Constructing & validating Kernels

For some Matrix X with traing samples $\vec{x}_i$

$\phi(x)$ is the feature transformation
for example $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \text{height} \\ \text{age} \end{bmatrix}$ $\phi(x)$ could be $\begin{bmatrix} \text{height} \\ \text{age} \\ \text{age}^2 \\ \text{height} \cdot \text{age} \\ \log(\text{height}) \end{bmatrix}$

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix} \quad \Phi(X) = F = \begin{bmatrix} | & | & & | \\ f_1 & f_2 & \cdots & f_d \\ | & | & & | \end{bmatrix} \quad K = F^T F = \begin{bmatrix} f_1^T f_1 & \cdots & f_1^T f_n \\ \vdots & f_i^T f_j & \vdots \\ f_n^T f_1 & \cdots & f_n^T f_n \end{bmatrix} = \begin{bmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & k(x_i,x_j) & \vdots \\ k(x_1,x_n) & \cdots & k(x_n,x_n) \end{bmatrix}$$

Feature matrix. Call it F for simplicity. Transformations $\phi(x)$ of samples X.

Kernel Matrix often also called Gram Matrix

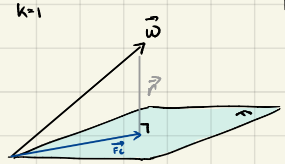
if for some F this mapping exists then $K(x,y)$ is valid.

Note: $\Leftrightarrow$ K is symmetric and p.s.d.

Kernelizing functions

represent w as a linear combination of f_i.

$$w = \sum_{k=1}^{d} f_k c_k + \vec{r} \quad \text{where } c \begin{bmatrix} c_1 \\ \vdots \\ c_d \end{bmatrix} \text{ is the coefficient vector.}$$

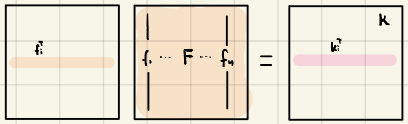


Step by Step easy kernelization:

1. Start with function $\mathcal{L}(w)$
2. Wherever you see $f^T w$ in $\mathcal{L}(w)$ replace w by $F\vec{c}$.
3. note that $f_j^T w = f_j^T F c = k_i^T c$

can ignore r bc: $f_i^T(Fc + r) = f_i^T Fc + f_i^T r = f_i^T Fc$

which is a column times z of the Kernel Matrix K



Examples:

Least squares

1. $\frac{1}{n}\sum_{i=1}^{n}(w_0 + f_i^T w - y_i)^2$
2. $= \frac{1}{n}\sum_{i=1}^{n}(w_0 + f_i^T Fc - y_i)^2$ $f_j^T r = 0$ all j
3. $= \frac{1}{n}\sum_{i=1}^{n}(w_0 + k_i^T c - y_i)^2$

Ridge = Least Squares + $\lambda\|w\|^2$

1. $\|w\|^2 = w^T w$
2. $(Fc)^T(Fc) = c^T F^T F c$
3. $c^T K c$

Softmax:

1. $\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(w_0 + f_i^T w)))$
2. $\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(w_0 + f_i^T Fc)))$
3. $\frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i(w_0 + k_i^T c)))$

Kernel Trick: Compute K without computing F and only using X.

Polynomial Features (sklearn.preprocessing.PolynomialFeatures)

$f = \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$ note multiplying some components by a scalar will just be absorbed by the weights. so we let $f = \begin{bmatrix} \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ \sqrt{2}x_1 x_2 \\ x_2^2 \end{bmatrix}$

Why not just use f_i, f_j? Feature / Combinatorial explosion for degree d and n features $\frac{(n+d)!}{n!d!}$. Try in sklearn. Even others are worse. Infinite!

let $u = x_i$ and $v = x_j$

$$K_{i,j} = f_i^T f_j = \begin{bmatrix} \sqrt{2}u_1 & \sqrt{2}u_2 & u_1^2 & \sqrt{2}u_1u_2 & u_2^2 \end{bmatrix} \begin{bmatrix} \sqrt{2}v_1 \\ \sqrt{2}v_2 \\ v_1^2 \\ \sqrt{2}v_1v_2 \\ v_2^2 \end{bmatrix}$$

$$= (1 + 2u_1v_1 + 2u_2v_2 + u_1^2v_1^2 + 2u_1u_2v_1v_2 + u_2^2v_2^2) - 1$$

$$= (1 + u_1v_1 + u_2v_2)^2 - 1$$

$$= (1 + u^T v)^2 - 1$$

only need x_i, x_j to compute $K_{i,j}$

Let $k^*(x,y) = \sqrt{\exp(-t)}\,(\exp(-t)+1)^2$ where $t = \|x-y\|^2$

is $k^*(x,y)$ a valid kernel?

$$\exp(-t)(\exp(-t)+1)^2 = \exp\left(\frac{-t}{2}\right)\left[\exp(-2t) + 2\exp(-t) + 1\right]$$

$$= \exp\left(\frac{-5t}{2}\right) + 2\exp\left(\frac{-3t}{2}\right) + \exp\left(\frac{-t}{2}\right)$$

Ugh!!! how does this help?

let $\sigma = \frac{\sqrt{5}}{5}, \frac{\sqrt{3}}{3}$ & 1 then apply Rule 5

Helpful Rules

Given $k_1(x,y)$ and $k_2(x,y)$ are valid Kernels the following will also be valid Kernels:

(1) $k(x,y) = c\,k_1(x,y)$
(2) $k(x,y) = f(x)\,k_1(x,y)\,f(y)$
(3) $k(x,y) = q(k_1(x,y))$
(4) $k(x,y) = \exp(k_1(x,y))$
(5) $k(x,y) = k_1(x,y) + k_2(x,y)$
(6) $k(x,y) = k_1(x,y) \cdot k_2(x,y)$

c is a scalar
f is any function
q is a polynomial with non-negative coefficients

Common Kernels

Linear: $K(x,y) = x^T y$
Polynomial: $K(x,y) = (t\,x^T y + c)^d$
Gaussian RBF: $K(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$
Sigmoid: $K(x,y) = \tanh(t\,x^T y + c)$
Laplacian: $K(x,y) = \exp(-\|x-y\|)$

How 2 check

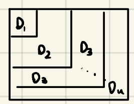
shortcuts:

inner product? Yes, $\Rightarrow$ valid Kernel. ! careful, converse does not hold!

can you compose it of known Kernels with the rules above? Yes? $\Rightarrow$ valid Kernel

if no shortcuts are applicable, check if K p.s.d

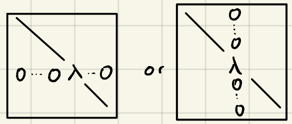
1. all eigenvalues $\geq 0 \Rightarrow$ p.s.d.
2. energy K is p.s.d if $x^T K x \geq 0$ for all $x=0$
3. All leading Determinants $D_1, \dots D_n$ are non-negative



generally easier to use these to show k is not a kernel.

Quick eigenvalues

$\sum\lambda = Tr(K)$ $\quad \prod\lambda = \det(K)$



if diagonal λs are entries useful if Block structure