KERNEL METHODS

→ Is MMLC (with convex loss) Convex program ?

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

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

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

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

In other words, $K = X^T X$, where $X = \left[\varphi(x_1) \; \cdots \; \varphi(x_m) \right]_{n \times m}$

→ In particular co-ordinate descent has global convergence.

Important observations:

(i) 'φ' has no restrictions ; whereas $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ has $(K \succeq 0)$.

(ii) Given 'φ', k is fixed ; k given $\not\Rightarrow \varphi$ is given.

(iii) Only 'k' is needed for MMLC and not 'φ'

(iv) Given K, 'φ' can be computed (Eigen Value & Cholesky decomposition)

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

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

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

(i) $k(x, z) = x^T z$, $\forall\; x, z \in \mathbb{R}^n$.

$$\left(= \sum_{i=1}^{n} x_i z_i\right)$$

$$\left(\because K = X^T X, \text{ where } X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}_{n \times m}\right)$$

(ii) $k(x, z) = x^T W z \quad \forall\; x, z \in \mathbb{R}^n$; W is diagonal $\succeq 0$.

$$\left(= \sum_{i=1}^{n} w_i x_i z_i\right)$$

$$\left(\because K = X^T W X, \text{ where } X = \begin{bmatrix} x_1 & \cdots & x_m \end{bmatrix}_{n \times m}\right)$$

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

$$\left(\because K = X^T W X, \text{ where } X = \begin{bmatrix} x_1, & \ldots, & x_m \end{bmatrix}_{n \times m}\right)$$

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

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

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

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 Kernality

(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, $\alpha_1, \ldots, \alpha_n \geq 0$, $k_1, \ldots k_n$ are kernels over X, then

$K = \sum_{i=1}^{n} \alpha_i k_i$ 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.

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

$$= \sum_i \sum_j \lambda_i \rho_j \left(v_i v_i^T \odot u_j u_j^T\right)$$

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

$$\gamma_{ij} = v_i \odot u_j$$

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

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

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

$\therefore \beta^T K_1 \beta, \ldots, \beta^T K_n \beta, \ldots \to \beta^T K \beta$

$\geq 0 \qquad \geq 0 \qquad \therefore \geq 0$

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

More Examples

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

POLYNOMIAL KERNEL.

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

(polynomial of degree atmost 'd' in x)

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

EXPONENTIAL KERNEL

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

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

GAUSSIAN KERNEL

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

→ Normalization preserves kernality.

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

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

Eg $k(x,z) = e^{x^T W^{-1} z}$, $W \succ 0$.

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

Gaussian kernel is normalized exponential kernel.

Utility of kernels

(1) 'k' is a means of "minimally" specifying 'φ'.

(2) Designing 'k' more intuitive for domain experts than 'φ'.

(3) Inducing non-linear prediction functions over 'x'.

(4) Not limited to SVMs or MMLC.

eg. kernelized nearest neighbours.