x_1, \dots, x_n \in \mathcal{X}$, the Gram matrix G

G_{ij} $\qquad G_{ij} = K(x_i, x_j)$

is positive semidefinite.

$\longrightarrow \boxed{q^T G q \geq 0,} \quad \forall q \in \mathbb{R}^n.$

How to show $K(\cdot,\cdot)$ is a kernel?

Easy way: Find ϕ mapping such that

$$\underbrace{\phi(x)}_{\text{new features.}}{}^T \phi(x') = K(x, x').$$

"similarity measure"

$\hookrightarrow$ implies the definition

$$G_{ij} = \phi(x_i)^T \phi(x_j)$$

$$G = \underbrace{\Phi^T \Phi}_{p.s.d?} , \quad \Phi = [\phi(x_1), \dots \phi(x_n)]$$

$$\forall q, \quad q^T \Phi^T \Phi q = (\Phi q)^2 \geq 0.$$

Examples:

① $K(x, x') = \underbrace{x^T x'}$

when is this useful?

$n << d$

② Polynomial: $K(x, x') = (1 + x^T x')^k$

③ Gaussian Kernel/
Radial Basis Function (RBF)

$$K(x, x') = \exp\left(\frac{-\|x - x'\|^2}{2\sigma^2}\right)$$

- - - - . . .

other examples in the note.

Dual SVM

$$\max_{\alpha, \lambda} \sum_i \lambda_i - \frac{1}{2} \sum_{i,j \in [n]} \lambda_i \lambda_j y_i y_j \underbrace{\phi^T(x_i) \phi(x_j)}_{K(x_i, x_j)}$$

$$\text{s.t.} \quad 0 \leq \lambda_i \leq C.$$

For RBF, w lives in $\mathbb{R}^\infty$?

$$w^* = \sum_{i=1}^{n} y_i \lambda_i^* \phi(x_i)$$

For a new test point x.

$$w^{*T} \phi(x) = \sum_{i=1}^{n} y_i \lambda_i^* \underbrace{\phi(x_i)^T \phi(x)}_{K(x_i, x)}$$

Kernel $\longleftrightarrow$ "similarity"

Build new kernels from old kernels.

$$\underbrace{K(x,y)}_{\text{Kernel.}} = c \underbrace{K_1(x,y)}_{\text{Kernel}}, \quad \text{for } c > 0.$$

HW 2

$$\underbrace{K(x,y)}_{\text{Kernel.}} = \underbrace{K_1(x,y) + K_2(x,y)}_{\text{Kernels}}$$

$$K(x,y) = K_1(x,y) K_2(x,y)$$

other examples in notes.

Kernel Ridge Regression

$$A = \begin{bmatrix} \leftarrow x_1 \rightarrow \\ \vdots \\ \leftarrow x_n \rightarrow \end{bmatrix} \qquad b = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Ridge Reg : $\hat{w} = (A^T A + \lambda I_d)^{-1} A^T b$

Alternative form : $(A^T A + \lambda I)^{-1} A^T$ (See note for a proof)

$$= A^T (G + \lambda I_n)^{-1}$$

$G \in \mathbb{R}^{n \times n}$, $G_{ij} = x_i^T x_j$

|

$$G_{ij} = \phi(x_i)^T \phi(x_j)$$

$$\hat{w} = A^T \underbrace{(\overset{n\times n}{\underline{G}} + \lambda I_n)^{-1} \overset{n\times 1}{b}}_{v}$$

$$\hat{w} = A^T v = \sum_{i=1}^{n} v_i x_i$$

New test point x

prediction $x^T \hat{w} = \sum_{i=1}^{n} v_i \underbrace{x^T x_i}_{K(x, x_i)}$

$$x^T \rightarrow \phi(x)^T$$

$$\phi(x)^T \hat{w} = \sum_{i=1}^{n} v_i \underbrace{\phi(x)^T \phi(x_i)}_{K(x, x_i)}$$