INNER PRODUCT KERNELS & THE KERNEL TRICK

$\boxed{\text{Issues With Nonlinear Feature Maps}}$

Suppose we transform our data via

$$x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} \longmapsto \Phi(x) = \begin{bmatrix} \varphi^{(1)}(x) \\ \vdots \\ \varphi^{(m)}(x) \end{bmatrix}, \quad m \gg d$$

where $\varphi^{(j)}$ are nonlinear.

We just discussed how this can lead to ill-conditioned problems, which can be mitigated via regularization.

In addition, since $m \gg d$, there can be an

(A) increased ______________ burden.

Example In ridge regression, we would need to invert an $m \times m$ matrix, or at least solve a linear system with m unknowns.

Fortunately, the following two facts offer a solution:

- Many machine learning algorithms only involve
(B). the data through ________ ________.

- For certain feature maps Φ, the function

$$k(x, x') :=$$

has a simple, closed form expression that can be evaluated without explicitly calculating $\Phi(x)$.

Example | Consider the case when $d = 2$ and

$$k(u, v) := (u^T v)^2$$

$$= \left(\begin{bmatrix} u^{(1)} & u^{(2)} \end{bmatrix} \begin{bmatrix} v^{(1)} \\ v^{(2)} \end{bmatrix} \right)^2$$

©

$$= \left(\sum_{i=1}^{2} \qquad \right)^2$$

$$= \left(\sum_{i=1}^{2} \qquad \right) \left(\sum_{j=1}^{2} \qquad \right)$$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2}$$

$$=$$

What feature mapping Φ does this correspond to?

$$\longrightarrow \quad \Phi = \begin{bmatrix} (u^{(1)})^2 \\ \sqrt{2}\, u^{(1)} u^{(2)} \\ (u^{(2)})^2 \end{bmatrix}$$

Now suppose d is arbitrary, and

$$\begin{aligned} k(u, v) &= (u^T v)^2 \\ &= \left(\sum_{i=1}^{d} u^{(i)} v^{(i)} \right)^2 \\ &= \left(\sum_{i=1}^{d} u^{(i)} v^{(i)} \right) \left(\sum_{j=1}^{d} u^{(j)} v^{(j)} \right) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} u^{(i)} u^{(j)} v^{(i)} v^{(j)} \end{aligned}$$

What is the dimension of the corresponding feature space?

$$\Phi(u) = \Big[\underbrace{(u^{(1)})^2, \ldots, (u^{(d)})^2}_{d}, \underbrace{\sqrt{2}\, u^{(1)} u^{(2)}, \ldots, \sqrt{2}\, u^{(d-1)} u^{(d)}}_{\frac{d(d-1)}{2}}\Big]^T$$

<u>Exercise 1</u> Describe the feature space associated with

$$k(u, v) = (u^T v)^3$$

when $d = 2$.

Solution 1

$$k(u,v) = \left(u^{(1)}v^{(1)} + u^{(2)}v^{(2)}\right)^3$$

$$= (u^{(1)})^3 (v^{(1)})^3 + 3(u^{(1)})^2 u^{(2)} (v^{(1)})^2 v^{(2)}$$

$$+ 3\, u^{(1)} (u^{(2)})^2 v^{(1)} (v^{(2)})^2 + (u^{(2)})^3 (v^{(2)})^3$$

$$= \sum_{i=0}^{3} \binom{3}{i} (u^{(1)})^{3-i} (u^{(2)})^{i} \cdot (v^{(1)})^{3-i} (v^{(2)})^{i}$$

$$\Rightarrow \Phi(u) = \left[(u^{(1)})^3,\ \sqrt{3}(u^{(1)})^2 u^{(2)},\ \sqrt{3}\, u^{(1)} (u^{(2)})^2,\ (u^{(2)})^3\right]^T$$

More generally,

$$k(u,v) = \left(\sum_{i=1}^{d} u^{(i)} v^{(i)}\right)^p$$

$$= \sum_{\substack{(j_1,\dots,j_d) \\ \sum j_k = p}} \binom{p}{j_1 \cdots j_d} (u^{(1)})^{j_1} \cdots (u^{(d)})^{j_d} \cdot (v^{(1)})^{j_1} \cdots (v^{(d)})^{j_d}$$

$$\Rightarrow \Phi(u) = \left[\dots,\ \sqrt{\binom{p}{j_1 \cdots j_d}}\,(u^{(1)})^{j_1} \cdots (u^{(d)})^{j_d},\ \dots\right]^T$$

$\Rightarrow$ all ____________ of degree p.

Ⓓ

Definition An ________ ________ ________ E

is a mapping $k : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$

for there exists an inner product space $\mathcal{H}$

and a mapping $\Phi : \mathbb{R}^d \longrightarrow \mathcal{H}$ such that

$$k(u, v) = \langle \Phi(u), \Phi(v) \rangle$$

for all $u, v \in \mathbb{R}^d$.

Given a function $k(u, v)$, when is it an IP kernel? There are two ways to verify the property

- Mercer's theorem (we won't cover)
- PSD property

Note The $\mathcal{H}, \Phi$ associated to an IP kernel is not necessarily unique, and these two methods will give distinct constructions of $\mathcal{H}, \Phi$.

PSD Kernels

Definition We say $k(u,v)$ is a ________

______-___________ kernel if k is (F)

symmetric and, for all n and all

$x_1, \ldots, x_n \in \mathbb{R}^d$, the ________ matrix

$$\Big[\qquad \qquad \Big]_{i,j=1}^{n}$$

is positive semi-definite.

Theorem k is an IP kernel $\Longleftrightarrow$ k is a PSD kernel

Proof Schölkopf and Smola, Learning with Kernels.

Examples of IP kernels

1. Homogeneous polynomial kernel

$$k(u,v) = \langle u, v \rangle^p, \quad p = 1, 2, \ldots$$

2. Inhomogeneous polynomial kernel

$$k(u,v) = (\langle u, v \rangle + c)^p, \quad p = 1, 2, \ldots \quad c > 0$$

(G) $\Phi \rightarrow$

3. Gaussian kernel

$$k(u,v) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left\{ -\frac{\|u-v\|^2}{2\sigma^2} \right\}, \quad \sigma > 0$$

→ constant can be dropped

→ also called radial basis function (RBF) kernel

→ $\mathcal{H}$ is ________ ________!

(H)

$\boxed{\text{Key}}$ A. computational

B. inner products, $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$

C. $\left(\sum_{i=1}^{2} u^{(i)} v^{(i)} \right)^2 = \left(\sum_{i=1}^{2} u^{(i)} v^{(i)} \right) \left(\sum_{j=1}^{2} u^{(j)} v^{(j)} \right)$

$$= \sum_{i=1}^{2} \sum_{j=1}^{2} u^{(i)} u^{(j)} v^{(i)} v^{(j)}$$

$$= (u^{(1)})^2 (v^{(1)})^2 + 2 u^{(1)} u^{(2)} v^{(1)} v^{(2)} + (u^{(2)})^2 (v^{(2)})^2$$

$$= \langle \Phi(u), \Phi(v) \rangle$$

where

$$\Phi(u) = \begin{bmatrix} (u^{(1)})^2 \\ \sqrt{2}\, u^{(1)} u^{(2)} \\ (u^{(2)})^2 \end{bmatrix}$$

D. monomials E. inner producting kernel

F. positive semi-definite, gram, $[k(x_i, x_j)]_{i,j=1}^{n}$

G. all monomials of degree $\leq p$

H. infinite dimensional