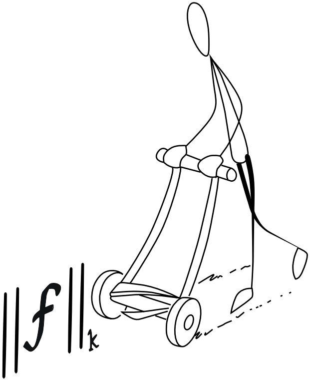
# KERNEL MACHINES

Marco Gori, Lisa Graziani



# FEATURE SPACE

- The linear machines are limited either in regression or in classification. The linearity assumption in some real-world problems is quite restrictive.
- We need to transform the input space to an enriched space (*feature space*) in order to deal with not-linear problem or not linearly-separable patterns.
- The features are determined by the *feature map*

$$\phi : \mathscr{X} \subset \mathbb{R}^d \rightarrow \mathscr{H} \subset \mathbb{R}^D,$$

where in most cases, $D \geq d$, and often $D \gg d$.

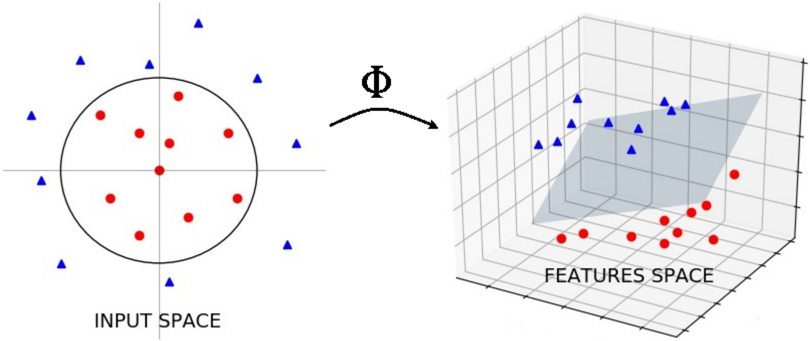
Suppose we are given a classification problem with patterns $x \in \mathscr{X} \subset \mathbb{R}^2$. We consider the associated feature space defined by the map $\phi : \mathscr{X} \subset \mathbb{R}^2 \rightarrow \mathscr{H} \subset \mathbb{R}^3$ such that $x \rightarrow z = (x_1^2, x_1 x_2, x_2^2)'$.

Linear-separability in $\mathscr{H}$ yields a quadratic separation in $\mathscr{X}$:

$$a_1 z_1 + a_2 z_2 + a_3 z_3 + a_4 = a_1 \cdot x_1^2 + a_2 \cdot x_1 x_2 + a_3 \cdot x_2^2 + a_4.$$

# FEATURE SPACE

## Example



# MAXIMUM MARGIN PROBLEM

## Classification under linear-separability

Let us consider a linear machine in the feature space

$$f(x) = w'\phi(x) + b = \hat{w}'\hat{\phi}(x),$$

where $\hat{\phi}(x) := (\phi_1(x), \ldots, \phi_D(x), 1)'$.

Let $\mathscr{L} = \{(x_\kappa, y_\kappa),\ \kappa = 1, \ldots, \ell\}$ be the training set, with $y_\kappa \in \{-1, +1\}$, and let us assume that the feature space $\mathscr{L}_\phi = \{(\phi(x_\kappa), y_\kappa),\ \kappa = 1, \ldots, \ell\}$ is linearly-separable.

The *maximum margin problem* is determining $\hat{w}^\star$ such that

$$\hat{w}^\star = \arg\max_{\hat{w}} \left\{ \frac{1}{\| w \|} \min_{\kappa} \left( y_\kappa \cdot \hat{w}'\hat{\phi}(x_\kappa) \right) \right\}. \qquad (1)$$

# MAXIMUM MARGIN PROBLEM

## Geometrical interpretation of the problem in the feature space

The distance of $\phi(x_\kappa)$ to the hyperplane defined by $\hat{w}$ is

$$d(\kappa, \hat{w}) := \frac{y_\kappa \cdot \hat{w}'\hat{\phi}(x_\kappa)}{\|w\|} = \frac{|\hat{w}'\hat{\phi}(x_\kappa)|}{\|w\|}.$$

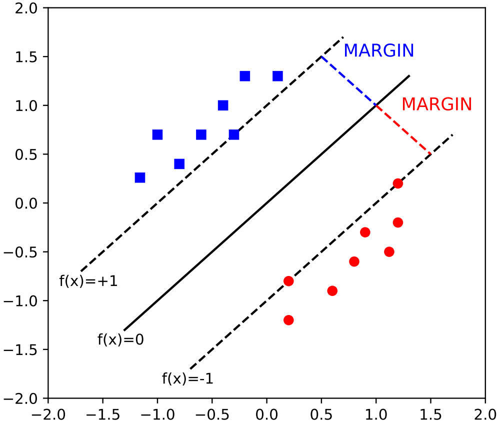
(The equivalence $y_\kappa \cdot \hat{w}'\hat{\phi}(x_\kappa) = |\hat{w}'\hat{\phi}(x_\kappa)|$ is due to hypothesis of linearly separable examples in the feature space.)

So we have to find the hyperplane defined by $\hat{w}$ such that the distance between the nearest $\phi(x_\kappa)$ and the hyperplane is maximized. This distance is called MARGIN.

# MAXIMUM MARGIN PROBLEM

Example

In 2-dimensional spaces we have to find the separation line such that the distance between the nearest point to the line in each side and the line is maximized.



# MAXIMUM MARGIN PROBLEM

The maximum margin problem (1) is equivalent to the following optimization problem:

$$\begin{cases} \min \frac{1}{2} w^2 \\ \qquad 1 - y_\kappa \cdot \hat{w}' \hat{\phi}(x_\kappa) \leq 0, \quad \kappa = 1, \dots, \ell \end{cases} \tag{2}$$

To solve it we consider the Lagrangian function:

$$\mathcal{L}(\hat{w}, \lambda) = \frac{1}{2} w^2 + \sum_{\kappa=1}^{\ell} \lambda_\kappa \left(1 - y_\kappa \cdot \hat{w}' \hat{\phi}(x_\kappa)\right), \quad \text{with } \lambda \geq 0. \tag{3}$$

# MAXIMUM MARGIN PROBLEM

If we impose $\nabla \mathcal{L}(\hat{w}, \lambda) = 0$ then we have

$$\partial_w \mathcal{L}(\hat{w}, \lambda) = w - \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa \phi(x_\kappa) = 0$$

$$\partial_b \mathcal{L}(\hat{w}, \lambda) = - \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa = 0.$$

Now we can re-write the Lagrangian as function of the Lagrangian multiplier only.
From the first equation we obtain $w = \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa \phi(x_\kappa)$.

# MAXIMUM MARGIN PROBLEM

$$\theta(\lambda) = \inf_{\hat{w}} \mathcal{L}(\hat{w}, \lambda) = \frac{1}{2}\Big(\sum_{h=1}^{\ell} \lambda_h y_h \phi(x_h)\Big)' \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa \phi(x_\kappa)$$

$$- \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa \Big( \sum_{h=1}^{\ell} (\lambda_h y_h \phi(x_h))' \phi(x_\kappa) + b \Big) + \sum_{\kappa=1}^{\ell} \lambda_\kappa$$

$$= \frac{1}{2} \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} \lambda_h \lambda_\kappa y_h y_\kappa \phi(x_h)' \phi(x_\kappa)$$

$$- \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} \lambda_h \lambda_\kappa y_h y_\kappa \phi(x_h)' \phi(x_\kappa) - b \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa + \sum_{\kappa=1}^{\ell} \lambda_\kappa$$

$$= -\frac{1}{2} \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} \lambda_h \lambda_\kappa y_h y_\kappa \phi(x_h)' \phi(x_\kappa) + \sum_{\kappa=1}^{\ell} \lambda_\kappa .$$

# MAXIMUM MARGIN PROBLEM

The maximum margin problem (2) is equivalent to the *dual optimization problem*:

$$\begin{cases} \max \theta(\lambda) = \sum_{\kappa=1}^{\ell} \lambda_\kappa - \frac{1}{2} \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} k(x_h, x_\kappa) y_h y_\kappa \cdot \lambda_h \lambda_\kappa \\ \lambda_\kappa \geq 0, \quad \kappa = 1, \ldots, \ell \\ \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa = 0 \end{cases} \tag{4}$$

where $k$ is the *kernel function*:

$$k : \mathscr{X} \times \mathscr{X} \to \mathbb{R} : \; k(x_h, x_\kappa) := \phi'(x_h)\phi(x_\kappa).$$

# MAXIMUM MARGIN PROBLEM

The optimal function turns out to be

$$f^\star(x) = (w^\star)'\phi(x) + b^\star = \sum_{\kappa=1}^{\ell} \left(\lambda_\kappa^\star y_\kappa \phi(x_\kappa)\right)' \phi(x) + b^\star$$

$$= \sum_{\kappa=1}^{\ell} y_\kappa \lambda_\kappa^\star k(x_\kappa, x) + b^\star.$$

If we define $\hat{\lambda} := (\lambda_1, \ldots, \lambda_\ell, b)'$ and $k_i(x) := k(x_i, x)$ then $f(x) = \hat{\lambda}' k(x)$.

- Primal: $f(x) = \hat{w}'\hat{\phi}(x)$, parameter $\hat{w}$.
- Dual: $f(x) = \hat{\lambda}' k(x)$, parameter $\hat{\lambda}$.

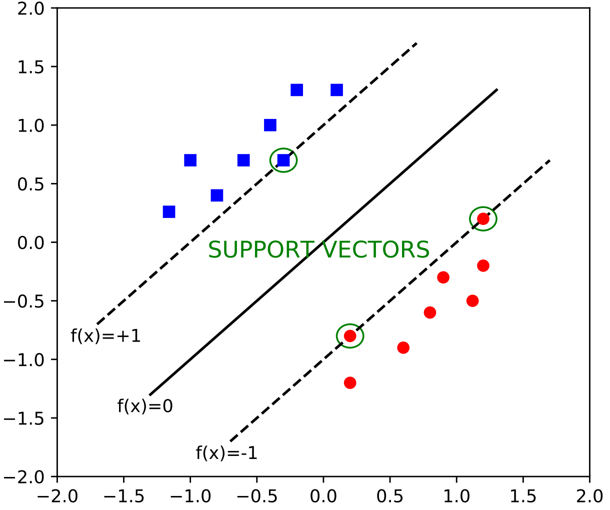
# MAXIMUM MARGIN PROBLEM

From the Karush Kuhn Tucker (KKT) conditions we have

$$\lambda_\kappa^\star(y_\kappa f^\star(x_\kappa) - 1) = 0, \quad \kappa = 1, \ldots, \ell.$$

- $\lambda_\kappa^\star = 0$. $\Rrightarrow$ $y_\kappa f^\star(x_\kappa) > 1$, and this means that the stationary condition is satisfied with an interior coordinate.
  $x_\kappa$ is called a *straw vector*.
- $\lambda_\kappa^\star > 0$. $\Rrightarrow$ $y_\kappa f^\star(x_\kappa) = 1$, and this means that the stationary condition is met on the border.
  $x_\kappa$ is called a *support vector*.

# MAXIMUM MARGIN PROBLEM



# MAXIMUM MARGIN PROBLEM

## Dealing with soft-constraints

In the previous margin problem (2) the patterns are assumed to be linearly-separable, but this is a critical assumption.
We relax the constraints: we introduce *slack variables* $\xi_\kappa$, $\kappa = 1, \ldots, \ell$, one for each example. They are used for tolerating the violation of the constraints as follows

$$\begin{cases} y_\kappa f(x_\kappa) \geq 1 - \xi_\kappa \\ \xi_\kappa \geq 0. \end{cases} \tag{5}$$

- $\xi_\kappa = 0 \Rrightarrow$ previous MMP formulation.
- $\xi_\kappa \in (0, 1) \Rrightarrow$ the solution is still correct.
- $\xi_\kappa = 1 \Rrightarrow f(x_\kappa) = 0$, so we have uncertain decision.
- $\xi_\kappa > 1 \Rrightarrow$ we have the strongest constraint relaxation, that might led to errors.

# MAXIMUM MARGIN PROBLEM

The constraints defined by (5) suggest us to define the following optimization problem:

$$\begin{cases} \min \frac{1}{2} w^2 + C \sum_{\kappa=1}^{\ell} \xi_\kappa \\ y_\kappa f(x_\kappa) \geq 1 - \xi_\kappa, \\ \xi_\kappa \geq 0, \quad \kappa = 1, \ldots, \ell. \end{cases}$$

# MAXIMUM MARGIN PROBLEM

The Lagrangian is

$$\mathcal{L}(\hat{w}, \xi, \lambda) = \frac{1}{2}w^2 + C\sum_{\kappa=1}^{\ell}\xi_\kappa - \sum_{\kappa=1}^{\ell}(y_\kappa f(x_\kappa) - 1 + \xi_\kappa)\lambda_\kappa - \sum_{\kappa=1}^{\ell}\mu_\kappa\xi_\kappa.$$

If we impose $\nabla\mathcal{L}(\hat{w}, \xi, \lambda) = 0$ then we have

$$\partial_w\mathcal{L} = 0 \Rightarrow w - \sum_{\kappa=1}^{\ell}\lambda_\kappa y_\kappa \phi(x_\kappa) = 0$$

$$\partial_b\mathcal{L} = 0 \Rightarrow \sum_{\kappa=1}^{\ell}\lambda_\kappa y_\kappa = 0$$

$$\partial_{\xi_\kappa}\mathcal{L} = 0 \Rightarrow C - \lambda_\kappa - \mu_\kappa = 0.$$

# MAXIMUM MARGIN PROBLEM

Now, the last condition make it possible to re-write the Lagrangian as

$$\mathcal{L}(\hat{w}, \xi, \lambda, \mu) = \frac{1}{2}w^2 - \sum_{\kappa=1}^{\ell} \lambda_\kappa (y_\kappa \hat{w}' \hat{\phi}(x_\kappa) - 1) + \sum_{\kappa=1}^{\ell} (C - \lambda_\kappa - \mu_\kappa)\xi_\kappa$$

$$= \frac{1}{2}w^2 - \sum_{\kappa=1}^{\ell} \lambda_\kappa (y_\kappa \hat{w}' \hat{\phi}(x_\kappa) - 1).$$

It is the same Lagrangian as the one of the primal formulation of MMP in case of hard constraints (3).

# MAXIMUM MARGIN PROBLEM

If we replace $\hat{w}$ into $\mathcal{L}(\hat{w}, \xi, \lambda)$, we obtain the dual problem:

$$\begin{cases} \max \sum_{\kappa=1}^{\ell} \lambda_\kappa - \frac{1}{2} \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} \lambda_h \lambda_\kappa y_h y_\kappa k(x_h, x_\kappa) \\ 0 \leq \lambda_\kappa \leq C, \quad \kappa = 1, \dots, \ell \\ \sum_{\kappa=1}^{\ell} \lambda_\kappa y_\kappa = 0. \end{cases}$$

As $C \to \infty$ this soft-constrains problem is turned into the correspondent hard formulation (4).

# MAXIMUM MARGIN PROBLEM



We have pairs $(x_\kappa, y_\kappa)$ where $y_\kappa \in \mathbb{R}$.
Let $\epsilon > 0$ be and consider the constraint $|y_\kappa - f(x_\kappa)| \leq \epsilon$.
Like for classification, we can introduce slack variables.

We formulate the regression problem as

$$\begin{cases} \min \frac{1}{2} w^2 + C \sum_{\kappa=1}^{\ell} (\xi_\kappa^- + \xi_\kappa^+) \\ [y_\kappa - f(x_\kappa) \geq 0](y_\kappa - f(x_\kappa) \leq \epsilon + \xi_\kappa^+) \\ + [f(x_\kappa) - y_\kappa < 0](f(x_\kappa) - y_\kappa \leq \epsilon + \xi_\kappa^-) \\ \xi_\kappa^+ \geq 0, \quad \xi_\kappa^- \geq 0. \end{cases}$$

# MAXIMUM MARGIN PROBLEM

The Lagrangian is

$$\mathcal{L} = \frac{1}{2}w^2 + C\sum_{\kappa=1}^{\ell}(\xi_\kappa^- + \xi_\kappa^+) + \sum_{\kappa=1}^{\ell}\lambda_\kappa^+(y_\kappa - \hat{w}'\hat{\phi}(x_\kappa) - \epsilon - \xi_\kappa^+)$$
$$+ \sum_{\kappa=1}^{\ell}\lambda_\kappa^-(\hat{w}'\hat{\phi}(x_\kappa) - y_\kappa - \epsilon - \xi_\kappa^-) - \sum_{\kappa=1}^{\ell}\mu_\kappa^+\xi_\kappa^+ - \sum_{\kappa=1}^{\ell}\mu_\kappa^-\xi_\kappa^-.$$

In order to pass to the dual space we determine the critical points

$$\partial_w\mathcal{L} = 0 \Rightarrow w - \sum_{\kappa=1}^{\ell}(\lambda_\kappa^+ - \lambda_\kappa^-)\hat{\phi}(x_\kappa) = 0$$

$$\partial_b\mathcal{L} = 0 \Rightarrow \sum_{\kappa=1}^{\ell}(\lambda_\kappa^+ - \lambda_\kappa^-) = 0$$

$$\partial_{\xi_\kappa^+}\mathcal{L} = 0 \Rightarrow C - \lambda_\kappa^+ - \mu_\kappa^+ = 0$$

$$\partial_{\xi_\kappa^-}\mathcal{L} = 0 \Rightarrow C - \lambda_\kappa^- - \mu_\kappa^- = 0.$$

# MAXIMUM MARGIN PROBLEM

We obtain the following dual problem

$$\begin{cases} \max \theta(\lambda^+, \lambda^-) = -\frac{1}{2} \sum_{h=1}^{\ell} \sum_{\kappa=1}^{\ell} (\lambda_h^+ - \lambda_h^-)(\lambda_\kappa^+ - \lambda_\kappa^-) k(x_h, x_\kappa) \\ \qquad - \epsilon \sum_{\kappa=1}^{\ell} (\lambda_\kappa^+ + \lambda_\kappa^-) + \sum_{\kappa=1}^{\ell} y_\kappa (\lambda_\kappa^+ - \lambda_\kappa^-) \\ \qquad \sum_{\kappa=1}^{\ell} \lambda_\kappa^+ = \sum_{\kappa=1}^{\ell} \lambda_\kappa^- \\ \qquad 0 \leq \lambda_\kappa^+ \leq C \\ \qquad 0 \leq \lambda_\kappa^- \leq C \end{cases}$$

where $k(x_h, x_\kappa) = \langle \hat{\phi}(x_h), \hat{\phi}(x_\kappa) \rangle$.

# KERNEL FUNCTIONS

We have already seen the definition of **kernel**:

$$k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$$

$$k(x, z) = \langle \phi(x), \phi(z) \rangle = \phi'(x)\phi(z).$$

*Kernel trick*: kernel functions return a similarity measure between any two points in the input space which is based on their mapping to the feature space, without its direct involvement in the computation.

We have $\mathscr{X} = \mathbb{R}^2$, $\mathscr{H} = \mathbb{R}^3$ and

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \xrightarrow{\phi} \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}.$$

$$\begin{aligned} k(x_h, x_\kappa) &= (x_{h1}^2, \sqrt{2}x_{h1}x_{h2}, x_{h2}^2) \cdot \begin{pmatrix} x_{\kappa 1}^2 \\ \sqrt{2}x_{\kappa 1}x_{\kappa 2} \\ x_{\kappa 2}^2 \end{pmatrix} \\ &= x_{h1}^2 x_{\kappa 1}^2 + 2x_{h1}x_{h2}x_{\kappa 1}x_{\kappa 2} + x_{h,2}^2 x_{\kappa,2}^2 \\ &= (x_{h1}x_{\kappa 1} + x_{h2}x_{\kappa 2})^2 \\ &= \langle x_h, x_k \rangle^2. \end{aligned}$$

# KERNEL FUNCTIONS

Now we define Gram matrix

$$K(\mathscr{X}_\ell^\sharp) = \begin{pmatrix} k(x_1, x_1), & \dots & k(x_1, x_\ell) \\ \vdots & & \vdots \\ k(x_\ell, x_1), & \dots & k(x_\ell, x_\ell) \end{pmatrix} \in \mathbb{R}^{\ell,\ell} \tag{6}$$

which is a structured organization of the image of $k$ over a sampling $\mathscr{X}_\ell^\sharp = \{x_1, x_2, \dots, x_\ell\}$ of $\mathscr{X}$. It allows us to replace functional analysis on the kernel with linear algebra on the associated Gram matrix.

$K(\mathscr{X}_\ell^\sharp) \geq 0 \quad \forall \mathscr{X}_\ell^\sharp$ (i.e. is a non-negative matrix) $\Leftrightarrow k$ is a kernel

# KERNEL FUNCTIONS

## Infinite dimensional feature spaces

As $D \to \infty$ the feature vector is

$$\phi(x) = (\phi_1(x), \phi_2(x), \ldots)' \in \mathbb{R}^\infty.$$

We introduce the functional operator

$$\mathcal{T}_k u(x) = \int_{\mathscr{X}} k(x, z) u(z) dz$$

which replaces the Gram matrix at finite dimension.

$\mathcal{T}_k \geq 0 \;\; \Leftrightarrow k$ is a kernel

# KERNEL FUNCTIONS

## Types of kernels

- Linear kernels: $k(x,z) = x'z$ $(\phi = id)$
- Polynomial kernels: $k(x,z) = (x'z)^p$
  Let be $x, z \in \mathbb{R}^d$.

$$\langle x, z\rangle^p = \left(\sum_{i=1}^{d} x_i z_i\right)^p = \sum_{|\alpha|=p} \frac{p!}{\alpha!}(x \circ z)^\alpha = \sum_{|\alpha|=p} \frac{p!}{\alpha!}\prod_{i=1}^{d}(x_i z_i)^{\alpha_i}$$

$$= \sum_{|\alpha|=p} \left(\frac{p!}{\alpha!}\right)^{1/2} \prod_{i=1}^{d}(x_i)^{\alpha_i} \cdot \left(\frac{p!}{\alpha!}\right)^{1/2} \prod_{i=1}^{d}(z_i)^{\alpha_i}$$

$$= \left\langle \left(\frac{p!}{\alpha!}\right)^{1/2} \prod_{i=1}^{d}(x_i)^{\alpha_i}, \left(\frac{p!}{\alpha!}\right)^{1/2} \prod_{i=1}^{d}(z_i)^{\alpha_i} \right\rangle_{|\alpha|=p}.$$

The feature vector is $\phi(u) = \left(\frac{p!}{\alpha!}\right)^{1/2} \prod_{i=1}^{d}(u_i)^{\alpha_i}$.

$(\alpha = [\alpha_1, \ldots, \alpha_d]$, $|\alpha| = \alpha_1 + \ldots + \alpha_d$, $\alpha! = \alpha_1! \ldots \alpha_d!$,
$x^\alpha = x_1^{\alpha_1} \ldots x_d^{\alpha_d})$

- Gaussian kernel: $k(x,z) = e^{-\frac{\|x-z\|^2}{2\sigma^2}}$
  Only infinite-dimensional feature representations are known.
  We propose one of those representations by using Taylor's expansion in the case of $\mathscr{X} = \mathbb{R}$. We have
  $e^{-\gamma(x-z)^2} = e^{-\gamma x^2+2\gamma xz-\gamma z^2} =$
  $e^{-\gamma x^2-\gamma z^2}\left(1+\frac{2\gamma xz}{1!}+\frac{(2\gamma xz)^2}{2!}+\dots\right) =$
  $e^{-\gamma x^2-\gamma z^2}\left(1\cdot 1+\frac{\sqrt{2\gamma}}{1!}x\cdot\frac{\sqrt{2\gamma}}{1!}z+\sqrt{\frac{(2\gamma)^2}{2!}}x^2\cdot\sqrt{\frac{(2\gamma)^2}{2!}}z^2+\dots\right).$
  The feature map is
  $\phi(y) = e^{-\gamma y^2}\left(1, \sqrt{\frac{2\gamma}{1!}}y, \sqrt{\frac{(2\gamma)^2}{2!}}y^2, \dots \sqrt{\frac{(2\gamma)^i}{i!}}y^i, \dots\right)'.$

# KERNEL FUNCTIONS

## Types of kernels

- Dot product kernels: $k(x,z) = K(\langle x, z\rangle)$.
  Linear and polynomial kernels are dot product kernels.
- Translation invariance kernels: $k(x,z) = K(x - z)$.
  Gaussian kernels are translation invariance kernels.
- Radial kernel: $k(x,z) = K(\|x - z\|)$.
- $B_n$-splines kernels: $k(x,z) = B_{2p+1}(\|x - z\|)$,

  where $B_n(u) := \bigotimes_{i=1}^{n} \left[|u| \le \frac{1}{2}\right]$ and $\bigotimes_{i=1}^{n}$ is the $n$-fold convolution of the characteristic function of the interval $[-\frac{1}{2}, \frac{1}{2}]$, and $\bigotimes_{i=1}^{0} \left[|u| \le \frac{1}{2}\right] := \left[|u| \le \frac{1}{2}\right]$.

  $B_n$-splines kernels are an example of translational invariance kernels.
  $B_n$-splines kernels approximate Gaussian kernels as $n \to \infty$.

Given $\alpha \in \mathbb{R}$, any two kernels $k_1, k_2$, and $f \colon \mathscr{X} \to \mathbb{R}$ then

- $k(x, z) = k_1(x, z) + k_2(x, z)$
- $k(x, z) = \alpha k_1(x, z)$
- $k(x, z) = k_1(x, z) \cdot k_2(x, z)$
- $k(x, z) = f(x)f(z)$