Kernel functions (Quick recap)

Feature Mapping

$$\tilde{\underline{x}} = \underline{\phi}(\underline{x})$$

$\tilde{\underline{x}}$: vector in the transformed space e.g., 3-D

$\underline{\phi}$: vector function

$\underline{x}$: vector in the original space e.g., 2-D

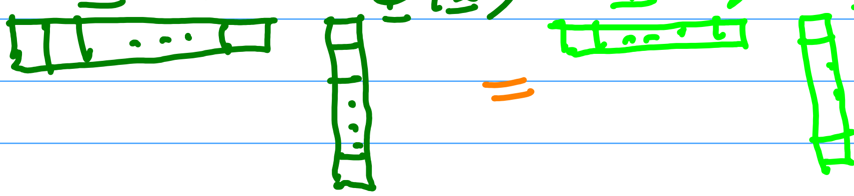
Kernel function:

$$\overbrace{k(\underline{x}, \underline{x}')}^{\text{scalar}} \triangleq \underline{\phi}^T(\underline{x})\,\underline{\phi}(\underline{x}')$$

$\underline{x}, \underline{x}'$: two vectors in the original space

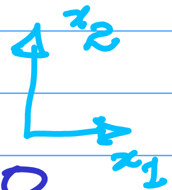
$\underline{x}$: first, $\underline{x}'$: second (original space)

$$\underbrace{\underline{\phi}^T(\underline{x})\,\underline{\phi}(\underline{x}') = \underline{\phi}^T(\underline{x}')\,\underline{\phi}(\underline{x})}_{\text{symmetric}}$$



Some common examples of Kernels

1) linear kernel $\underline{\phi}(\underline{x}) = \underline{x}$ itself.

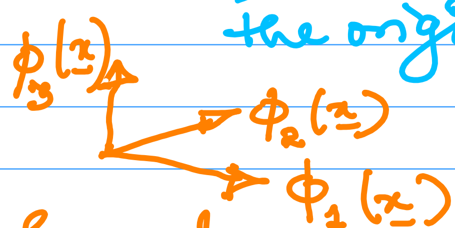


e.g., linear decision boundary $\underbrace{\underline{w}^T\underline{x} + b = 0}_{\text{linearity in the original space}}$

In the transformed domain

$$\tilde{\underline{w}}^T\underline{\phi}(\underline{x}) + b = 0$$

$\tilde{\underline{w}}$: appropriate dimension



linearity in the transformed space.

(for a linear kernel: the transformation is an identity transformation)

(2) <u>Stationary Kernel</u> $k(\underline{x}, \underline{x}') = k(\underline{x} - \underline{x}')$

invariant to translations in the input pattern space.

e.g., music information retrieval
male, female voice ('pitch transposition')

(2a) <u>RBF</u> $k(\underline{x}, \underline{x}') = k(\|\underline{x} - \underline{x}'\|)$
(Radial Basis Function)

<u>Key point</u>: <u>'Kernel Trick'</u>

(*) Kernel function: Why? Feature transform - decision boundary: linear in a higher dimension, or at least the linear decision boundary in the transformed space could give a better separation as compared to the original space.

(*) Computation: 'trick': make computations in the lower dimensional space itself.

Dual Representations (Regression)

Earlier example: SVM: classification

(*) Many linear models for classification and regression: which can be reformulated in terms of a dual representation in which kernels arise naturally.

Regularised Linear Regression

$$J(\underline{w}) = \frac{1}{2}\sum_{i=1}^{N}\{\underbrace{\underline{w}^T\Phi(\underline{x})}_{\text{model (linear)}} - \underbrace{t_i}_{\text{target value}}\}^2 + \frac{\lambda}{2}\underline{w}^T\underline{w}$$

"cosmetic purposes"

function, to minimise

in $\Phi(\underline{x})$ space

difference b/w the target and the model

"fidelity" term

Summation: for all training data points

"cosmetic purposes"

Regulariser

Lagrange multiplier

+/- discrepancies b/w the model and the target are treated in the same way

Regulariser → drive the system to favour low weights, unless it is supported by the data (fidelity)

Recap

$$\frac{\partial(\underline{x}^T\underline{a})}{\partial\underline{x}} = \underline{a} \equiv \frac{\partial(\underline{a}^T\underline{x})}{\partial\underline{x}} = \underline{a}$$

optimum → minimum

$$\frac{\partial J(\underline{w})}{\partial\underline{w}} = 0$$

$\lambda \geq 0$ Lagrange Multiplier

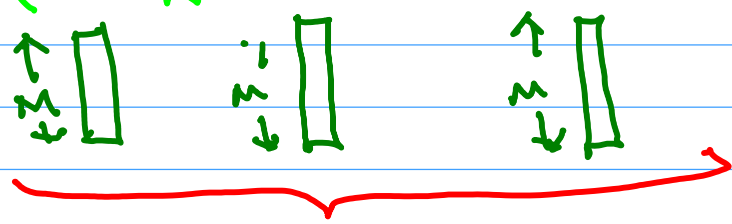
$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = 0 \Rightarrow \frac{1}{\cancel{2}} \times \cancel{2} \sum_{i=1}^{N} \{\underline{w}^T \underline{\phi}(\underline{x}_i) - t_i\} \underline{\phi}(\underline{x}_i) + \frac{\lambda}{\cancel{2}} \cancel{2}\, \underline{w} = 0$$

$$\Rightarrow \underline{w} = \left(-\frac{1}{\lambda}\right) \sum_{i=1}^{N} \underbrace{\{\underline{w}^T \underline{\phi}(\underline{x}_i) - t_i\}}_{a_i} \underline{\phi}(\underline{x}_i)$$

$$a_i \triangleq \left(-\frac{1}{\lambda}\right) \{\underline{w}^T \underline{\phi}(\underline{x}_i) - t_i\}$$

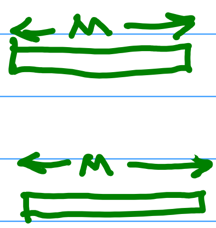
$$\Rightarrow \underline{w} = \underbrace{\sum_{i=1}^{N} a_i \underline{\phi}(\underline{x}_i)}_{\text{inner product representation}}$$

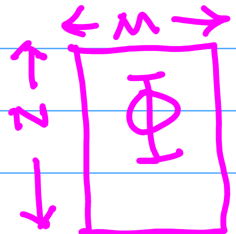
$$\overbrace{[\underline{\phi}(\underline{x}_1)\ \ \underline{\phi}(\underline{x}_2)\ \cdots\ \underline{\phi}(\underline{x}_N)]}^{}_{\longleftarrow N \longrightarrow} \ \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\underline{a}}$$



Φ : transpose of the "Φ"
'Design Matrix'

$$\Phi \text{ was } \begin{array}{c} \\ N \end{array} \begin{bmatrix} \underline{\phi}^T(\underline{x}_1) \\ \vdots \\ \underline{\phi}^T(\underline{x}_N) \end{bmatrix}$$





$$\Rightarrow \boxed{\underline{w} = \Phi^T \underline{a}}$$

Reformulate the problem in terms of $\underline{a}$, instead of $\underline{w}$

Substitute $\underline{w} = \Phi^T \underline{a}$ into the expression for $J(\underline{w})$ to try to eliminate w altogether, and attempt to replace $J(\underline{w})$ with an expression which involves $\underline{a}$ alone, at the optimum.

$$J(\underline{a}) = \frac{1}{2}\sum_{i=1}^{N}\left\{ (\Phi^T \underline{a})^T \underline{\phi}(\underline{x}_i) - t_i \right\}^2 + \frac{\lambda}{2}(\Phi^T \underline{a})^T(\Phi^T \underline{a})$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left\{ (\underline{a}^T \Phi)\, \underline{\phi}(\underline{x}_i) \right\}^2 - \frac{1}{2}2\sum_{i=1}^{N}(\underline{a}^T \Phi)\, \underline{\phi}(\underline{x}_i)\, t_i$$

$$+ \frac{1}{2}\sum_{i=1}^{N} t_i^2 + \underbrace{\frac{\lambda}{2}\underline{a}^T \Phi \Phi^T \underline{a}}$$

fourth term, is in its final form (we will see this later!)

The third term $= \frac{1}{2}\sum_{i=1}^{N} t_i^2$ we write this as an inner product

$$= \frac{1}{2}\begin{bmatrix} t_1 & t_2 & \cdots & t_N \end{bmatrix}\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix} = \frac{1}{2}\underline{t}^T \underline{t}$$

The second term $= -\sum_{i=1}^{N} \boxed{\underline{a}^T \Phi}\, \underline{\phi}(\underline{x}_i)\, t_i$

$$= -(\underline{a}^T \Phi)\sum_{i=1}^{N} t_i\, \underline{\phi}(\underline{x}_i)$$

$$= -(\underline{a}^T \Phi)\underbrace{\begin{bmatrix} \underline{\phi}(\underline{x}_1) & \underline{\phi}(\underline{x}_2) & \cdots & \underline{\phi}(\underline{x}_N) \end{bmatrix}}_{\Phi^T}\overbrace{\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}}^{\underline{t}} = -\underline{a}^T \Phi \Phi^T \underline{t}$$