INNER PRODUCT KERNELS & THE KERNEL TRICK

Issues With Nonlinear Feature Maps

Suppose we transform our data via

\[
x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} \quad \rightarrow \quad \Psi(x) = \begin{bmatrix} \phi^{(1)}(x) \\ \vdots \\ \phi^{(m)}(x) \end{bmatrix}, \quad m \gg d
\]

where \( \phi^{(j)} \) 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 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 the data through $\Phi$.

- For certain feature maps $\Phi$, 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

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

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

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

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

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

What feature mapping $\Phi$ does this correspond to?
\[
\mathbf{H} = \begin{bmatrix}
(U^{(1)})^2 \\
\sqrt{2} \cdot U^{(1)} \cdot U^{(2)} \\
(U^{(2)})^2
\end{bmatrix}
\]

Now suppose \( d \) is arbitrary, and

\[
k(u, v) = (u^Tv)^2
\]

\[
= \left( \sum_{i=1}^{d} u^{(i)} v^{(i)} \right)^2
\]

\[
= \left( \sum_{i=1}^{d} u^{(i)} v^{(i)} \right) \left( \sum_{j=1}^{d} u^{(j)} v^{(j)} \right)
\]

\[
= \sum_{i=1}^{d} \sum_{j=1}^{d} u^{(i)} u^{(j)} v^{(i)} v^{(j)}
\]

What is the dimension of the corresponding feature space?
\[ \Phi(u) = [(u_1)^2, \ldots, (u_d)^2, \sqrt{2} u_1 u_2, \ldots, \sqrt{2} u_{d-1} u_d]^T \]

\[ \text{Exercise 1} \quad \text{Describe the feature space associated with} \]
\[ k(u,v) = (u^T v)^3 \]
\[ \text{when } d = 2. \]
Solution

\[ k(u, v) = (u^{(1)} v^{(1)} + u^{(2)} v^{(2)})^3 \]

\[ = (u^{(1)})^3 (v^{(1)})^3 + 3 (u^{(1)})^2 u^{(2)} (v^{(1)})^2 v^{(2)} + \]
\[ 3 u^{(1)} (u^{(2)})^2 v^{(1)} (v^{(2)})^2 + (u^{(2)})^3 (v^{(2)})^3 \]

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

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

More generally,

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

\[ = \sum_{(j_1, \ldots, j_d)} \binom{p}{j_1 \ldots j_d} u^{(j_1)} v^{(j_2)} \ldots u^{(j_d)} v^{(j_1)} \ldots u^{(j_d)} v^{(j_1)} \ldots \]

\[ \sum_{\sum j_k = p} \sum_{(j_1, \ldots, j_d)} \binom{p}{j_1 \ldots j_d} u^{(j_1)} v^{(j_2)} \ldots u^{(j_d)} v^{(j_1)} \ldots u^{(j_d)} v^{(j_1)} \ldots \]

\[ \Rightarrow \Phi(u) = \left[ \ldots, \sqrt{\binom{p}{j_1 \ldots j_d}} u^{(j_1)} v^{(j_2)} \ldots u^{(j_d)} v^{(j_1)} \ldots \right]^T \]

\[ \Rightarrow \text{all } \ldots \text{ of degree } p. \]
**Definition**

An is a mapping $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ for there exists an inner product space $\mathcal{H}$ and a mapping $\Phi: \mathbb{R}^d \rightarrow \mathcal{H}$ such that

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

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

**Note**

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

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
PSD kernels

Definition | We say \( k(u,v) \) is a \( \blacksquare \) kernel if \( k \) is symmetric and, for all \( n \) and all \( x_1, \ldots, x_n \in \mathbb{R}^d \), the \( \blacksquare \) matrix

\[
\begin{bmatrix}
    k(x_i, x_j)
\end{bmatrix}_{i,j=1}^n
\]

is positive semi-definite.

Theorem | \( k \) is an IP kernel \( \iff \) \( k \) 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

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

\( \Phi \rightarrow \)

3. Gaussian kernel

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

\( \rightarrow \) constant can be dropped

\( \rightarrow \) also called radial basis function (RBF) kernel

\( \rightarrow \) \( \Phi \) is \ ______ ______ !
**Key**

A. computational

B. inner products,

\[ k(x,x') = \langle \Phi(x), \Phi(x') \rangle \]

C. 

\[
\left( \sum_{i=1}^{2} u^{(c)} u^{(i)} \right)^2 = \left( \sum_{i=1}^{2} u^{(c)} v^{(c)} \right) \left( \sum_{j=1}^{2} u^{(j)} v^{(j)} \right)
\]

\[
= \sum_{i=1}^{2} \sum_{j=1}^{2} u^{(c)} u^{(j)} v^{(c)} v^{(j)}
\]

\[
= (u^{(1)})^2 (v^{(1)})^2 + 2 u^{(1)} u^{(2)} v^{(1)} v^{(2)} + (u^{(2)})^2 (v^{(2)})^2
\]

\[ = \langle \Phi(u), \Phi(v) \rangle \]

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