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**

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

- 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 $\text{dim}(\mathcal{Z}) \gg \text{dim}(\mathcal{X})$

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”

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

   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 $\Phi : \mathbb{R}^d \to \mathbb{R}^{d^2}$

\[
\begin{pmatrix}
x_1 \\
\vdots \\
x_d
\end{pmatrix} \rightarrow (x_1 x_1, x_1 x_2, \ldots, x_1 x_d, \ldots, x_d x_1, x_d x_2, \ldots, x_d x_d)^T
\]

- while $K(x,z)$ has complexity $O(d)$, $\Phi(x)^T \Phi(z)$ is $O(d^2)$

- for $K(x,z) = (x^T z)^k$ 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 $\gamma$ 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
    \[
    K(x, z) = \left(1 + x^T z\right)^k, \quad k \in \{1, 2, \ldots\}
    \]
Question

“this problem of mine is really asking for the kernel $k'(x,z) = ...$”

- how do I know if this is a dot-product kernel?

let’s start by the definition

**Definition:** a mapping

$$k: \mathcal{X} \times \mathcal{X} \to \mathcal{H}$$

$$(x,y) \to k(x,y)$$

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 $\langle..,..\rangle$ can be abstract, not necessarily $\mathbb{R}^d$

**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) $x+x' = x'+x$  
  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 $\mathbb{R}^d$ with standard vector addition and scalar multiplication

- another example is the space of mappings $\mathcal{X} \rightarrow \mathbb{R}$ with

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

\[
Q: \mathcal{H} \times \mathcal{H} \to \mathbb{R} \\
(x, x') \to Q(x, x')
\]

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

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

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

- in \( \mathbb{R}^d \) 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

\[
<.,.>: \mathcal{H} \times \mathcal{H} \to \mathbb{R}
\]

\[
(x,x') \to <x,x'>
\]

such that

1. \( <x,x> \geq 0, \ \forall x \in \mathcal{H} \)
2. \( <x,x> = 0 \) if and only if \( x = 0 \)

Note that for the canonical bilinear form in \( \mathbb{R}^d \)

\[
<x,x> = x^T A x
\]

This means that \( A \) must be positive definite

\[
x^T A x > 0, \ \forall x \neq 0
\]
Positive definite matrices

 recalled 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:

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

$\ldots$
Positive definite matrices

property iv) is particularly interesting

- in \( \mathbb{R}^d \), \( \langle x, x \rangle = x^T A x \) 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 \( A = R^T R \)
- hence

\[
\langle x, y \rangle = x^T A y = (R x)^T (R y) = \Phi(x)^T \Phi(y)
\]

with

\[
\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^d
\]

\[
x \rightarrow R x
\]

i.e. the dot-product kernel

\[
k(x, z) = x^T A z, \quad (A \text{ positive definite})
\]

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

- There are positive semidefinite matrices
  \[ x^T Ax \geq 0 \]
  and positive definite matrices
  \[ x^T Ax > 0 \]

- 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 $\forall \{x_1, \ldots, x_l\}, x_i \in \mathcal{X}$, the Gram matrix

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

is positive definite.

Note: this implies that

- $k(x, x) \geq 0 \; \forall x \in \mathcal{X}$
- \[ \begin{bmatrix} k(x, x) & k(x, y) \\ k(y, x) & k(y, y) \end{bmatrix} \text{PD} \; \forall x, y \in \mathcal{X} \] (*) etc...
Positive definite kernels

This proves some simple properties:

- A PD kernel is symmetric

\[ k(x, y) = k(y, x), \quad \forall x, y \in X \]

Proof:
Since PD means symmetric (*), implies \( k(x, y) = k(y, x) \) \( \forall x, y \in X \)

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

\[ k(x, y)^2 \leq k(x, x)k(y, y), \quad \forall x, y \in X \]

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

\[ k(x, x)k(y, y) - k(x, y)^2 \geq 0 \]
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 \( \langle ., . \rangle \) such that
    \[
    k(x, y) = \langle \Phi(x), \Phi(y) \rangle
    \]
  - this implies that if:
    - we pick any \( l \), and any sequence \( \{x_1, \ldots, x_l\} \),
    - and let \( K \) be the associated Gram matrix
    \[
    K = \begin{bmatrix}
    \vdots \\
    \cdots k(x_i, x_j) \cdots \\
    \vdots 
    \end{bmatrix}
    \]
    - then, for \( \forall c \neq 0 \)
Positive definite kernels

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

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

\[ = \left\langle \sum_i c_i \Phi(x_i), \sum_j c_j \Phi(x_j) \right\rangle \quad \text{(}\langle ., . \rangle\text{ is a bilinear form)} \]

\[ = \left\| \sum_i c_i \Phi(x_i) \right\|^2 \geq 0 \quad \text{(from def of dot product)} \]
Positive definite kernels

- the converse is also true but more difficult to prove

Lemma 2: Let \( k(x,y), \ x, y \in X \), be a positive definite kernel. Then \( k(x,y) \) is a dot product kernel

proof:

- we need to show that there is a transformation \( \Phi \), a vector space \( \mathcal{H} = \Phi(X) \), and a dot product \( \langle ., . \rangle^* \) in \( \mathcal{H} \) such that

\[
k(x,y) = \langle \Phi(x), \Phi(y) \rangle^*
\]

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

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

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

- 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) $f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i)$
  $$f(.) + f'(.) = f'(.) + f(.) \in \mathcal{H}$$

  $f'(.) = \sum_{i=1}^{m} \beta_j k(., x'_j)$
Example

- when we use the Gaussian kernel
  \[ K(., x_i) = e^{-\frac{\|x - x_i\|^2}{\sigma^2}} \]

- \( k(., x_i) \) is a Gaussian centered on \( x_i \) with covariance \( \sigma I \)

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

  is the space of all linear combinations of Gaussians

- note that these are not mixtures but close
The operator $\langle ., . \rangle_*$

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

\[
f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i) \quad g(.) = \sum_{j=1}^{m'} \beta_j k(., x'_j)
\]  

(**)

- we define the operator $\langle ., . \rangle_*$ as

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

(***)

\[20\]
Example

when we use the Gaussian kernel

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

the operator \( <.,.>_* \) is a weighted sum of Gaussian terms

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

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 $<.,.>^*$

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

has the property

$$
\langle k(., x_i), k(., x'_j) \rangle^* = k(x_i, x'_j)
$$

proof: just make

$$
\begin{align*}
\alpha_i &= 1, \quad \alpha_k = 0 \quad \forall k \neq i \\
\beta_j &= 1, \quad \beta_k = 0 \quad \forall k \neq j
\end{align*}
$$
The operator $<.,.>^*$

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

- since

$$\left\langle k(.,x_i), k(.,x_j) \right\rangle^* = k(x_i, x_j)$$

- then, clearly

$$k(x_i, x_j) = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle^*$$

with

$$\Phi : X \rightarrow \mathcal{H}$$

$$x \rightarrow k(.,x)$$

- i.e. the kernel is a dot-product on $\mathcal{H}$, which results from the feature transformation $\Phi$

- this proves **Lemma 2**
Example

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

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

- is a dot product on

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

- 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,

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

from (***)

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

from (****)

equality of the two left hand sides is the definition of bilinearity
The operator $\langle \cdot , \cdot \rangle^*$

2) is $\langle \cdot , \cdot \rangle^*$ symmetric?

Note that

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

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, $\langle \cdot , \cdot \rangle^*$ is symmetric.
The operator <..,*>

1. 3) is \(<f,f>* ≥ 0, \forall f \in H?>

2. 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 \(\alpha \in \mathbb{R}^m\) and \(K\) is the Gram matrix

3. since \(k(x,y)\) is positive definite, \(K\) is positive definite by definition and \(<f,f>* ≥ 0\)  

4. the only non-trivial part of the proof is to show that \(<f,f>* = 0 \Rightarrow f = 0\)

5. we need two more results
The operator $<.,.>*$

Lemma 3: $<.,.>*$ is itself a positive definite kernel on $\mathcal{H} \times \mathcal{H}$

proof:

- consider any sequence $\{f_1, \ldots, f_m\}, f_i \in \mathcal{H}$
- then

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

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

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

- hence the Gram matrix is always PD and the kernel $<.,.>*$ is PD
The operator $<.,.>^*$

Lemma 4: $\forall f \in \mathcal{H}, \langle k(.,x),f(.)\rangle^* = f(x)$

proof:

\[
\langle k(.,x),f(.)\rangle^* = \left\langle k(.,x), \sum_i \alpha_i k(.,x_i) \right\rangle^* \quad \text{(by (**))}
\]

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

\[= \sum_i \alpha_i k(x,x_i) = f(x) \quad \text{(by (****))}
\]
The operator \(<.,.>^*\)

4) we are now ready to prove that \(<f,f>^* = 0 \Rightarrow f = 0\)

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

\[ \left\langle k(., x), k(., x) \right\rangle^* \left\langle f, f \right\rangle^* \geq \left( \left\langle k(., x), f \right\rangle^* \right)^2 \]

- and using lemma 4

\[ k(x, x) \left\langle f, f \right\rangle^* \geq f^2(x) \]

- from which \(<f,f>^* = 0 \Rightarrow f = 0\)
In summary

we have shown that

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

is a dot product on

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

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

**Theorem:** \( k(x,y), x,y \in \mathcal{X}, \) 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 \( \{x_1, \ldots, x_l\}, x_i \in \mathcal{X} \)

- 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
    \[ \Phi: \mathcal{X} \rightarrow \mathcal{H} \]
    \[ x \rightarrow k(.,x) \]
  - To the vector space
    \[ \mathcal{H} = \left\{ f(.) \mid f(.) = \sum_{i=1}^{m} \alpha_i k(., x_i), \quad \forall m, \forall x_i \in \mathcal{X} \right\} \]
  - Where the kernel implements the dot product
    \[ \langle f, g \rangle_* = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \]
Dot product kernels

- the dot product

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

- has the reproducing property

\[
\langle k(., x), f(.) \rangle_* = f(x)
\]

- 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, \langle ., . \rangle_* is itself a positive definite kernel on \mathcal{H} \times \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 \( x_i \in \mathbb{R}^d \) is mapped into the Gaussian \( G(x, x_i, \sigma) \)
  - \( \mathcal{H} \) 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?