

# KERNEL METHODS

→ Is MMLC (with convex loss) convex program?

→ Yes, as hessian  $K$  (gram matrix) is always a psd matrix ( $K \succeq 0$ )

Proof:  $K \succeq 0 \Leftrightarrow \beta^T K \beta \geq 0 \quad \forall \beta$

$$\Leftrightarrow \sum_i \sum_j \beta_i \beta_j \varphi(x_i)^T \varphi(x_j) \geq 0 \quad \forall \beta$$

$$\Leftrightarrow \left\| \sum_i \beta_i \varphi(x_i) \right\|^2 \geq 0 \quad \forall \beta$$

In other words,  $K = X^T X$ , where  $X = \begin{bmatrix} \varphi(x_1) & \dots & \varphi(x_m) \end{bmatrix}_{n \times m}$

→ In particular co-ordinate descent has global convergence.

Important observations:

- (i)  $\varphi$  has no restrictions; whereas  $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  has ( $K \succeq 0$ ).
- (ii) Given  $\varphi$ ,  $K$  is fixed;  $K$  given  $\nRightarrow \varphi$  is given.
- (iii) Only  $K$  is needed for MMLC and not  $\varphi$ .
- (iv) Given  $K$ ,  $\varphi$  can be computed (Eigen Value or Cholesky decomposition)



KERNEL: A function  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  is called as a kernel over  $\mathcal{X}$  iff every gram matrix induced by  $k$  is  $\geq 0$ .

i.e. if  $K_{ij} = k(x_i, x_j)$ , then  $K \geq 0 \quad \forall x_1, \dots, x_m \in \mathcal{X}, \quad \forall m$ .

### EXAMPLES OF KERNELS over $\mathbb{R}^n$

(i)  $k(x, z) = x^T z, \quad \forall x, z \in \mathbb{R}^n$   
 $\left( = \sum_{i=1}^n x_i z_i \right)$

$\left( \because K = X^T X, \text{ where } X = [x_1 \ x_2 \ \dots \ x_m]_{n \times m} \right)$

(ii)  $k(x, z) = x^T W z \quad \forall x, z \in \mathbb{R}^n; \quad W \text{ is diagonal } \geq 0$   
 $\left( = \sum_{i=1}^n \omega_i x_i z_i \right)$

$\left( \because K = X^T W X, \text{ where } X = [x_1 \ \dots \ x_m]_{n \times m} \right)$

(iii)  $k(x, z) = x^T W z \quad \forall x, z \in \mathbb{R}^n; \quad W \geq 0$

$\left( \because K = X^T W X, \text{ where } X = [x_1, \dots, x_m]_{n \times m} \right)$



Interestingly,  $g(x) = \omega^T \phi(x) - b$

$$= \sum \alpha_i k(x_i, x) - b$$

$$= \sum \alpha_i x_i^T W x - b,$$

which is linear in  $x$ !

So,  $k(x, z) \equiv x^T W z$  is called as linear kernel.

$W$  is the kernel's parameter or hyperparameter.

## Operations preserving kernelity

(i) If  $k_1, k_2$  are kernels over  $X$ , then  $k \equiv k_1 + k_2$  is a kernel (over  $X$ )

$$\therefore \alpha^T K \alpha = \underbrace{\alpha^T K_1 \alpha}_{\geq 0} + \underbrace{\alpha^T K_2 \alpha}_{\geq 0} \geq 0 \quad \forall \alpha \in \mathbb{R}^m, \forall m.$$

(ii) If  $\alpha \geq 0$ ,  $k_1$  is kernel (over  $X$ ), then  $k \equiv \alpha k_1$  is also a kernel (over  $X$ )

(iii) If  $\alpha_1, \dots, \alpha_n \geq 0$ ,  $k_1, \dots, k_n$  are kernels over  $X$ , then

$$k = \sum_{i=1}^n \alpha_i k_i \text{ is a kernel over } X.$$



(iv)  $k_1, k_2$  are kernels over  $X$ , then  $k \equiv k_1 k_2$  is a kernel over  $X$ .

$$\therefore K = K_1 \circ K_2 = \left( \sum_i \lambda_i v_i v_i^T \right) \circ \left( \sum_j \rho_j u_j u_j^T \right), \quad (\lambda, \rho \geq 0)$$

$$= \sum_i \sum_j \lambda_i \rho_j (v_i v_i^T \circ u_j u_j^T)$$

$$= \sum_i \sum_j \sigma_{ij} z_{ij} z_{ij}^T, \quad \sigma_{ij} = \lambda_i \rho_j \geq 0$$

$$z_{ij} = v_i \circ u_j$$

$$\beta^T K \beta = \sum_i \sum_j \sigma_{ij} (\beta^T z_{ij})^2 \geq 0 \quad \forall \beta.$$

Given a set of kernels, all polynomials (with  $\geq 0$  coeff.) over them are kernels.

(v) If  $k_1, k_2, \dots, k_n, \dots$  are kernels,  $\lim_{n \rightarrow \infty} k_n$  exists and is itself a  $k: X \times X \rightarrow \mathbb{R}$ , then,  $k$  is a kernel.

$$\therefore \underbrace{\beta^T K_1 \beta}_{\geq 0}, \dots, \underbrace{\beta^T K_n \beta}_{\geq 0}, \dots \rightarrow \beta^T K \beta \quad \therefore \geq 0$$



Given some kernels, all analytic functions (with  $\geq 0$  coeffs.) over them are kernels.

### More Examples

$$(iv) k(x, z) \equiv (1 + x^T W z)^d \quad \forall x, z \in \mathcal{X}, \quad W \succeq 0, \quad d \in \mathbb{N}.$$

POLYNOMIAL KERNEL.

$$g(x) = w^T \phi(x) - b = \sum_i \alpha_i k(x_i, x) - b = \sum_i \alpha_i (1 + x_i^T W x)^d - b$$

(polynomial of degree at most  $d$  in  $x$ )

$$(v) k(x, z) \equiv e^{x^T W z} \quad \forall x, z \in \mathcal{X}, \quad W \succeq 0$$

EXPONENTIAL KERNEL

$$g(x) = \sum_i \alpha_i e^{x_i^T W x} - b,$$

$$(vi) k(x, z) \equiv e^{-\frac{1}{2}(x-z)^T W (x-z)}, \quad W \succeq 0$$

GAUSSIAN KERNEL

$$g(x) = \sum_i \alpha_i e^{-\frac{1}{2}(x_i - x)^T W (x_i - x)} - b$$



→ Normalization preserves kernelity

(vi) If  $k$  is a kernel, then  $\bar{k}(x, z) \equiv \frac{k(x, z)}{\sqrt{k(x, x) k(z, z)}}$  is a kernel.

$\therefore \tilde{k}(x, z) = \frac{1}{\sqrt{k(x, x)}} \frac{1}{\sqrt{k(z, z)}}$  is a kernel.

eg  $k(x, z) = e^{x^T W z}$   $W \succeq 0$ .

$$\bar{k}(x, z) = \frac{e^{x^T W z}}{\sqrt{e^{x^T W x} e^{z^T W z}}} = e^{-\frac{1}{2} (x-z)^T W (x-z)}$$

Gaussian kernel is normalized exponential kernel.

## Utility of kernels

- ①  $k$  is a means of "minimally" specifying  $\varphi$ .
- ② Designing  $k$  more intuitive for domain experts than  $\varphi$ .
- ③ Inducing non-linear prediction functions over  $X$ .
- ④ Not limited to SVMs or MM/LC.  
eg. kernelized nearest neighbours.