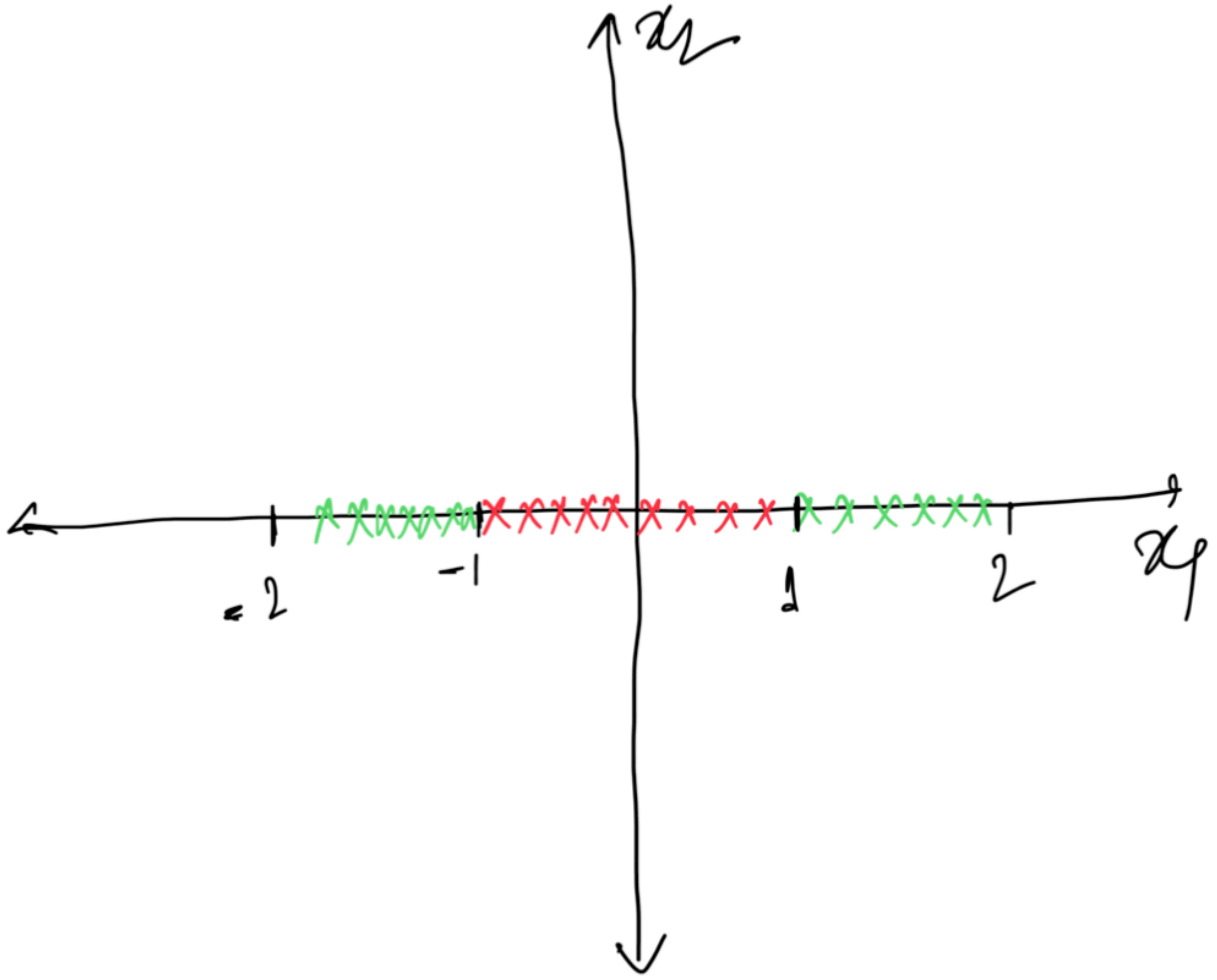
Kernel methods

Kernel methods

- Data is not linearly separable!



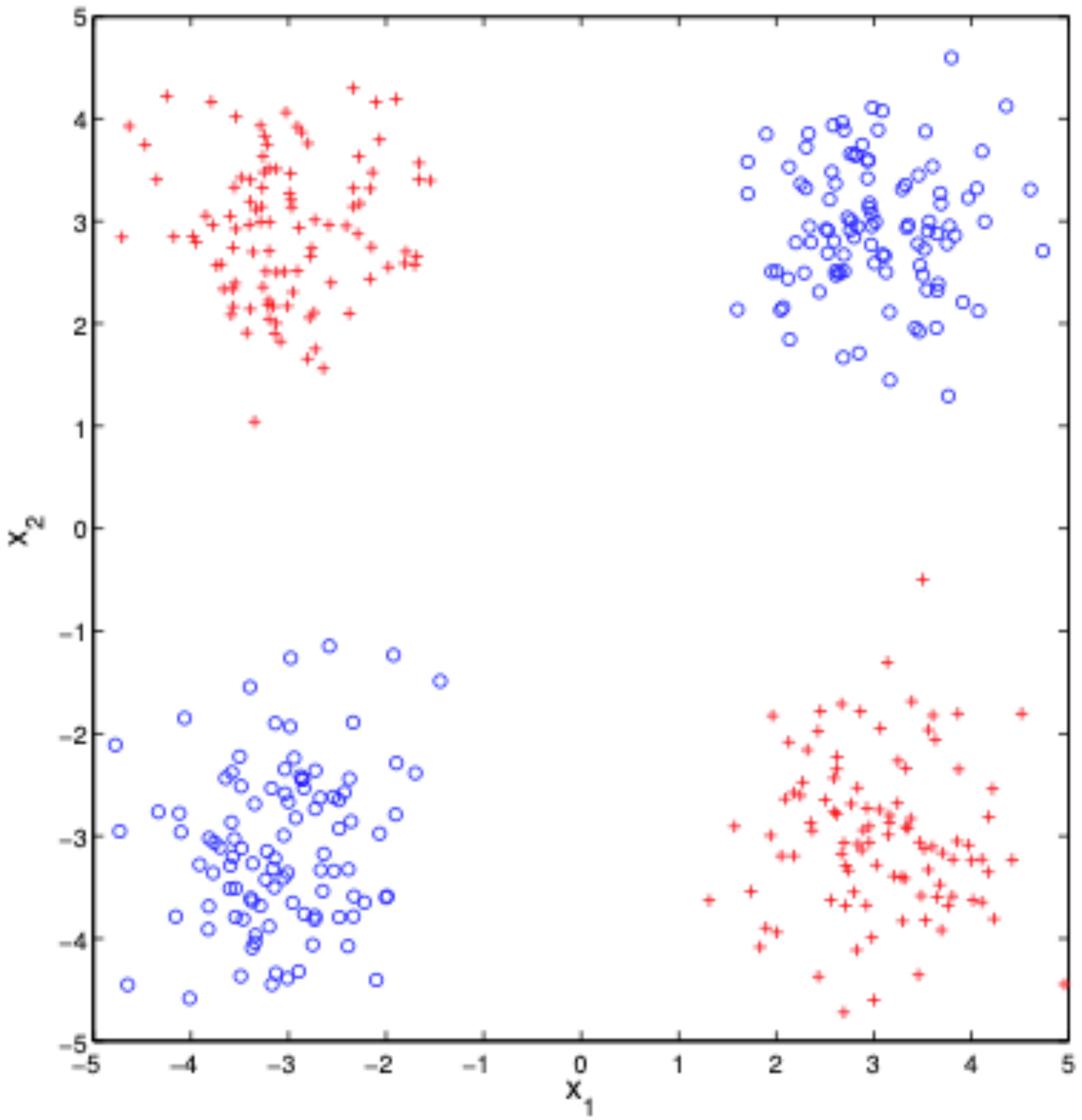
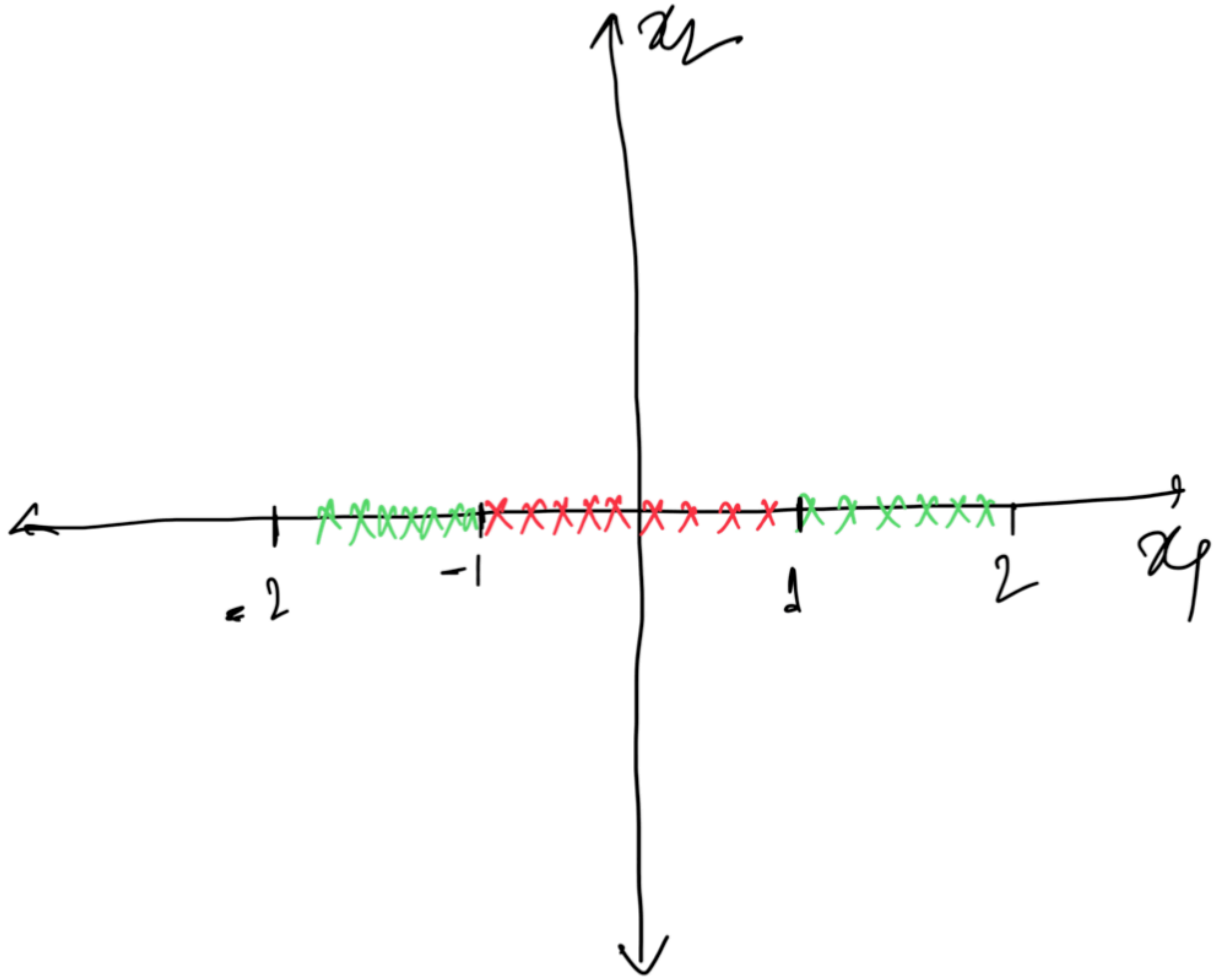


Image: Arthur Gretton's tutorial , 2019

Kernel methods (cont…)

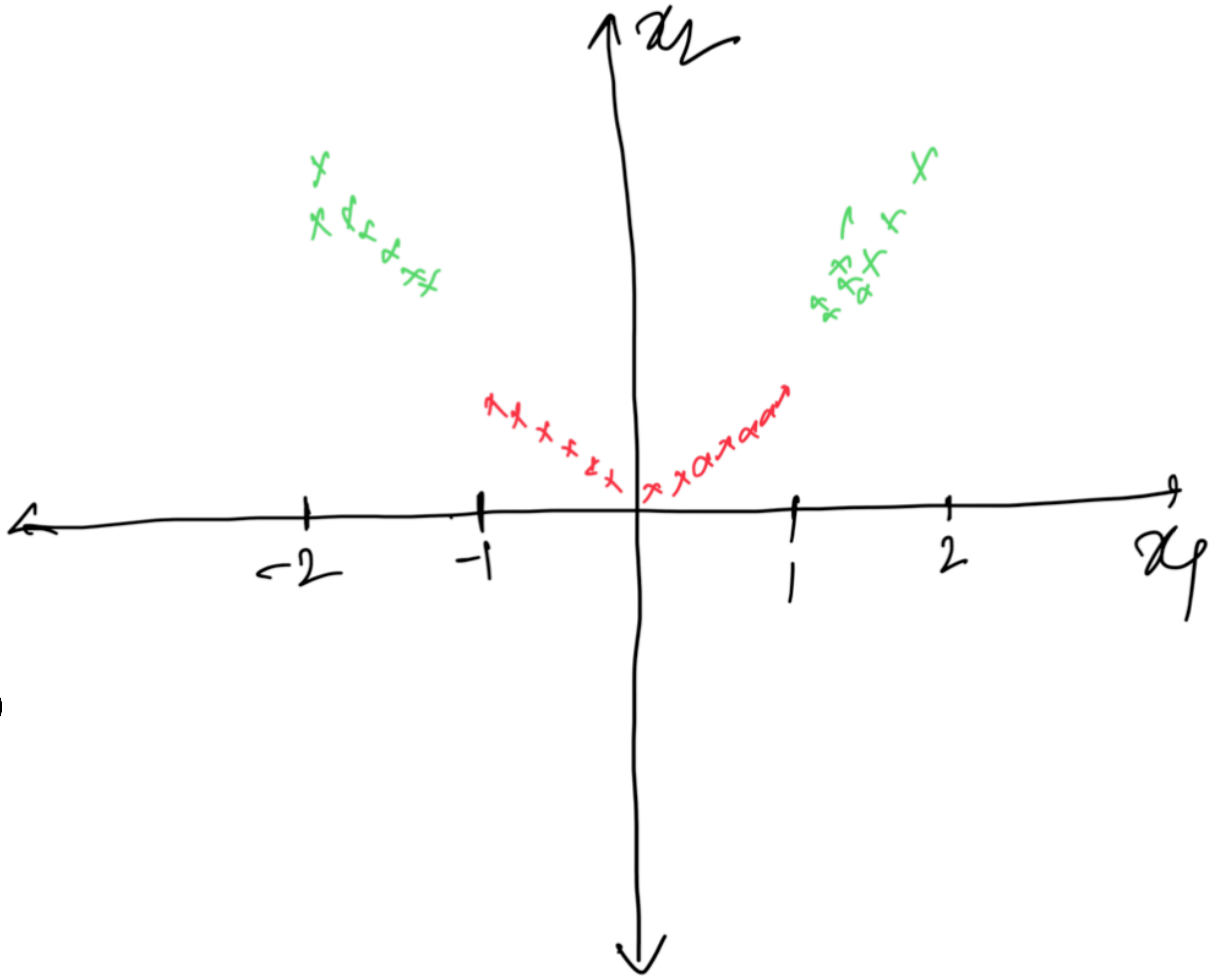
- Data is not linearly separable!
- Transform the data points to other space: map $\phi : \mathscr{X} \rightarrow \mathscr{H}$



=>

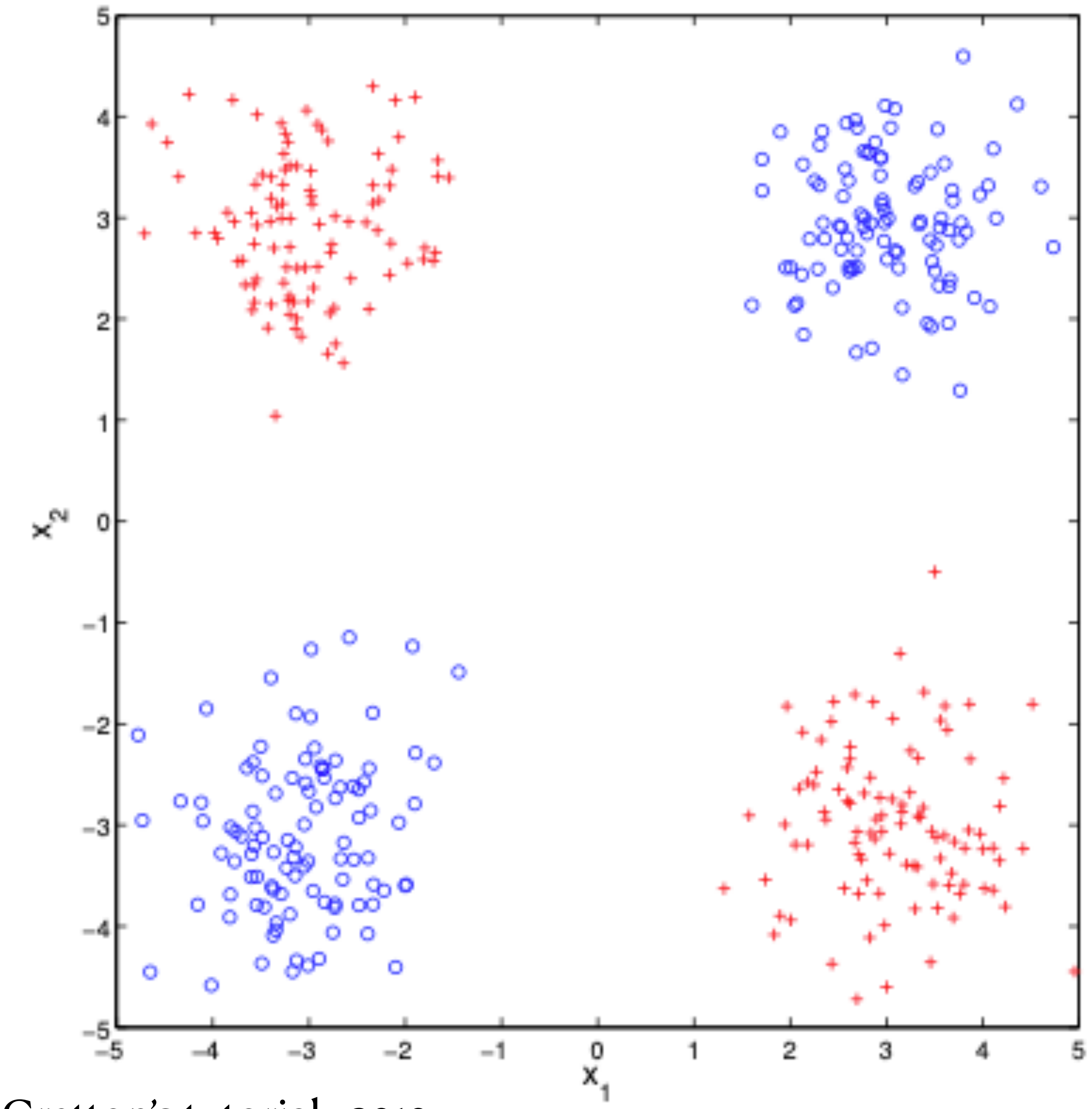
?

$(x_1) => (x_1, x_1^2)$



Kernel methods (cont…)

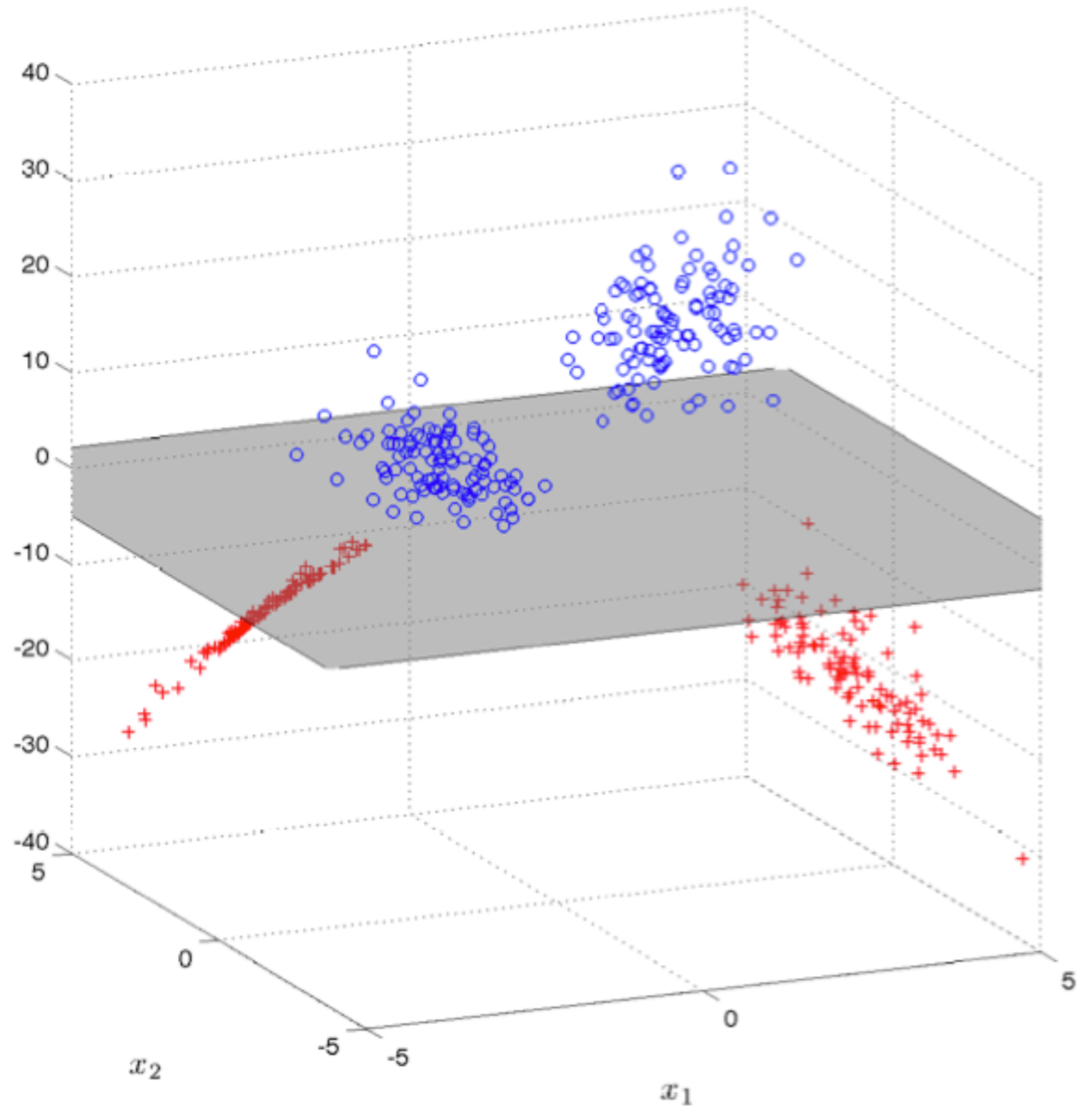
- Data is not linearly separable!
- Transform the data points to other space: map $\phi : \mathscr{X} \rightarrow \mathscr{H}$



=>

?

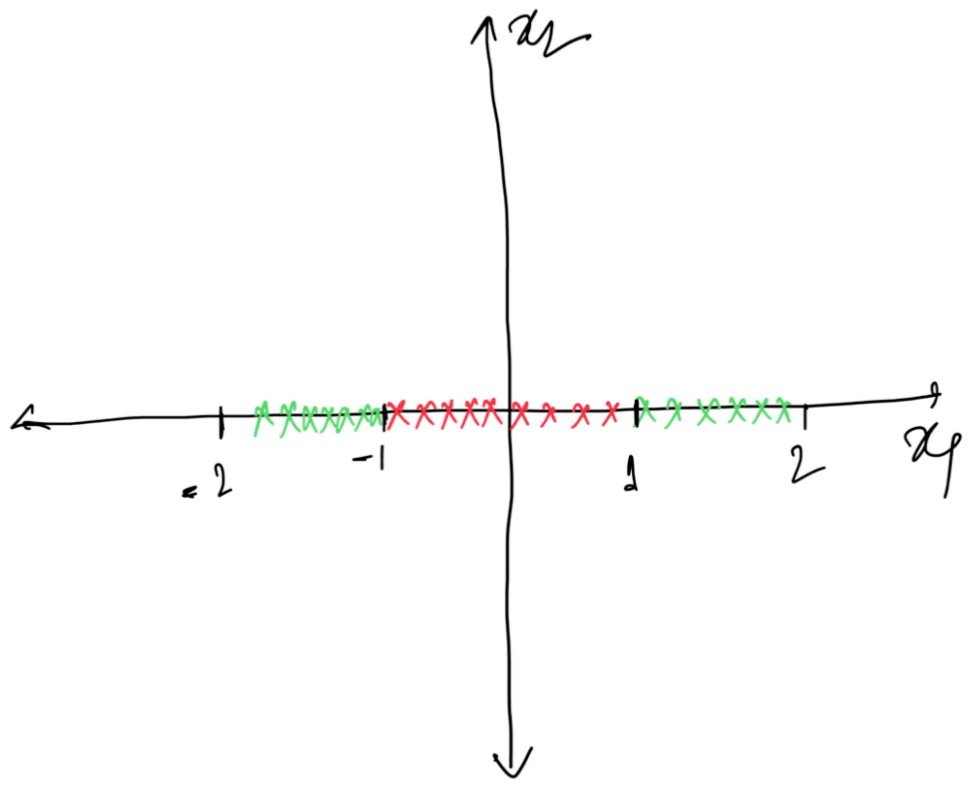
$(x_1, x_2) => (x_1, x_2, x_1x_2)$



Images: Arthur Gretton's tutorial , 2019

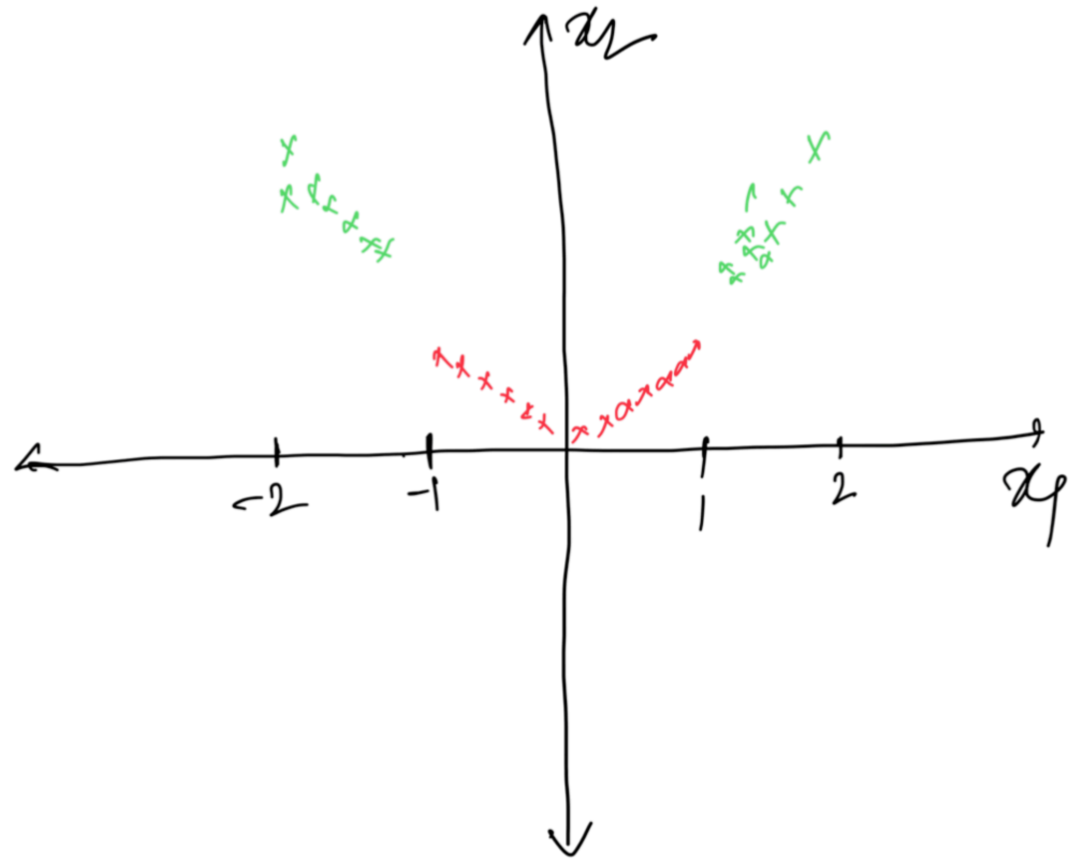
Kernel methods (cont…)

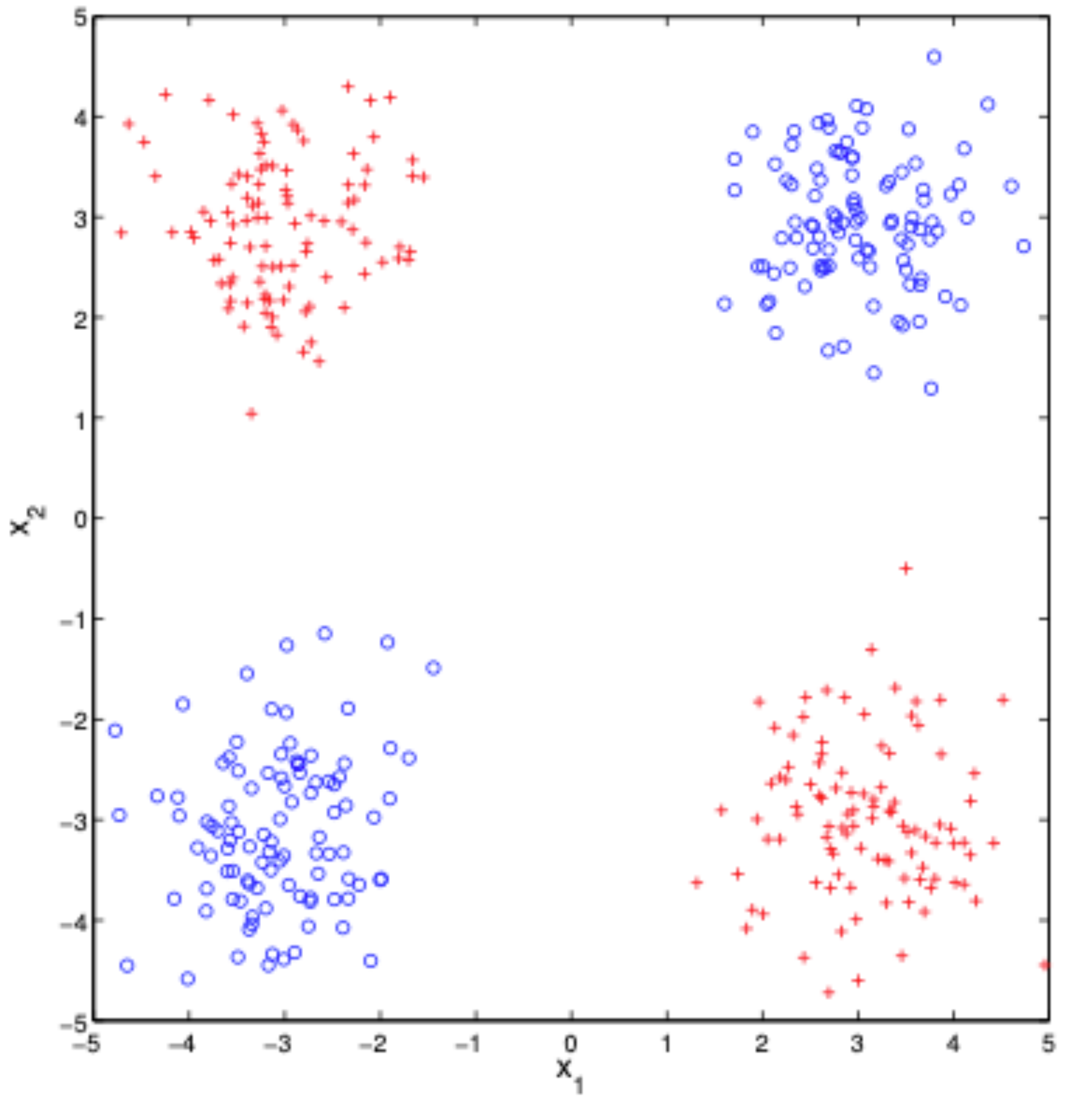
- Mapping is fine, then what is the use of kernel?
- For mapping we have to explicitly define ϕ!
- Can we do the same thing without ϕ (explicitly) ?
  - Yes and that is the role of kernel
- How?
  - $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$



=>

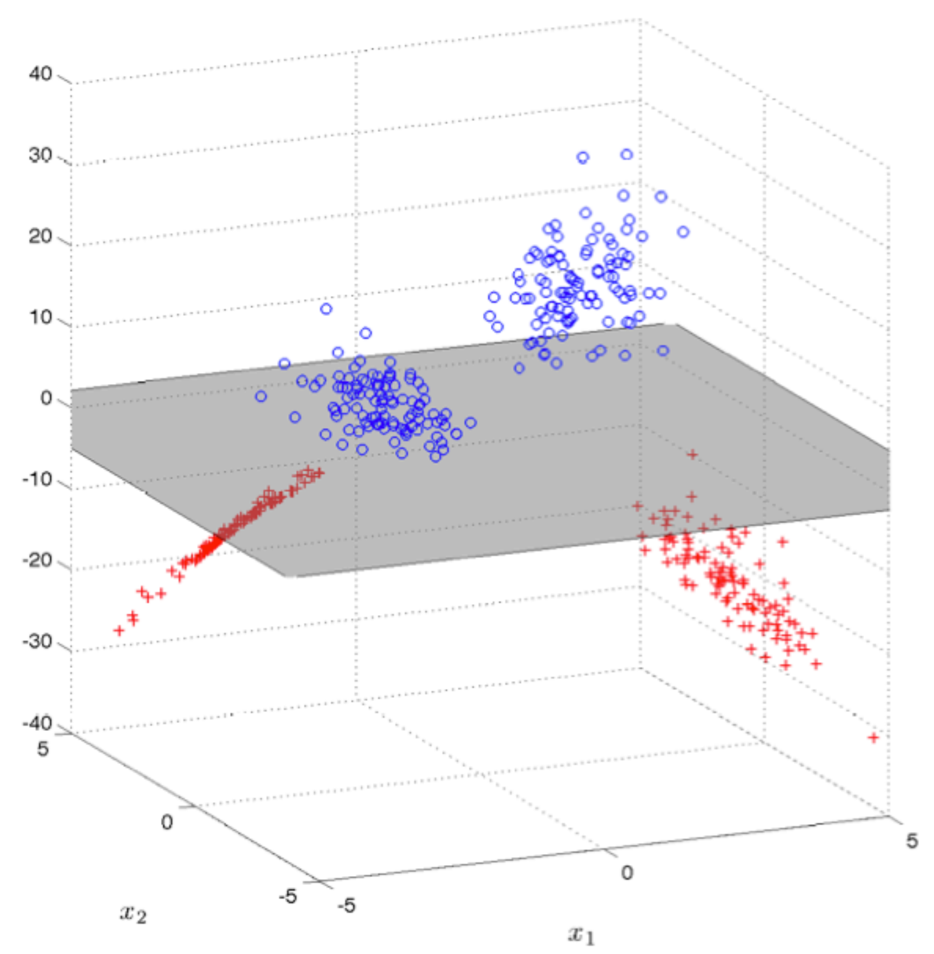
$(x_1) => (x_1, x_1^2)$





=>

$(x_1, x_2) => (x_1, x_2, x_1 x_2)$



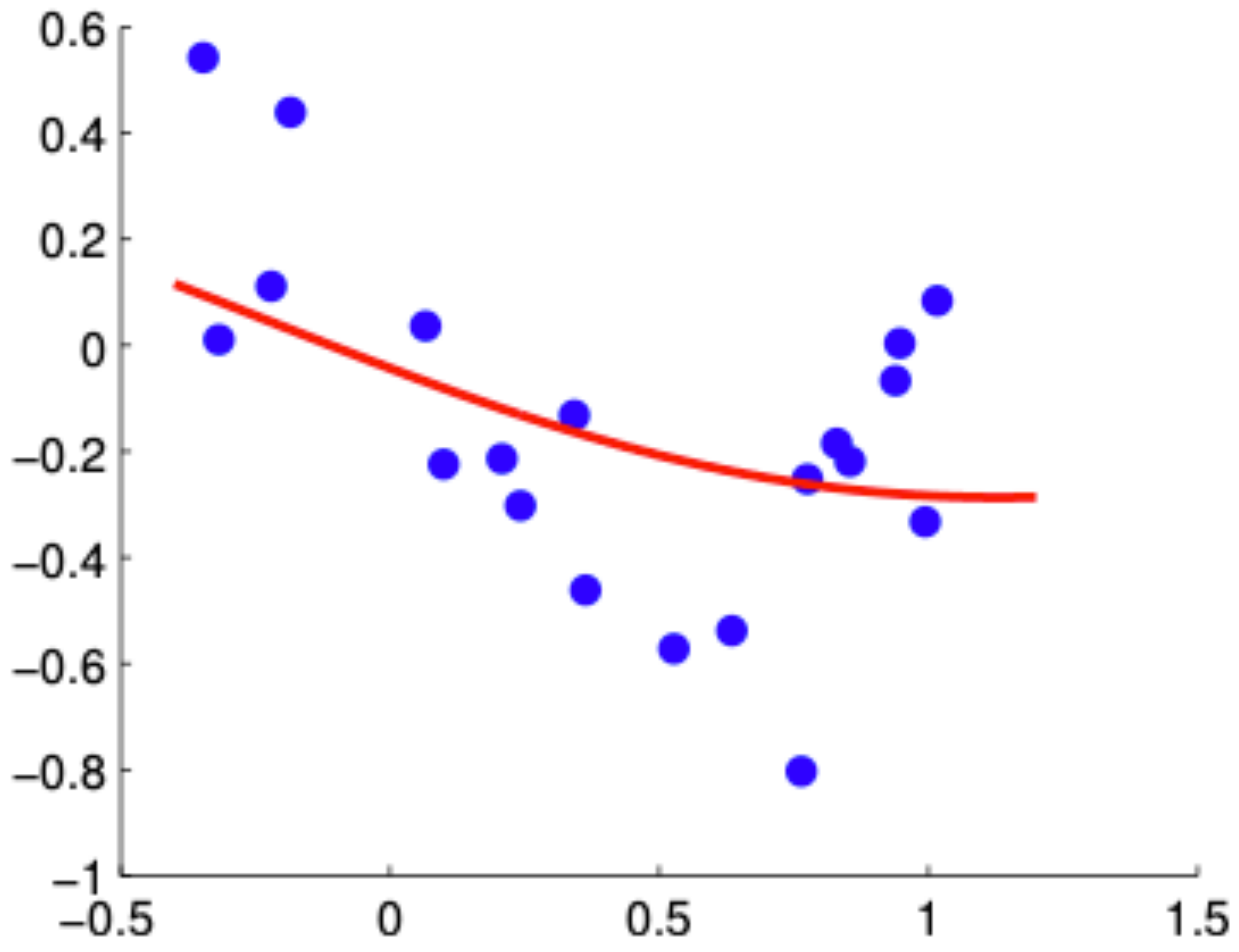
Images (bottom): Arthur Gretton's tutorial , 2019

Kernel methods (cont…)

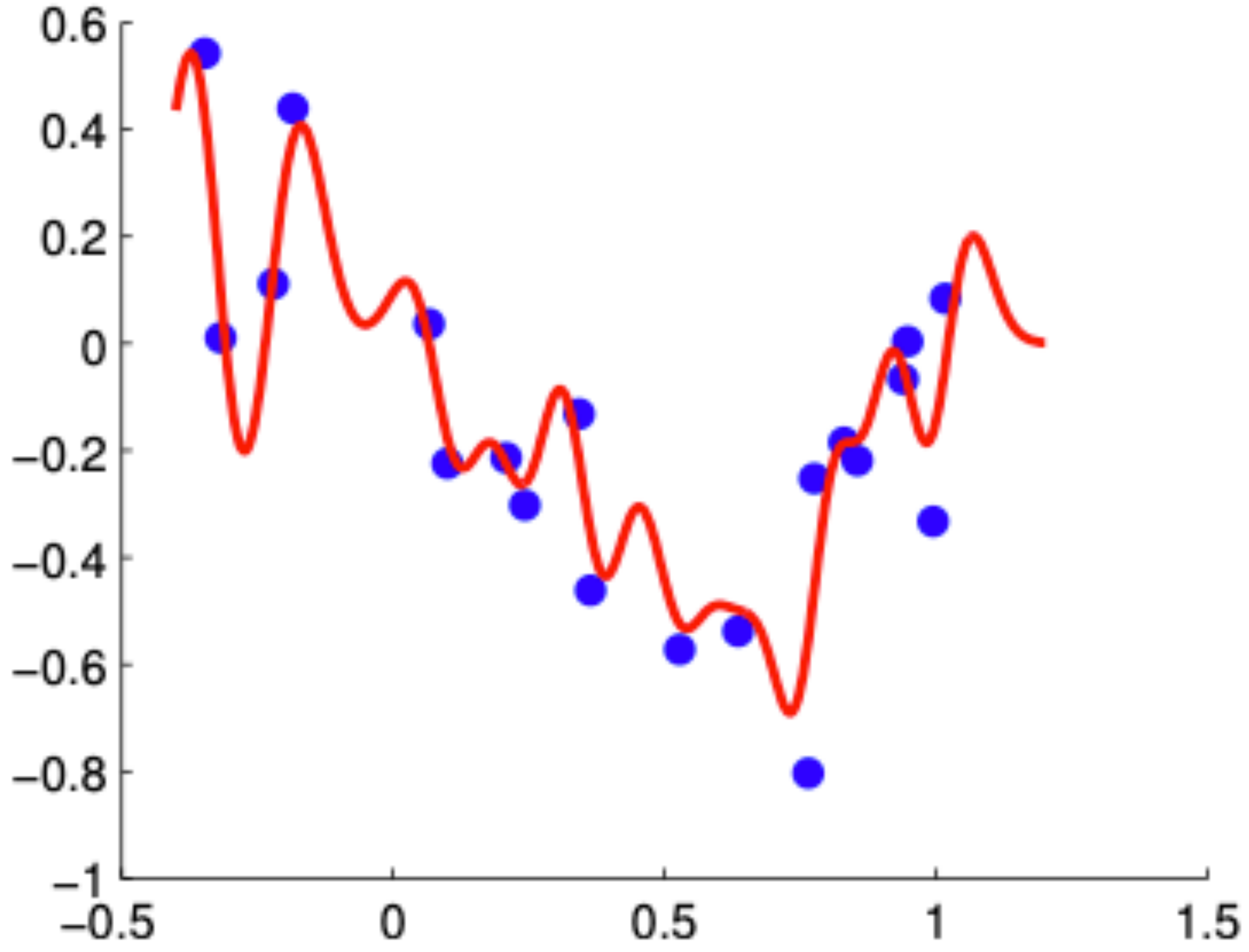
- What is kernel?
  - $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$
  - Dot product (loosely speaking)
  - When you have dot product in your learning algorithm: you can use kernel tricks
- In our SVM (hard) settings:
  - $\max_{\lambda} \left\{ \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j Y_i Y_j X_i^T X_j \right\}$
  - Subject to $\sum_{i=1}^{n} \lambda_i Y_i = 0$ and $\lambda_i \geq 0, i = 1 \cdots n$
- $w = \sum_{i=1}^{n} \lambda_i Y_i X_i$ and $b = Y_i - W^T X_i = Y_i - \sum_{j=1}^{n} \lambda_j Y_j X_j^T X_i$
- For the test data X
  - $\text{sgn}(W^T X + b) = \text{sgn}(\sum_{i=1}^{n} \lambda_i Y_i X_i^T X + b)$

Kernel methods (cont…)

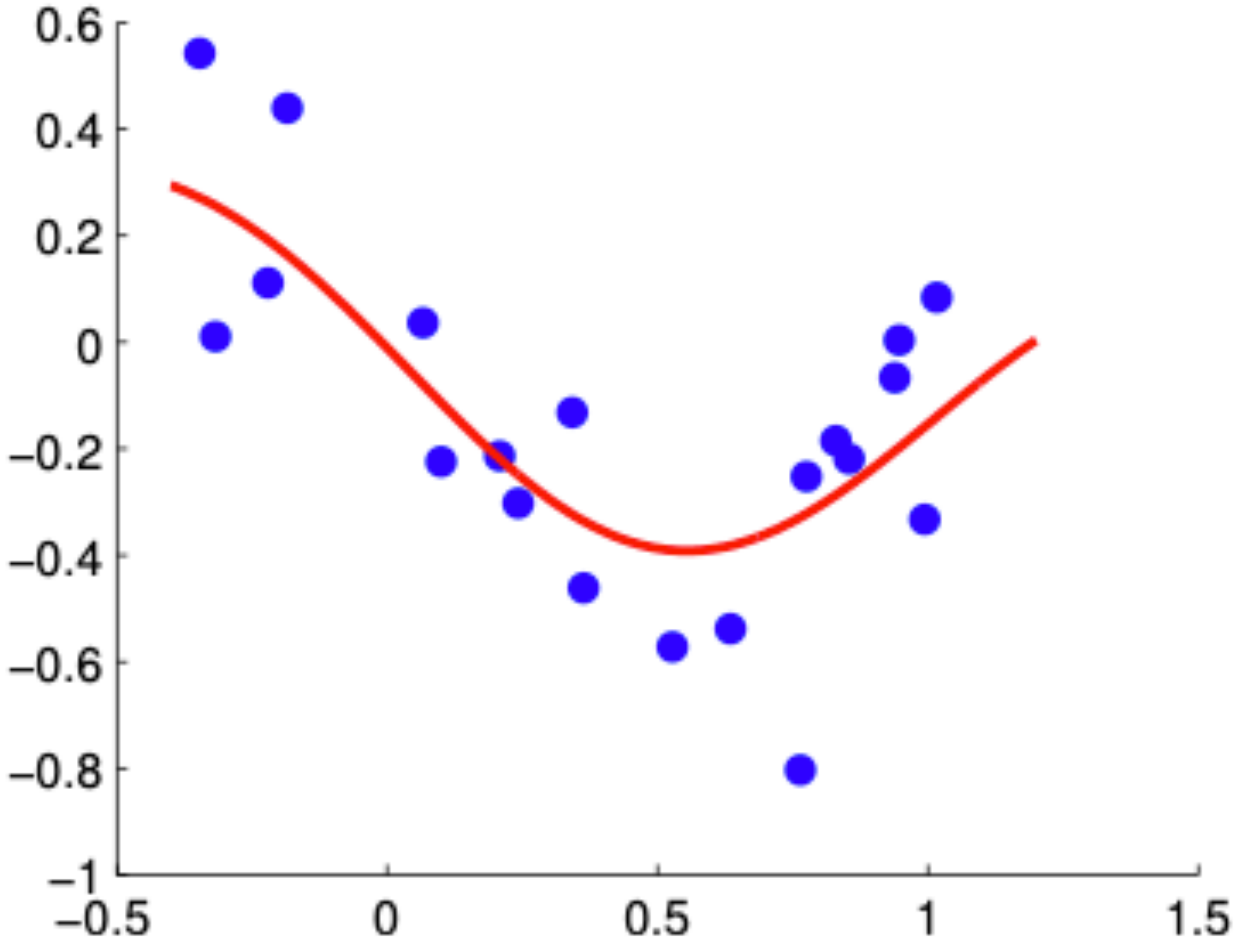
- Data is not filled well!



Under fitted



Over fitted



Good fitted

Image: Arthur Gretton's tutorial , 2019

What is kernel?

- Def-1: Let $\mathscr{X}$ be any space. A symmetric function $k : \mathscr{X} \times \mathscr{X} \to R$ is called a kernel function if for all $n \geq 1, X_1, X_2, \ldots, X_n \in \mathscr{X}$ and $c_1, c_2, \ldots, c_n \in R$ we have
$$\sum_{i,j=1}^{n} c_i c_j k(X_i, X_j) \geq 0$$
    - $C^T K C \geq 0$
    - Symmetric
    - Positive semi-definite (PSD)
- Def-2: Let $\mathscr{X}$ be a non-empty set. A function $k : \mathscr{X} \times \mathscr{X} \to R$ is called a kernel function if there exists an R-Hilbert Space and a mapping $\phi : \mathscr{X} \to \mathscr{H}$ such that $\forall X_i, X_j \in \mathscr{X}$
$$k(X_i, X_j) := \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$$
    - R-Hilbert Space: reproducing kernel Hilbert space (RKHS)
    - Hilbert space: Complete Inner product space (cocktail party definition)
- In Def-2, PSD comes automatically ?

Hilbert space

- Def-2: Let $\mathscr{X}$ be a non-empty set. A function $k : \mathscr{X} \times \mathscr{X} \rightarrow R$ is called a kernel function if there exists an R-Hilbert Space and a mapping $\phi : \mathscr{X} \rightarrow \mathscr{H}$ such that $\forall X_i, X_j \in \mathscr{X}\ k(X_i, X_j) := \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$
  - R-Hilbert Space: reproducing kernel Hilbert space (RKHS)
  - Hilbert space: Complete Inner product space (cocktail party definition)
- Vector space
  - Can you recall the axioms over a field say R (in our setting) ?
- Inner product
  - Map from $V \times V$ to R
  - Can you recall the three conditions?
    - $\langle x_1, x_2 \rangle = \overline{\langle x_2, x_1 \rangle}$
    - $\langle \alpha_1 x_1 + \alpha_2 x_2, x_3 \rangle = \alpha_1 \langle x_1, x_3 \rangle + \alpha_2 \langle x_2, x_3 \rangle$
    - $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ iff $x = 0$
- Norm using inner product: Map $V \rightarrow R$
  - $\|x\|_V = \sqrt{\langle x, x \rangle}$

Hilbert space (cont…)

- Vector space
  - Can you recall the axioms over a field say R (in our setting) ?
- Inner product
  - Map from $V \times V$ to R
  - Can you recall the three conditions?
    - $\langle x_1, x_2 \rangle = \overline{\langle x_2, x_1 \rangle}$
    - $\langle \alpha_1 x_1 + \alpha_2 x_2, x_3 \rangle = \alpha_1 \langle x_1, x_3 \rangle + \alpha_2 \langle x_2, x_3 \rangle$
    - $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ iff $x = 0$
- Norm using inner product: Map $V \to R$
  - $\|x\|_V = \sqrt{\langle x, x \rangle}$
- Normed space := vector space + Norm
- Cauchy sequence: A sequence $\left\{ x_i \right\}_{i=1}^{n}$ of elements in a normed space $\mathscr{H}$ is said to be a Cauchy sequence if for every $\epsilon > 0$, there exists a positive integer N such that for all $m, n \geq N$ if $\|x_n - x_m\|_{\mathscr{H}} < \epsilon$

Some standard kernels

- Linear kernel
  - $k(X_i, X_j) := X_i^T X_j$
- Polynomial kernel
  - $k(X_i, X_j) := (X_i^T X_j + c)^p$
- Gaussian kernel (RBF- radial basis function):
  - $k(X_i, X_j) := \exp^{\left\{ -\frac{\|X_i - X_j\|^2}{2\sigma^2} \right\}}$

Operations on kernels

- Let $k_1, k_2 : \mathscr{X} \times \mathscr{X} \to R$ are two kernel functions, $X_i, X_j \in \mathscr{X}$ and $f : \mathscr{X} \to R$ be any function, then
  - Is $\lambda \times k_1$ for some $\lambda > 0$ a kernel ?
    - Yes
  - Is $k_1 + k_2$ a kernel ?
    - Yes
  - Is $k_1 \times k_2$ a kernel ?
    - Yes
  - Is $f(X_i)k_1(X_i, X_j)f(X_j)$ a kernel?
    - Yes
    - Particularly $f(X_i)f(X_j)$ is a kernel
- Why we need these?
  - We can define new kernel using well known kernels
  - Useful to proof a function is a kernel or not?

13-04-2024

Kernel as similarity functions

- Linear kernel
  - $k(X_i, X_j) := X_i^T X_j$
    - Measure the similarity between X_i and X_j
    - Cosine similarity?
- What about the Gaussian kernel (RBF- radial basis function) then?
  - $k(X_i, X_j) := \exp\left\{ -\frac{\|X_i - X_j\|^2}{2\sigma^2} \right\}$ measure the similarity here?

Reproducing Kernel Hilbert space (RKHS)

- Theorem (Kernel implies embedding):
  - A function $k : \mathscr{X} \times \mathscr{X} \to R$ is a kernel if and only if there exists a Hilbert space $\mathscr{H}$ and a map $\phi : \mathscr{X} \to \mathscr{H}$ such that $k(X_i, X_j) := \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$
- Proof:
  - If part ($\Leftarrow$):
    - Given Hilbert space $\mathscr{H}$ and a map $\phi : \mathscr{X} \to \mathscr{H}$ such that $k(X_i, X_j) := \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$
    - Its our Def-2 ?
  - Only if part ($\Rightarrow$)
    - Given $\mathscr{X}$ and k, there exists a Hilbert space $\mathscr{H}$ and a map $\phi : \mathscr{X} \to \mathscr{H}$ such that $k(X_i, X_j) := \langle \phi(X_i), \phi(X_j) \rangle_{\mathscr{H}}$

Representer theorem

- Given $X_1, X_2, \ldots, X_n, X_i \in \mathscr{X}$ are n data points and $Y_1, Y_2, \ldots, Y_n, Y_i \in R$ are their corresponding outputs, $k : \mathscr{X} \times \mathscr{X} \to R$ is kernel on $\mathscr{X}$ with a corresponding reproducing kernel Hilbert space $\mathscr{H}$. Consider a regularised risk minimisation problem of the form:

$$\min_{w \in \mathscr{H}} E\{W, (X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\} + \lambda\Omega(\|W\|_{\mathscr{H}})$$

Where E is any arbitrary error/loss function and $\Omega : [0,\infty) \to R$ is a strictly monotonic increasing function.

Then the above optimization problem has always has an optimal solution of the form

$W^* = \sum_{i=1}^{n} \alpha_i k(X_i, :)$, where $\alpha_i \in R$ for all $1 \leq i \leq n$

Kernel PCA

- Gram matrix: $M = X^T X$
  - $$M(i,j) = X_i^T X_j$$ $$= k(X_i, K_j)$$

Kernel k-means

- In our k-means algorithm we have calculate the distance between X_i and the cluster centres $\bar{X}_k$
  - $d(X_i, \bar{X}_k) = \|X_i - \bar{X}_k\|^2$
- Replace Euclidean distance calculation by the kernelled version
  - $$\begin{aligned} \|X_i - \bar{X}_k\|^2 &= \|\phi(X_i) - \phi(\bar{X}_k)\|^2 \\ &= k(X_i, X_i) + k(\bar{X}_k, \bar{X}_k) - 2k(X_i, \bar{X}_k) \end{aligned}$$

Kernel regression

- In LLSR: $Y = W^T X$
  - $W = (XX^T)^{-1} X Y^T$
- Ridge LLSR:
  - $W = (XX^T + \lambda I)^{-1} X Y^T$