Dot-product kernels

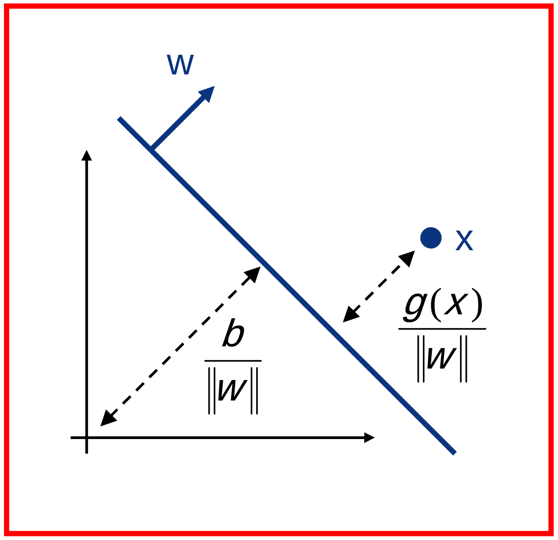
Nuno Vasconcelos

ECE Department, UCSD

Classification

- a classification problem has two types of variables
  - X - vector of observations (features) in the world
  - Y - state (class) of the world
- Perceptron: classifier implements the linear decision rule

$$\boxed{h(x) = \mathrm{sgn}[g(x)]} \quad \text{with} \quad \boxed{g(x) = w^T x + b}$$



- appropriate when the classes are linearly separable
- to deal with non-linear separability we introduce a kernel

Kernel summary

1. D not linearly separable in $\mathcal{X}$, apply feature transformation $\Phi: \mathcal{X} \rightarrow \mathcal{Z}$, such that $dim(\mathcal{Z}) >> dim(\mathcal{X})$

2. computing $\Phi(x)$ too expensive:
   - write your learning algorithm in dot-product form
   - instead of $\Phi(x_i)$, we only need $\Phi(x_i)^T \Phi(x_j) \ \forall_{ij}$

3. instead of computing $\Phi(x_i)^T \Phi(x_j) \ \forall_{ij}$, define the "dot-product kernel"

$$\boxed{K(x,z) = \Phi(x)^T \Phi(z)}$$

   and compute $K(x_i,x_j) \ \forall_{ij}$ directly
   - note: the matrix

$$K = \begin{bmatrix} & \vdots & \\ \cdots & K(x_i,x_j) & \cdots \\ & \vdots & \end{bmatrix}$$

     is called the "kernel" or Gram matrix

4. forget about $\Phi(x)$ and use $K(x,z)$ from the start!

Polynomial kernels

- this makes a significant difference when $K(x,z)$ is easier to compute that $\Phi(x)^T\Phi(z)$
- e.g., we have seen that

$$K(x,z) = \left(x^T z\right)^2 = \Phi(x)^T \Phi(z)$$

with $\Phi: \quad \Re^{d} \rightarrow \Re^{d^2}$

$$\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \rightarrow \left(x_1 x_1, x_1 x_2, \cdots, x_1 x_d, \cdots, x_d x_1, x_d x_2, \cdots, x_d x_d\right)^T$$

- while $K(x,z)$ has complexity $O(d)$, $\Phi(x)^T\Phi(z)$ is $O(d^2)$
- for $K(x,z) = (x^Tz)^k$ we go from $O(d)$ to $O(d^k)$

Question

- what is a good dot-product kernel?
  - intuitively, a good kernel is one that maximizes the margin γ in range space
  - however, nobody knows how to do this effectively
- in practice:
  - pick a kernel from a library of known kernels
  - we talked about
    - the linear kernel $K(x,z) = x^T z$
    - the Gaussian family

$$K(x,z) = e^{-\frac{\|x-z\|^2}{\sigma}}$$

    - the polynomial family

$$K(x,z) = \left(1 + x^T z\right)^k, \quad k \in \{1,2,\cdots\}$$

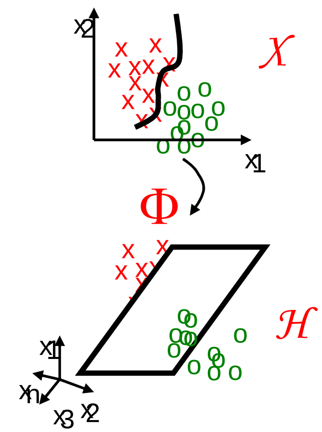
Question

- “this problem of mine is really asking for the kernel $k'(x,z) = \ldots$”
  - how do I know if this is a dot-product kernel?
- let’s start by the definition
- **Definition:** a mapping

$$k: \mathcal{X} \times \mathcal{X} \to \Re$$
$$(x,y) \to k(x,y)$$

is a dot-product kernel if and only if

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle$$



where $\Phi: \mathcal{X} \to \mathcal{H}$, $\mathcal{H}$ is a vector space and, $\langle .,. \rangle$ a dot-product in $\mathcal{H}$

Vector spaces

- note that both $\mathcal{H}$ and <.,.> can be abstract, not necessarily $\Re^d$
- **Definition:** a vector space is a set $\mathcal{H}$ where
  - addition and scalar multiplication are defined and satisfy:

    1) $x+(x'+x'') = (x+x')+x''$
    2) $x+x' = x'+x \in \mathcal{H}$
    3) $0 \in \mathcal{H}$, $0 + x = x$
    4) $-x \in \mathcal{H}$, $-x + x = 0$
    5) $\lambda x \in \mathcal{H}$
    6) $1x = x$
    7) $\lambda(\lambda' x) = (\lambda\lambda')x$
    8) $\lambda(x+x') = \lambda x + \lambda x'$
    9) $(\lambda+\lambda')x = \lambda x + \lambda' x$
- the canonical example is $\Re^d$ with standard vector addition and scalar multiplication
- another example is the space of mappings $\mathcal{X} \rightarrow \Re$ with

$$(f+g)(x) = f(x) + g(x) \qquad (\lambda f)(x) = \lambda f(x)$$

Bilinear forms

- to define dot-product we first need to recall the notion of a bilinear form
- **Definition:** a bilinear form on a vector space $\mathcal{H}$ is a mapping

$$Q: \mathcal{H} \times \mathcal{H} \rightarrow \Re$$
$$(x,x') \rightarrow Q(x,x')$$

such that $\forall x,x',x'' \in \mathcal{H}$

$$i) \quad Q[(\lambda x+\lambda x'),x''] = \lambda Q(x,x'') + \lambda' Q(x',x'')$$
$$ii) \quad Q[x'',(\lambda x+\lambda x')] = \lambda Q(x'',x) + \lambda' Q(x'',x')$$

- in $\Re^d$ the canonical bilinear form is

$$Q(x,x') = x^T A x'$$

- if $Q(x,x') = Q(x',x) \;\; \forall x,x' \in \mathcal{H}$, the form is symmetric

Dot products

- **Definition:** a dot-product on a vector space $\mathcal{H}$ is a symmetric bilinear form

$$<.,.>: \mathcal{H} \times \mathcal{H} \rightarrow \Re$$
$$(x,x') \rightarrow <x,x'>$$

such that

i) $<x,x> \geq 0, \; \forall x \in \mathcal{H}$
ii) $<x,x> = 0$ *if and only if* $x = 0$

- note that for the canonical bilinear form in $\Re^d$

$$<x,x> = x^T A x$$

this means that A must be positive definite

$$x^T A x > 0, \; \forall x \neq 0$$

Positive definite matrices

- recall that (e.g. Linear Algebra and Applications, Strang)
- **Definition:** each of the following is a necessary and sufficient condition for a real symmetric matrix A to be (semi) positive definite:
  i) $x^T A x \geq 0, \ \forall x \neq 0$
  ii) all eigenvalues of A satisfy $\lambda_i \geq 0$
  iii) all upper-left submatrices A_k have non-negative determinant
  iv) there is a matrix R with independent rows such that
  $A = R^T R$
- upper left submatrices:

$$A_1 = a_{1,1} \qquad A_2 = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \qquad A_3 = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix} \qquad \cdots$$

Positive definite matrices

- property iv) is particularly interesting
  - in $\Re^d$, $<x,x> = x^TAx$ is a dot-product kernel if and only if A is positive definite
  - from iv) this holds if and only if there is R such that $A = R^TR$
  - hence

    $$<x,y> = x^TAy = (Rx)^T(Ry) = \Phi(x)^T\Phi(y)$$

    with

    $$\begin{aligned} \Phi: \Re^d &\rightarrow \Re^d \\ x &\rightarrow Rx \end{aligned}$$

- i.e. the dot-product kernel

$$k(x,z) = x^TAz, \quad (A \text{ positive definite})$$

- is the standard dot-product in the range space of the mapping $\Phi(x) = Rx$

Note

- there are positive semidefinite matrices

$$x^T A x \geq 0$$

  and positive definite matrices

$$x^T A x > 0$$

- we will work with semidefinite but, to simplify, will call definite
- if we really need > 0 we will say "strictly positive definite"

Positive definite kernels

- how do we define a positive definite function?
- Definition: a function $k(x,y)$ is a positive definite kernel on $\mathcal{X} \times \mathcal{X}$ if $\forall l$ and $\forall \{x_1, ..., x_l\}$, $x_i \in \mathcal{X}$, the Gram matrix

$$K = \begin{bmatrix} & \vdots & \\ \cdots & k(x_i, x_j) & \cdots \\ & \vdots & \end{bmatrix}$$

is positive definite.

- Note: this implies that
  - $k(x,x) \geq 0 \quad \forall x \in \mathcal{X}$
  - $\begin{bmatrix} k(x,x) & k(x,y) \\ k(y,x) & k(y,y) \end{bmatrix}$ PD $\forall x,y \in \mathcal{X}$ (*)

etc...

Positive definite kernels

- this proves some simple properties
  - a PD kernel is symmetric

$$\boxed{k(x,y)=k(y,x), \quad \forall x,y \in X}$$

Proof:
since PD means symmetric (*) implies $k(x,y) = k(y,x)$ $\forall x,y \in \mathcal{X}$ ∎

  - Cauchy-Schwarz inequality for kernels: if $k(x,y)$ is a PD kernel, then

$$\boxed{k(x,y)^2 \leq k(x,x)k(y,y), \quad \forall x,y \in X}$$

Proof:
from (*), and property iii) of PD matrices, the determinant of the 2x2 matrix of (*) is non-negative. This means that

$$k(x,x)k(y,y) - k(x,y)^2 \geq 0 \quad \blacksquare$$

Positive definite kernels

- it is not hard to show that all dot product kernels are PD
- Lemma 1: Let $k(x,y)$ be a dot-product kernel. Then $k(x,y)$ is positive definite
- proof:
  - $k(x,y)$ dot product kernel *implies that*
  - $\exists \Phi$ and some dot product $<.,.>$ such that

    $$k(x,y) = <\Phi(x), \Phi(y)>$$
  - this implies that if:
    - we pick any l, and any sequence $\{x_1, \ldots, x_l\}$,
    - and let K be the associated Gram matrix
    - then, for $\forall c \neq 0$

$$K = \begin{bmatrix} & \vdots & \\ \cdots & k(x_i, x_j) & \cdots \\ & \vdots & \end{bmatrix}$$

Positive definite kernels

$$c^T K c = \sum_{ij} c_i c_j k(x_i, x_j)$$

$$= \sum_{ij} c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle$$ (k is dot product)

$$= \left\langle \sum_i c_i \Phi(x_i), \sum_j c_j \Phi(x_j) \right\rangle$$ (<.,.> is a bilinear form)

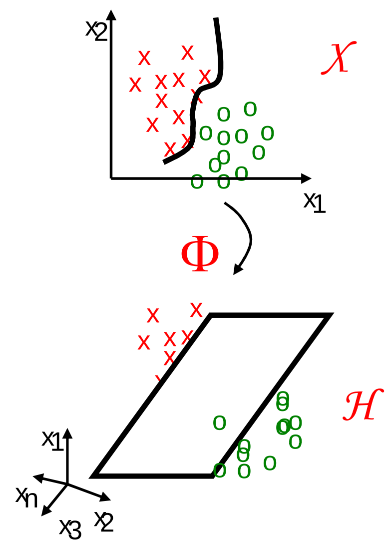
$$= \left\| \sum_i c_i \Phi(x_i) \right\|^2 \geq 0$$ (from def of dot product)

$$\blacksquare$$

Positive definite kernels

- the converse is also true but more difficult to prove
- Lemma 2: Let $k(x,y)$, $x,y \in \mathcal{X}$, be a positive definite kernel. Then $k(x,y)$ is a dot product kernel
- proof:
  - we need to show that there is a transformation Φ, a vector space $\mathcal{H} = \Phi(\mathcal{X})$, and a dot product $<.,.>_*$ in $\mathcal{H}$ such that

$$k(x,y) = <\Phi(x), \Phi(y)>_*$$



  - we proceed in three steps
    1. construct a vector space $\mathcal{H}$
    2. define the dot-product $<.,.>_*$ on $\mathcal{H}$
    3. show that $k(x,y) = <\Phi(x), \Phi(y)>_*$ holds

The vector space $\mathcal{H}$

- we define $\mathcal{H}$ as the space spanned by linear combinations of $k(.,x_i)$

$$\boxed{\mathcal{H} = \left\{ f(.) \,|\, f(.) = \sum_{i=1}^{m} \alpha_i k(.,x_i), \quad \forall m, \forall x_i \in X \right\}}$$

- notation: by $k(.,x_i)$ we mean a function of $g(y) = k(y,x_i)$ of y, x_i is fixed.
- homework: check that $\mathcal{H}$ is a vector space
  - e.g. 2)

$$\left.\begin{array}{l} f(.) = \sum_{i=1}^{m} \alpha_i k(.,x_i) \\ f'(.) = \sum_{i=1}^{m'} \beta_j k(.,x'_j) \end{array}\right\} f(.) + f'(.) = f'(.) + f(.) \in \mathcal{H}$$

Example

- when we use the Gaussian kernel

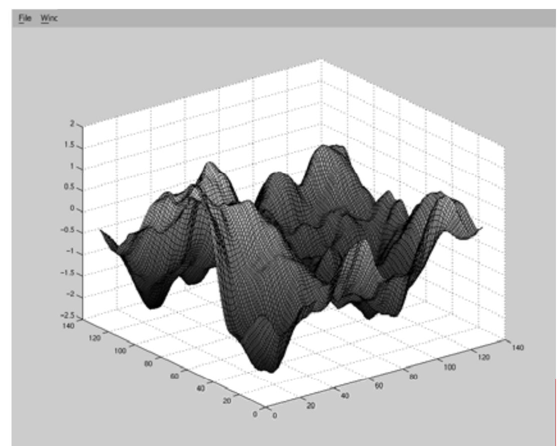
$$K(.,x_i) = e^{-\frac{\|.-x_i\|^2}{\sigma^2}}$$

- $k(.,x_i)$ is a Gaussian centered on x_i with covariance σI
- and

$$H = \left\{ f(.) \,|\, f(.) = \sum_{i=1}^{m} \alpha_i e^{-\frac{\|.-x_i\|}{\sigma^2}}, \ \forall m, \forall x_i \right\}$$

  is the space of all linear combinations of Gaussians

  e.g.



- note that these are not mixtures but close

The operator $<.,.>_*$

- if *f(.)* and *g(.)* $\in \mathcal{H}$, with

$$\boxed{f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \qquad g(.) = \sum_{j=1}^{m'} \beta_j k(., x'_j)} \qquad (**)$$

- we define the operator $<.,.>_*$ as

$$\boxed{\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)} \qquad (***)$$

Example

- when we use the Gaussian kernel

$$K(.,x_i) = e^{-\frac{\left\| . - x_i \right\|^2}{\sigma^2}}$$

- the operator <.,.>$_*$ is a weighted sum of Gaussian terms

$$\boxed{\left\langle f, g \right\rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j e^{-\frac{\left\| x_i - x'_j \right\|^2}{\sigma^2}}}$$

- you can look at this as either:
  - a dot product in $\mathcal{H}$ (still need to prove this)
  - a non-linear measure of similarity in $\mathcal{X}$, somewhat related to likelihoods

The operator <.,.>$_*$

- important note: for *f(.)* and *g(.)* $\in \mathcal{H}$, the operator <.,.>$_*$

$$\boxed{\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k\left(x_i, x'_j\right)}$$

has the property

$$\boxed{\langle k(., x_i), k(., x'_j) \rangle_* = k\left(x_i, x'_j\right)} \qquad (****)$$

- proof: just make

$$\begin{cases} \alpha_i = 1, & \alpha_k = 0 \ \ \forall k \neq i \\ \beta_j = 1, & \beta_k = 0 \ \ \forall k \neq j \end{cases} \qquad \blacksquare$$

The operator <.,.>$_*$

- assume that <.,.>$_*$ is a dot product in $\mathcal{H}$ (proof in moments)
- since

$$\langle k(.,x_i), k(.,x_j) \rangle_* = k(x_i, x_j)$$

- then, clearly

$$\boxed{k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_*}$$

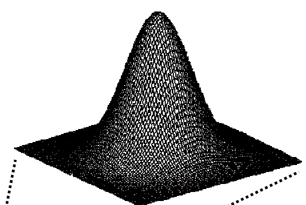
with

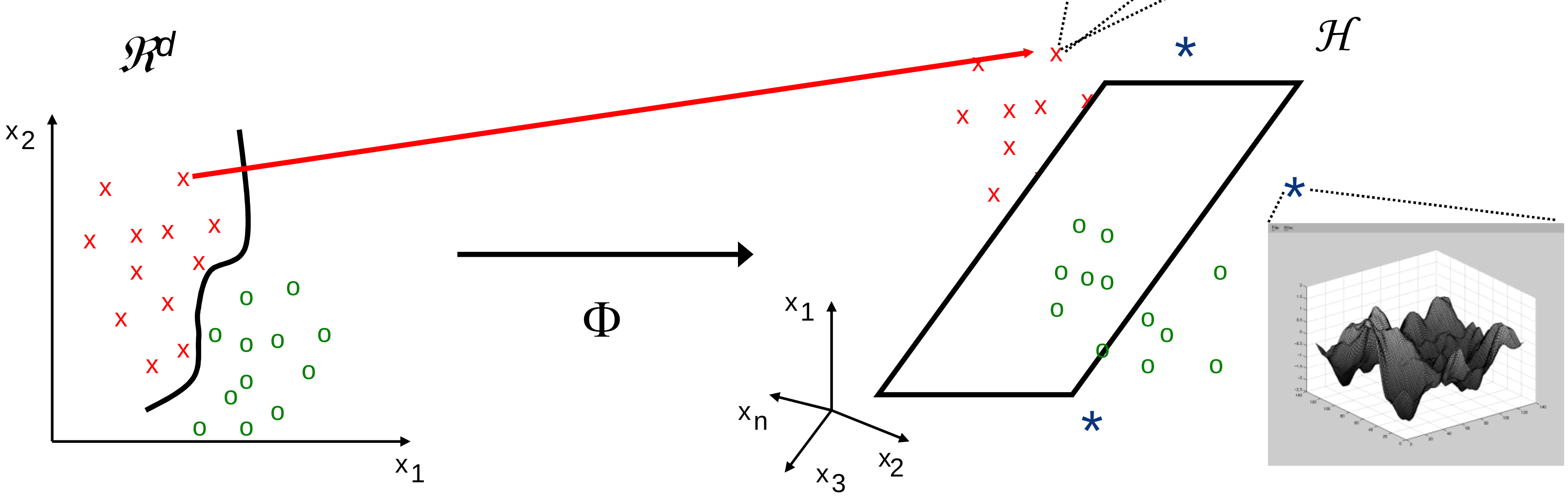
$$\boxed{\begin{aligned} \Phi: \mathcal{X} &\rightarrow \mathcal{H} \\ x &\rightarrow k(.,x) \end{aligned}}$$

- i.e. the kernel is a dot-product on $\mathcal{H}$, which results from the feature transformation Φ
- this proves Lemma 2 ■

Example

- when we use the Gaussian kernel $K(x, x_i) = e^{-\frac{\|x - x_i\|^2}{\sigma}}$
  - the point $x_i \in \Re^d$ is mapped into the Gaussian $G(x, x_i, \sigma I)$
  - $\mathcal{H}$ is the space of all functions that are linear combinations of Gaussians
  - this has infinite dimension
  - the kernel is a dot product in $\mathcal{H}$, and a non-linear similarity on $\mathcal{X}$





In summary

- to show that $k(x,y)$, $x,y \in \mathcal{X}$, positive definite $\Rightarrow$ $k(x,y)$ is a dot product kernel
- we need to show that

$$\boxed{\langle f,g\rangle_* = \sum_{i=1}^{m}\sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)}$$

- is a dot product on

$$\boxed{\mathcal{H} = \left\{ f(.) \,|\, f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i), \quad \forall m, \forall x_i \in X \right\}}$$

- this reduces to verifying the dot product conditions

The operator <.,.>$_*$

- 1) is <.,.>$_*$ a bilinear form on $\mathcal{H}$?
- by definition of *f(.)* and *g(.)* in (**)

$$\langle f, g \rangle_* = \left\langle \sum_{i=1}^{m} \alpha_i k(., x_i), \sum_{j=1}^{m'} \beta_j k(., x'_j) \right\rangle_*$$

- on the other hand,

$$\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \qquad \text{from (***)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j \langle k(., x_i), k(., x'_j) \rangle_* \qquad \text{from (****)}$$

- equality of the two left hand sides is the definition of bilinearity ■

The operator $<.,.>_*$

- 2) is $<.,.>_*$ symmetric?
- note that

$$\langle g, f \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k\left(x'_j, x_i\right) = \langle f, g \rangle_*$$

- if and only if $k(x_i, x_j') = k(x_j', x_i)$ for all x_i, x_j'.
- but this follows from the positive definiteness of $k(x,y)$
- we have seen that a PD kernel is always symmetric
- hence, $<.,.>_*$ is symmetric ■

The operator <.,.>$_*$

- 3) is $\langle f,f \rangle_* \geq 0, \quad \forall f \in \mathcal{H}$?
- by definition of $f(.)$ in (**)

$$\langle f, f \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha$$

  where $\alpha \in \Re^m$ and K is the Gram matrix

- since $k(x,y)$ is positive definite, K is positive definite by definition and $\langle f,f \rangle_* \geq 0$ ■ (x)
- the only non-trivial part of the proof is to show that $\langle f,f \rangle_* = 0 \Rightarrow f = 0$
- we need two more results

The operator <.,.>$_*$

- Lemma 3: <.,.>$_*$ is itself a positive definite kernel on $\mathcal{H}$ x $\mathcal{H}$
- proof:
  - consider any sequence $\{f_1, \ldots, f_m\}$, $f_i \in \mathcal{H}$
  - then

$$\sum_{ij} \gamma_i \gamma_j \langle f_i, f_j \rangle_* = \left\langle \sum_i \gamma_i f_i, \sum_j \gamma_j f_j \right\rangle_* \quad \text{(by bilinearity of <.,.>*)}$$

$$= \langle g_1, g_2 \rangle_* \quad \text{(for some } g_1, g_2 \in \mathcal{H}\text{)}$$

$$\geq 0 \quad \text{(by (x))}$$

  - hence the Gram matrix is always PD and the kernel <.,.>$_*$ is PD ■

The operator <.,.>$_*$

- Lemma 4: $\forall f \in \mathcal{H}$, $<k(.,x),f(.)>_* = f(x)$
- proof:

$$\langle k(.,x), f(.)\rangle_* = \left\langle k(.,x), \sum_i \alpha_i k(.,x_i) \right\rangle_* \quad \text{(by (**))}$$

$$= \sum_i \alpha_i \langle k(.,x), k(.,x_i)\rangle_* \quad \text{(by bilinearity of <.,.>*)}$$

$$= \sum_i \alpha_i k(x,x_i) = f(x) \blacksquare \quad \text{(by (****))}$$

The operator <.,.>$_*$

- 4) we are now ready to prove that $\langle f,f\rangle_* = 0 \Rightarrow f = 0$
- proof:
  - since <.,.>* is a PD kernel (lemma 3) we can apply Cauchy-Schwarz $k(x,y)^2 \leq k(x,x)k(y,y), \quad \forall x,y \in X$
  - using k(.,x) as x and f(.) as y this becomes

$$\langle k(.,x),k(.,x)\rangle_* \langle f,f\rangle_* \geq \left(\langle k(.,x),f\rangle_*\right)^2$$

  - and using lemma 4

$$k(x,x)\langle f,f\rangle_* \geq f^2(x)$$

  - from which $\langle f,f\rangle_* = 0 \Rightarrow f = 0$ ■

In summary

- we have shown that

$$\boxed{\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)}$$

- is a dot product on

$$\boxed{\mathcal{H} = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i), \quad \forall m, \forall x_i \in X \right\}}$$

- and this shows that if $k(x,y)$, $x,y \in \mathcal{X}$, is a positive definite kernel, then $k(x,y)$ is a dot product kernel.
- since we had initially proven the converse, we have the following theorem.

Dot product kernels

- $\boxed{\textbf{Theorem: } k(x,y),\ x,y \in \mathcal{X}, \text{ is a dot-product kernel if and only if it is a positive definite kernel}}$

- this is interesting because it allows us to check whether a kernel is a dot product or not!
  - check if the Gram matrix is positive definite for all possible sequences $\{x_1, ..., x_l\}, x_i \in \mathcal{X}$
- but the proof is much more interesting than this result alone
- it actually gives us insight on what the kernel is doing
- let's summarize

Dot product kernels

- a dot product kernel $k(x,y),\ x,y \in \mathcal{X}$:
  - applies a feature transformation

$$\boxed{\begin{array}{l} \Phi: \mathcal{X} \rightarrow \mathcal{H} \\ \quad x \quad \rightarrow k(.,x) \end{array}}$$

  - to the vector space

$$\boxed{\mathcal{H} = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i), \quad \forall m, \forall x_i \in X \right\}}$$

  - where the kernel implements the dot product

$$\boxed{\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j)}$$

Dot product kernels

- the dot product

$$\boxed{\langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k\left(x_i, x'_j\right)}$$

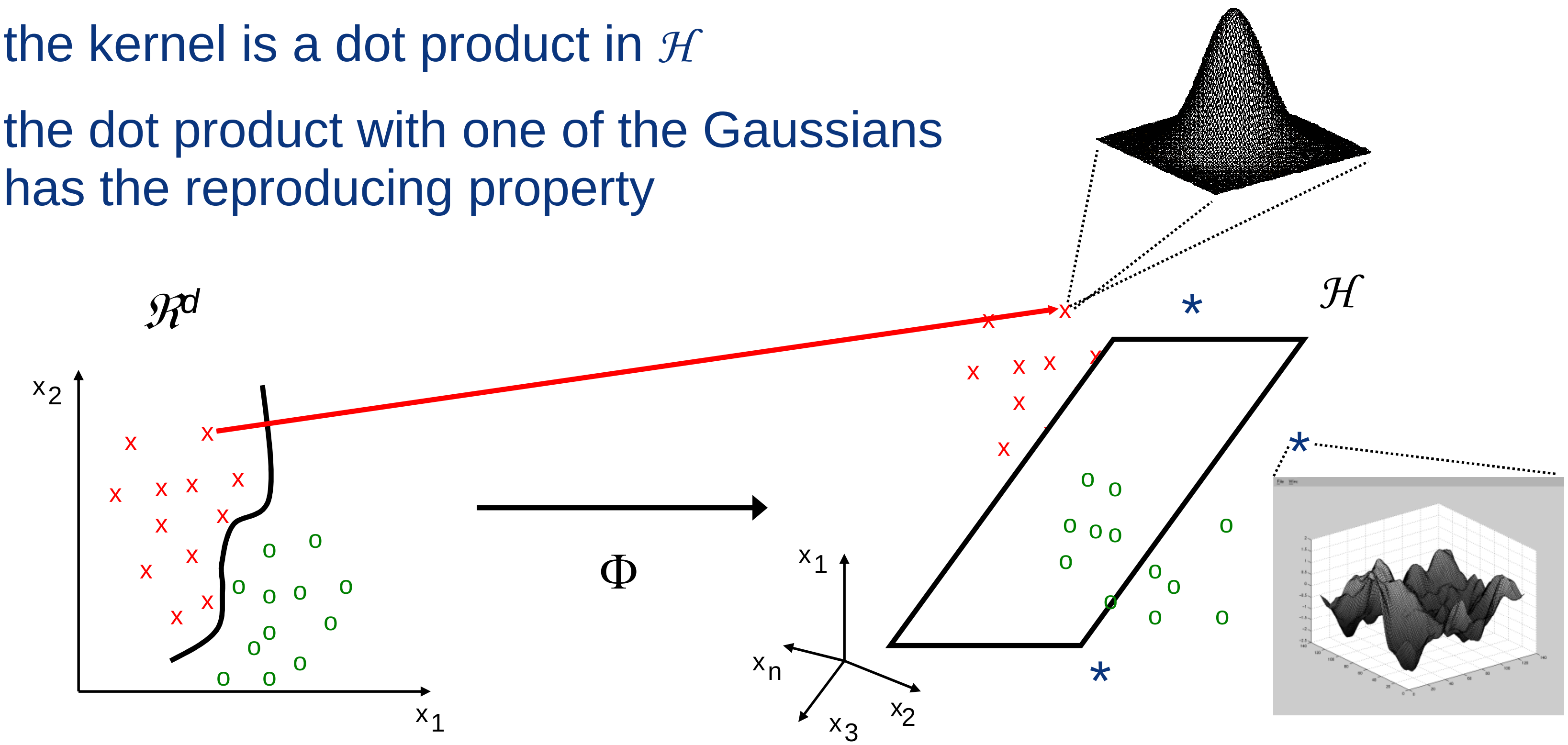
- has the reproducing property

$$\boxed{\langle k(.,x), f(.) \rangle_* = f(x)}$$

- you can think of this as analog to the convolution with a Dirac delta
- we will talk about this a lot in the coming lectures
- finally, <.,.>$_*$ is itself a positive definite kernel on $\mathcal{H}$ x $\mathcal{H}$

A good picture to remember

- when we use the Gaussian kernel $K(x,x_i) = e^{-\frac{\|x-x_i\|^2}{\sigma}}$
  - the point $x_i \in \mathcal{R}^d$ is mapped into the Gaussian $G(x,x_i,\sigma I)$
  - $\mathcal{H}$ is the space of all functions that are linear combinations of Gaussians
  - the kernel is a dot product in $\mathcal{H}$
  - the dot product with one of the Gaussians has the reproducing property



Any Questions?