Lecture 14: Kernels

Prof. Dr. Mario Fritz

2022 06 08

https://fritz.cispa.saarland
Trustworthy AI
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

Outline

Bibliography

Motivation

Kernels

Properties

Vector space structure

Appendix

Main references

- Learning with Kernels - Chapter 2
- Bishop - Chapter 6

Outline

Bibliography

Motivation

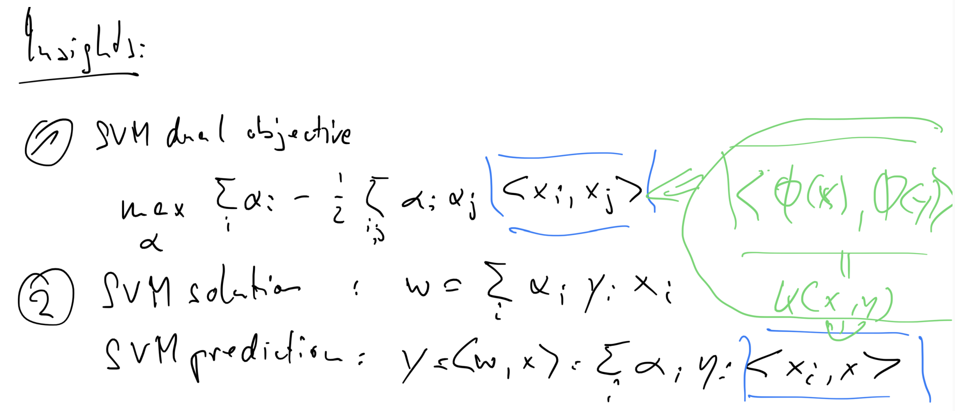
Kernels

Properties

Vector space structure

Appendix

Motivation



Insights:

① SVM dual objective

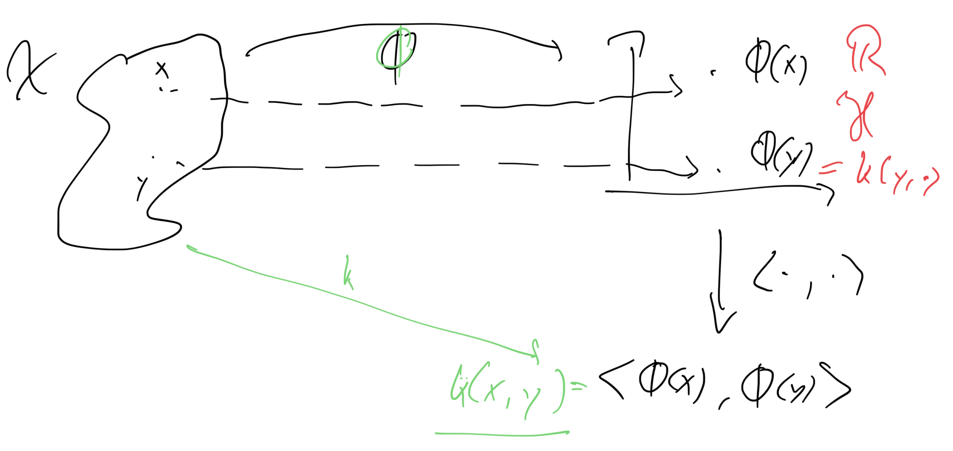
$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \boxed{\langle x_i, x_j \rangle} \leftarrow \boxed{\begin{matrix} \langle \phi(x), \phi(y) \rangle \\ \| \\ k(x,y) \end{matrix}}$$

② SVM solution : $w = \sum_i \alpha_i y_i x_i$

SVM prediction : $y = \langle w, x \rangle = \sum_i \alpha_i y_i \boxed{\langle x_i, x \rangle}$

... refer to video lecture ...

Motivation


$\mathcal{X}$
x
y
Φ
$\cdot\ \Phi(x)$
$\cdot\ \Phi(y) = k(y,\cdot)$
$\mathbb{R}$
$\mathcal{H}$
$\langle \cdot , \cdot \rangle$
k
$k(x,y) = \langle \Phi(x), \Phi(y) \rangle$

... refer to video lecture ...

Recap (I): Classification Problems

- An input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\mathcal{Y} = \{-1, +1\}$.
- A training data $\mathcal{D} = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \ldots, (\mathrm{x}_n, y_n)\}$
- So far, we considered

$$\mathcal{F} = \left\{ f(\mathrm{x}) = \mathrm{w}^\top \mathrm{x} \mid \mathrm{w} \in \mathbb{R}^d \right\}.$$

- Given any $f \in \mathcal{F}$, our decision function is

$$g(\mathrm{x}) = \begin{cases} -1 & \text{if } f(\mathrm{x}) < 0 \\ +1 & \text{if } f(\mathrm{x}) \geq 0 \end{cases}.$$

- We also write $g(\mathrm{x}) = \mathrm{sgn}(f(\mathrm{x})) = \mathrm{sgn}(\mathrm{w}^\top \mathrm{x})$.
- **Goal:** Learn a predictor $f^* \in \mathcal{F}$ from the training data such that
    1. $\mathrm{sgn}(f^*(\mathrm{x})) = y_i$ for $i = 1, \ldots, n$ and
    2. the predictor f **generalizes** well to previously unseen data.

Perceptron Algorithm

Perceptron learning rule:

- Initialize $\mathrm{w}_0 = 0$. For $t = 1, \ldots, T$:
  1. $\hat{y}_t = \mathrm{sgn}(\mathrm{w}_t^\top \mathrm{x}_t)$.
  2. If $\hat{y}_t = y_t$, do nothing.
  3. If $\hat{y}_t \neq y_t$, update $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + y_t \mathrm{x}_t$.

Perceptron Algorithm

Perceptron learning rule:

- Initialize $\mathrm{w}_0 = 0$. For $t = 1, \ldots, T$:
  1. $\hat{y}_t = \mathrm{sgn}(\mathrm{w}_t^\top \mathrm{x}_t)$.
  2. If $\hat{y}_t = y_t$, do nothing.
  3. If $\hat{y}_t \neq y_t$, update $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + y_t \mathrm{x}_t$.
- Any solution can be expressed as

$$\mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i.$$

Perceptron Algorithm

Perceptron learning rule:

- Initialize $\mathrm{w}_0 = 0$. For $t = 1, \ldots, T$:
  1. $\hat{y}_t = \mathrm{sgn}(\mathrm{w}_t^\top \mathrm{x}_t)$.
  2. If $\hat{y}_t = y_t$, do nothing.
  3. If $\hat{y}_t \neq y_t$, update $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + y_t \mathrm{x}_t$.
- Any solution can be expressed as

$$\mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i.$$

- Perceptron **prediction** rule:

$$\hat{y} = \mathrm{sgn}(\mathrm{w}^\top \mathrm{x}) = \mathrm{sgn}\left(\left(\sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i\right)^\top \mathrm{x}\right) = \mathrm{sgn} \sum_{i=1}^{n} \alpha_i y_i \langle \mathrm{x}_i, \mathrm{x} \rangle.$$

Perceptron Algorithm

Perceptron learning rule:

- Initialize $\mathrm{w}_0 = 0$. For $t = 1, \ldots, T$:
  1. $\hat{y}_t = \mathrm{sgn}(\mathrm{w}_t^\top \mathrm{x}_t)$.
  2. If $\hat{y}_t = y_t$, do nothing.
  3. If $\hat{y}_t \neq y_t$, update $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + y_t \mathrm{x}_t$.
- Any solution can be expressed as

$$\mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i .$$

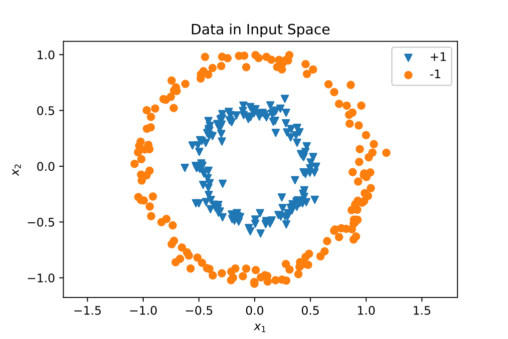
- Perceptron **prediction** rule:

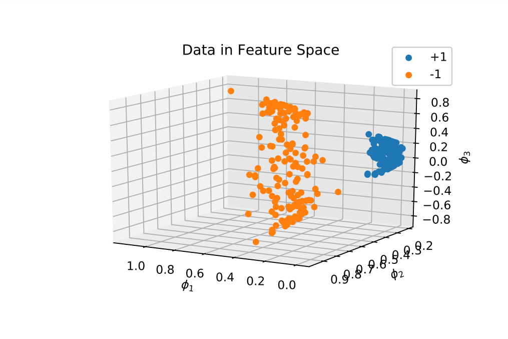
$$\hat{y} = \mathrm{sgn}(\mathrm{w}^\top \mathrm{x}) = \mathrm{sgn}\left(\left(\sum_{i=1}^{n} \alpha_i y_i \mathrm{x}_i\right)^\top \mathrm{x}\right) = \mathrm{sgn} \sum_{i=1}^{n} \alpha_i y_i \langle \mathrm{x}_i, \mathrm{x} \rangle .$$

- **Step 3:** (Dual form) If $\hat{y}_t \neq y_i$, update $\alpha_i \leftarrow \alpha_i + 1$.

Non-linear Classification Problem

Question: How would you solve this classification problem?





Solution: Feature Map (Embedding).

$$\phi \ : \ \Re^2 \longrightarrow \Re^3$$

$$(x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

Solving Non-linear Classification Problems

1. Construct a non-linear feature map $\phi : \Re^d \rightarrow \Re^m$.
2. Evaluate $D_\phi = \{\phi(\mathrm{x}_1), \phi(\mathrm{x}_2), \ldots, \phi(\mathrm{x}_n)\}$.
3. Learn classifier using D_ϕ.

Example: In the dual perceptron:

$$\mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathrm{x}_i).$$

$$\hat{y} = \operatorname{sgn} \sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle.$$

Observation: In general, $m >> d$ or even $m = \infty$!

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \mathrm{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle\right).$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \operatorname{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle\right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned} \langle \phi(\mathrm{x}), \phi(\mathrm{z}) \rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 \end{aligned}$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \mathrm{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle \right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned} \langle \phi(\mathrm{x}), \phi(\mathrm{z}) \rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 \end{aligned}$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \mathrm{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle\right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned} \langle \phi(\mathrm{x}), \phi(\mathrm{z}) \rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 \end{aligned}$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \operatorname{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle \right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned} \langle \phi(\mathrm{x}), \phi(\mathrm{z}) \rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 \\ &= (\mathrm{x} \cdot \mathrm{z})^2. \end{aligned}$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \operatorname{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle\right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned} \langle \phi(\mathrm{x}), \phi(\mathrm{z}) \rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 \\ &= (\mathrm{x} \cdot \mathrm{z})^2. \end{aligned}$$

- For $\mathrm{x}, \mathrm{x}' \in \mathcal{X}$, define a **kernel** function:

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}_i), \phi(\mathrm{x}) \rangle = (\mathrm{x} \cdot \mathrm{x}')^2$$

Dual Perceptron Revisited

- Recall our feature map $\phi : (x_1, x_2) \longmapsto (x_1^2, x_2^2, \sqrt{2}x_1 x_2)$
- The prediction rule of **dual perceptron algorithm** becomes

$$\hat{y} = \operatorname{sgn}(\mathrm{w}^\top \phi(\mathrm{x})) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \phi(\mathrm{x}_i), \phi(\mathrm{x})\rangle\right).$$

- An inner product between $\phi(\mathrm{x})$ and $\phi(\mathrm{z})$ in $\mathbb{R}^3$

$$\begin{aligned}\langle \phi(\mathrm{x}), \phi(\mathrm{z})\rangle_{\mathbb{R}^3} &= (x_1^2, x_2^2, \sqrt{2}x_1x_2)^\top (z_1^2, z_2^2, \sqrt{2}z_1z_2) \\ &= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 \\ &= (\mathrm{x} \cdot \mathrm{z})^2.\end{aligned}$$

- For $\mathrm{x}, \mathrm{x}' \in \mathcal{X}$, define a **kernel** function:

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}_i), \phi(\mathrm{x})\rangle = (\mathrm{x} \cdot \mathrm{x}')^2$$

- Hence, $\hat{y} = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(\mathrm{x}_i, \mathrm{x})\right)$. **We do not need the embedding ϕ but only a kernel!**

Outline

Bibliography

Motivation

Kernels

Properties

Vector space structure

Appendix

Kernel Trick

Theorem (Kernel implies embedding)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel *on $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a (feature) map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathrm{x}, \mathrm{x}' \in \mathcal{X}$ we have*

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle_{\mathcal{H}}. \tag{1}$$

Kernel Trick

Theorem (Kernel implies embedding)

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel *on $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a (feature) map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathrm{x}, \mathrm{x}' \in \mathcal{X}$ we have*

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle_{\mathcal{H}}. \tag{1}$$

Main properties of a **Hilbert space** $\mathcal{H}$ to keep in mind:

- $\mathcal{H}$ is a vector space with scarlar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.
- Space is complete (all Cauchy sequences converge).
- Scalar product leads to a norm: $\|\mathrm{x}\|_{\mathcal{H}} = \langle \mathrm{x}, \mathrm{x} \rangle_{\mathcal{H}}$

Observation: If you are not familiar with Hilbert spaces, think of $\mathbb{R}^d$.

Reproducing kernel Hilbert spaces

Definition

A **reproducing kernel Hilbert space (RKHS)** $\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space of linear functions from $\mathcal{X}$ to $\mathbb{R}$ with a reproducing kernel $k(\mathrm{x}, \mathrm{x}')$ on $\mathcal{X} \times \mathcal{X}$ such that

$$\begin{aligned} &\forall \mathrm{x} \in \mathcal{X}, \quad k(\mathrm{x}, \cdot) \in \mathcal{H} \\ &\forall f \in \mathcal{H}, \quad \langle f, k(\mathrm{x}, \cdot) \rangle_{\mathcal{H}} = f(\mathrm{x}) \quad \text{(reproducing property)} \end{aligned}$$

Important: RKHS defines a space of pointwise defined functions! In ML, we do have to do predictions for each point!

Uniqueness of RKHS

Theorem (Moore)

If k is a positive definite kernel then there exists a unique *reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is k.*

Uniqueness of RKHS

Theorem (Moore)

If k is a positive definite kernel then there exists a unique *reproducing kernel Hilbert space $\mathcal{H}$ whose kernel is k.*

Important observations:

- there is a **one-to-one** relation between reproducing kernel Hilbert spaces and positive definite kernels.
- there exist an embedding $\Phi : \mathcal{X} \to \mathcal{H}$, with $x \to \Phi(\mathrm{x}) \in \mathcal{H}$ of the input space into a Hilbert space (a.k.a. **feature space**),
- the embedding is not unique given a kernel but there exists one and they are all isometric isomorphic, the easiest is
$$\Phi : \mathrm{x} \to k(\mathrm{x}, \cdot) \in \mathcal{H}_k.$$
- kernel allows us to compute the inner product between the embedded vectors $k(\mathrm{x}, \mathrm{x}') = \langle \Phi(\mathrm{x}), \Phi(\mathrm{x}') \rangle_{\mathcal{H}}$,

Positive Definite Kernels

Definition (Positive definite (PD) kernel)

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is **positive definite (PD)** if for all $m \geq 1$, $\mathrm{x}_1, \ldots, \mathrm{x}_m \in \mathcal{X}$, $c_1, \ldots, c_m \in \mathbb{R}$

$$\sum_{i,j=1}^{m} c_i c_j k(\mathrm{x}_i, \mathrm{x}_j) \geq 0$$

The set of all real-valued positive definite kernels on $\mathcal{X}$ is denoted $\mathbb{R}_+^{\mathcal{X}\times\mathcal{X}}$.

Remark:

- In this lecture a kernel is always positive definite if not stated otherwise.
- Note that $\mathcal{X}$ is a general set $\implies$ later on we will define kernels on structured domains (graphs, histograms, etc.).

Gram Matrix

Definition (Gram Matrix)

Given a kernel k and a set of n points $\mathrm{x}_1, \ldots, \mathrm{x}_n \in \mathcal{X}$ the $n \times n$ matrix

$$K = (k(\mathrm{x}_i, \mathrm{x}_j))_{ij},$$

is called the **kernel matrix** (or **Gram Matrix**) K of the kernel k with respect to $\mathrm{x}_1, \ldots, \mathrm{x}_n$.

$$\phi^\top \phi = K = \begin{bmatrix} k(\mathrm{x}_1, \mathrm{x}_1) & k(\mathrm{x}_1, \mathrm{x}_2) & \ldots & k(\mathrm{x}_1, \mathrm{x}_n) \\ k(\mathrm{x}_2, \mathrm{x}_1) & k(\mathrm{x}_2, \mathrm{x}_2) & \ldots & k(\mathrm{x}_2, \mathrm{x}_n) \\ \vdots & & \vdots & \\ k(\mathrm{x}_n, \mathrm{x}_1) & k(\mathrm{x}_n, \mathrm{x}_2) & \ldots & k(\mathrm{x}_n, \mathrm{x}_n) \end{bmatrix}$$

Observation: The Gram matrix is therefore positive definite!

Examples of Kernels

- **Linear kernel:** $k(\mathsf{x}, \mathsf{x}') = \mathsf{x} \cdot \mathsf{x}'$
  - $\phi(\mathsf{x}) = \mathsf{x}$ and $\mathcal{H} = \mathbb{R}$
  - $\phi(\mathsf{x}) = (\mathsf{x}/\sqrt{2}, \mathsf{x}/\sqrt{2})$ and $\mathcal{H} = \mathbb{R}^2$

Examples of Kernels

- **Linear kernel:** $k(\mathrm{x}, \mathrm{x}') = \mathrm{x} \cdot \mathrm{x}'$
  - $\phi(\mathrm{x}) = \mathrm{x}$ and $\mathcal{H} = \mathbb{R}$
  - $\phi(\mathrm{x}) = (\mathrm{x}/\sqrt{2}, \mathrm{x}/\sqrt{2})$ and $\mathcal{H} = \mathbb{R}^2$
- **Polynomial kernel:** $k(\mathrm{x}, \mathrm{x}') = (\mathrm{x} \cdot \mathrm{x}' + c)^m$ for $c \geq 0$
  - $\phi(\mathrm{x}) = \left(\sqrt{\binom{m}{n_1, \ldots, n_{d+1}}} x_1^{n_1} \cdots x_d^{n_d} \cdot c^{n_{d+1}/2}\right)_{n_i \geq 0, \sum_{i=1}^{d+1} n_i = m}$
  - $\dim(\mathcal{H}) = \binom{d+m}{m}$

Examples of Kernels

- **Linear kernel:** $k(\mathrm{x}, \mathrm{x}') = \mathrm{x} \cdot \mathrm{x}'$
  - $\phi(\mathrm{x}) = \mathrm{x}$ and $\mathcal{H} = \mathbb{R}$
  - $\phi(\mathrm{x}) = (\mathrm{x}/\sqrt{2}, \mathrm{x}/\sqrt{2})$ and $\mathcal{H} = \mathbb{R}^2$
- **Polynomial kernel:** $k(\mathrm{x}, \mathrm{x}') = (\mathrm{x} \cdot \mathrm{x}' + c)^m$ for $c \geq 0$
  - $\phi(\mathrm{x}) = \left(\sqrt{\binom{m}{n_1, \dots, n_{d+1}}} x_1^{n_1} \cdots x_d^{n_d} \cdot c^{n_{d+1}/2}\right)_{n_i \geq 0, \sum_{i=1}^{d+1} n_i = m}$
  - $\dim(\mathcal{H}) = \binom{d+m}{m}$
- **Exponential kernel:** $k(\mathrm{x}, \mathrm{x}') = \exp(\langle \mathrm{x}, \mathrm{x}' \rangle)$
  - Assume $\mathrm{x} \in \mathbb{R}$ and use Taylor series expansion of e^{x},

$$\phi(x) = \left[1, x, \sqrt{\frac{1}{2!}} x^2, \sqrt{\frac{1}{3!}} x^3, \dots\right], \quad \mathcal{H} = \mathbb{R}^\infty$$

Examples of Kernels

- **Linear kernel:** $k(\mathrm{x}, \mathrm{x}') = \mathrm{x} \cdot \mathrm{x}'$
  - $\phi(\mathrm{x}) = \mathrm{x}$ and $\mathcal{H} = \mathbb{R}$
  - $\phi(\mathrm{x}) = (\mathrm{x}/\sqrt{2}, \mathrm{x}/\sqrt{2})$ and $\mathcal{H} = \mathbb{R}^2$
- **Polynomial kernel:** $k(\mathrm{x}, \mathrm{x}') = (\mathrm{x} \cdot \mathrm{x}' + c)^m$ for $c \geq 0$
  - $\phi(\mathrm{x}) = \left(\sqrt{\binom{m}{n_1, \ldots, n_{d+1}}} x_1^{n_1} \cdots x_d^{n_d} \cdot c^{n_{d+1}/2}\right)_{n_i \geq 0, \sum_{i=1}^{d+1} n_i = m}$
  - $\dim(\mathcal{H}) = \binom{d+m}{m}$
- **Exponential kernel:** $k(\mathrm{x}, \mathrm{x}') = \exp(\langle \mathrm{x}, \mathrm{x}' \rangle)$
  - Assume $\mathrm{x} \in \mathbb{R}$ and use Taylor series expansion of e^{x},

$$\phi(x) = \left[1, x, \sqrt{\frac{1}{2!}}x^2, \sqrt{\frac{1}{3!}}x^3, \ldots\right], \quad \mathcal{H} = \mathbb{R}^\infty$$

- **Gaussian RBF kernel:** $k(\mathrm{x}, \mathrm{x}') = \exp(-\gamma \|\mathrm{x} - \mathrm{x}'\|_2^2)$
  - Assume $\mathrm{x} \in \mathbb{R}$ and use Taylor series expansion of e^{x},

$$\phi(x) = \exp(-\gamma x^2) \left[1, \sqrt{\frac{2\gamma}{1!}}x, \sqrt{\frac{(2\gamma)^2}{2!}}x^2, \sqrt{\frac{(2\gamma)^3}{3!}}x^3, \ldots\right], \quad \mathcal{H} = \mathbb{R}^\infty$$

Outline

Bibliography

Motivation

Kernels

Properties

Vector space structure

Appendix

Properties of kernels I

Properties of kernels: Transformations of kernels which preserve the property of positive definiteness are important for

1. the construction of new kernels,
2. the verification that a given function $k(\mathrm{x}, \mathrm{x}')$ is positive definite, as we often have a similarity measure, we would like to use.

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.
2. $k(\mathrm{x}, \mathrm{x}') = f(\mathrm{x}) k_1(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')$, where $f(\cdot)$ is any function.
3. $k(\mathrm{x}, \mathrm{x}') = f(k_1(\mathrm{x}, \mathrm{x}'))$, where $f(\cdot)$ is a polynomial with non-negative coefficients.
4. $k(\mathrm{x}, \mathrm{x}') = \exp(k_1(\mathrm{x}, \mathrm{x}'))$

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.
2. $k(\mathrm{x}, \mathrm{x}') = f(\mathrm{x}) k_1(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')$, where $f(\cdot)$ is any function.
3. $k(\mathrm{x}, \mathrm{x}') = f(k_1(\mathrm{x}, \mathrm{x}'))$, where $f(\cdot)$ is a polynomial with non-negative coefficients.
4. $k(\mathrm{x}, \mathrm{x}') = \exp(k_1(\mathrm{x}, \mathrm{x}'))$
5. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') + k_2(\mathrm{x}, \mathrm{x}')$.
6. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') k_2(\mathrm{x}, \mathrm{x}')$.

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.
2. $k(\mathrm{x}, \mathrm{x}') = f(\mathrm{x}) k_1(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')$, where $f(\cdot)$ is any function.
3. $k(\mathrm{x}, \mathrm{x}') = f(k_1(\mathrm{x}, \mathrm{x}'))$, where $f(\cdot)$ is a polynomial with non-negative coefficients.
4. $k(\mathrm{x}, \mathrm{x}') = \exp(k_1(\mathrm{x}, \mathrm{x}'))$
5. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') + k_2(\mathrm{x}, \mathrm{x}')$.
6. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') k_2(\mathrm{x}, \mathrm{x}')$.
7. $k(\mathrm{x}, \mathrm{x}') = k_1(\phi(\mathrm{x}), \phi(\mathrm{x}'))$, where $\phi : \mathcal{X} \mapsto \mathbb{R}^m$.
8. $k(\mathrm{x}, \mathrm{x}') = \mathrm{x}^T A \mathrm{x}'$ for a symmetric positive definite matrix A.

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.
2. $k(\mathrm{x}, \mathrm{x}') = f(\mathrm{x}) k_1(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')$, where $f(\cdot)$ is any function.
3. $k(\mathrm{x}, \mathrm{x}') = f(k_1(\mathrm{x}, \mathrm{x}'))$, where $f(\cdot)$ is a polynomial with non-negative coefficients.
4. $k(\mathrm{x}, \mathrm{x}') = \exp(k_1(\mathrm{x}, \mathrm{x}'))$
5. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') + k_2(\mathrm{x}, \mathrm{x}')$.
6. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') k_2(\mathrm{x}, \mathrm{x}')$.
7. $k(\mathrm{x}, \mathrm{x}') = k_1(\phi(\mathrm{x}), \phi(\mathrm{x}'))$, where $\phi : \mathcal{X} \mapsto \mathbb{R}^m$.
8. $k(\mathrm{x}, \mathrm{x}') = \mathrm{x}^T A \mathrm{x}'$ for a symmetric positive definite matrix A.
9. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}_a, \mathrm{x}'_a) + k_2(\mathrm{x}_b, \mathrm{x}'_b)$ or $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}_a, \mathrm{x}'_a) k_2(\mathrm{x}_b, \mathrm{x}'_b)$, with $\mathrm{x} = (\mathrm{x}_a, \mathrm{x}_b)$ and variables x_a and x_b are not necessarily disjoint.

Properties

Let k_1, k_2 be kernels on $\mathcal{X}$. Then, the following properties allow us to build complicated kernels from simpler ones (k_1, k_2).

1. $k(\mathrm{x}, \mathrm{x}') = \alpha k_1(\mathrm{x}, \mathrm{x}')$ for $\alpha > 0$.
2. $k(\mathrm{x}, \mathrm{x}') = f(\mathrm{x}) k_1(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')$, where $f(\cdot)$ is any function.
3. $k(\mathrm{x}, \mathrm{x}') = f(k_1(\mathrm{x}, \mathrm{x}'))$, where $f(\cdot)$ is a polynomial with non-negative coefficients.
4. $k(\mathrm{x}, \mathrm{x}') = \exp(k_1(\mathrm{x}, \mathrm{x}'))$
5. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') + k_2(\mathrm{x}, \mathrm{x}')$.
6. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}, \mathrm{x}') k_2(\mathrm{x}, \mathrm{x}')$.
7. $k(\mathrm{x}, \mathrm{x}') = k_1(\phi(\mathrm{x}), \phi(\mathrm{x}'))$, where $\phi : \mathcal{X} \mapsto \mathbb{R}^m$.
8. $k(\mathrm{x}, \mathrm{x}') = \mathrm{x}^T A \mathrm{x}'$ for a symmetric positive definite matrix A.
9. $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}_a, \mathrm{x}'_a) + k_2(\mathrm{x}_b, \mathrm{x}'_b)$ or $k(\mathrm{x}, \mathrm{x}') = k_1(\mathrm{x}_a, \mathrm{x}'_a) k_2(\mathrm{x}_b, \mathrm{x}'_b)$, with $\mathrm{x} = (\mathrm{x}_a, \mathrm{x}_b)$ and variables x_a and x_b are not necessarily disjoint.

Outline

Bibliography

Motivation

Kernels

Properties

Vector space structure

Appendix

Vector space structure in $\mathcal{H}$ (I)

Norm of a feature vector $\Phi(x)$

$$\|\Phi(\mathrm{x})\|_{\mathcal{H}} = \sqrt{\|\Phi(\mathrm{x})\|_{\mathcal{H}}^2} = \sqrt{k(\mathrm{x},\mathrm{x})}.$$

Distance/Metric:

$$\begin{aligned} d(\mathrm{x},\mathrm{x}') &= \sqrt{\|\Phi(\mathrm{x}) - \Phi(\mathrm{x}')\|_{\mathcal{H}}^2} = \sqrt{\|\Phi(\mathrm{x})\|_{\mathcal{H}}^2 + \|\Phi(\mathrm{x}')\|_{\mathcal{H}}^2 - 2\left\langle \Phi(\mathrm{x}), \Phi(\mathrm{x}')\right\rangle_{\mathcal{H}}} \\ &= \sqrt{k(\mathrm{x},\mathrm{x}) + k(\mathrm{x}',\mathrm{x}') - 2k(\mathrm{x},\mathrm{x}')}. \end{aligned}$$

Angle:

$$\cos\left(\angle(\Phi(\mathrm{x}), \Phi(\mathrm{x}'))\right) = \frac{\langle \Phi(\mathrm{x}), \Phi(\mathrm{x}')\rangle}{\|\Phi(\mathrm{x})\|\,\|\Phi(\mathrm{x}')\|} = \frac{k(\mathrm{x},\mathrm{x}')}{\sqrt{k(\mathrm{x},\mathrm{x})k(\mathrm{x}',\mathrm{x}')}}.$$

Vector space structure in $\mathcal{H}$ (II)

The kernel function as similarity measure between x and x':

This interpretation is motivated by the fact that the kernel function $k(\mathrm{x}, \mathrm{x}')$ can be seen as an inner product

$$k(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle \,,$$

where $\phi : \mathcal{X} \rightarrow \mathcal{H}$ and $\mathcal{H}$ is a **Hilbert space**.

A new kernel $\tilde{k}$ defined as

$$\tilde{k}(\mathrm{x}, \mathrm{x}') = \frac{k(\mathrm{x}, \mathrm{x}')}{\sqrt{k(\mathrm{x}, \mathrm{x}) k(\mathrm{x}', \mathrm{x}')}} = \frac{\langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle}{\|\phi(\mathrm{x})\| \, \|\phi(\mathrm{x}')\|} = \cos(\angle(\phi(\mathrm{x}), \phi(\mathrm{x}'))),$$

is again a positive definite kernel.

- The cosine is a common similarity measure (text classification).
- $|\tilde{k}(\mathrm{x}, \mathrm{x}')| \leq 1$ and $\tilde{k}(\mathrm{x}, \mathrm{x}') = 1$ if and only if $\mathrm{x} = \mathrm{x}'$.

Vector space structure in $\mathcal{H}$ (III)

Why are (dis)similarity measures useful for learning?

- one needs only to define the similarity $k(\mathrm{x}, \mathrm{x}')$ of two points x and x' instead of a set of functions $\phi(\mathrm{x})$ in the feature maps approach.
- construction of a similarity measure between structured objects e.g. graphs is conceptually much easier than defining a certain set of feature maps on these structured objects.

Vector space structure in $\mathcal{H}$ (IV)

Center of mass/Centroid/Mean vector:

$$\Phi_m = \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathrm{x}_i).$$

Distance of $\Phi(\mathrm{x})$ to the center of mass:

$$\begin{aligned} \left\|\Phi(\mathrm{x}) - \Phi_m\right\|^2 &= \langle \Phi(\mathrm{x}), \Phi(\mathrm{x}) \rangle + \langle \Phi_m, \Phi_m \rangle - 2 \langle \Phi(\mathrm{x}), \Phi_m \rangle \\ &= k(\mathrm{x}, \mathrm{x}) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(\mathrm{x}_i, \mathrm{x}_j) - \frac{2}{n} \sum_{i=1}^{n} k(\mathrm{x}, \mathrm{x}_i). \end{aligned}$$

Centering of datapoints in the feature space:

$$\tilde{\Phi}(\mathrm{x}) = \Phi(\mathrm{x}) - \Phi_m = \Phi(\mathrm{x}) - \frac{1}{n} \sum_{i=1}^{n} \Phi(\mathrm{x}_i).$$

$$\left\langle \tilde{\Phi}(\mathrm{x}), \tilde{\Phi}(\mathrm{x}') \right\rangle = k(\mathrm{x}, \mathrm{x}') - \frac{1}{n} \sum_{i=1}^{n} k(\mathrm{x}, \mathrm{x}_i) - \frac{1}{n} \sum_{i=1}^{n} k(\mathrm{x}', \mathrm{x}_i) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(\mathrm{x}_i, \mathrm{x}_j).$$

Example: Parzen window classifier

Center of mass/Centroid/Mean vector:

$$\Phi_m^+ = \frac{1}{n_+}\sum_{i=1}^{n_+}\Phi(\mathrm{x}_i^+), \quad \Phi_m^- = \frac{1}{n_-}\sum_{j=1}^{n_-}\Phi(\mathrm{x}_j^-).$$

Classify by assigning point to class of closest centroid:

$$\begin{aligned} f(x) &= \operatorname{sign}\left(\left\|\Phi(\mathrm{x}) - \Phi_m^-\right\|^2 - \left\|\Phi(\mathrm{x}) - \Phi_m^+\right\|^2\right) \\ &= \operatorname{sign}\left(\frac{2}{n_+}\sum_{i=1}^{n_+} k(\mathrm{x}, \mathrm{x}_i^+) - \frac{2}{n_-}\sum_{j=1}^{n_-} k(\mathrm{x}, \mathrm{x}_j^-) + b\right) \end{aligned}$$

where

$$b = \frac{1}{n_-^2}\sum_{i,j=1}^{n_-} k(\mathrm{x}_i^-, \mathrm{x}_i^-) - \frac{1}{n_+^2}\sum_{i,j=1}^{n_+} k(\mathrm{x}_i^+, \mathrm{x}_j^+).$$

This is a so called **Parzen window classifier** (with offset).

Outline

Bibliography

Motivation

Kernels

Properties

Vector space structure

Appendix

Purpose

Disclosure: The following appendix aims to provide further mathematical details on how to construct RKHS for those that are interested. However, the following material is not necessary or required to prepare for the exam.

Basics from functional analysis

Definition

A **metric space** is a set $\mathcal{X}$ with a distance function $d : \mathcal{X} \times X \to \mathbb{R}$ such that:

- $d(x, x') \geq 0$,
- $d(x, x') = 0$ if and only if $x = x'$,
- $d(x, x') = d(\mathrm{x}, \mathrm{x}')$, (symmetry)
- $d(x, x') \leq d(x, z) + d(z, x')$. (triangle inequality)

It is denoted as $(\mathcal{X}, d)$.
For a **semi-metric** $d(x, x') = 0$ does not imply $x = x'$.

Remark: any semi-metric space $(\mathcal{X}, d)$ can be turned into a metric space by identifying points which have zero distance.

Convergence and Cauchy sequences

Definition

A sequence of elements $\{x_n\}_{n \in \mathbb{N}}$ of a metric space $(\mathcal{X}, d)$ is said to **converge** to an element $x \in \mathcal{X}$ if $\lim_{n\to\infty} d(x, x_n) = 0$. We will denote this either as $x_n \xrightarrow{d} x$ or $\lim_{n\to\infty} x_n = x$.

Definition

A sequence of elements $\{x_n\}$ of a metric space $(\mathcal{X}, d)$ is called a **Cauchy sequence** if $\forall\, \epsilon > 0, \quad \exists N$ such that $d(x_n, x_m) < \epsilon, \; \forall\, n, m > N$.

Proposition: Every convergent sequence is a Cauchy sequence.

Definition

A metric space in which all Cauchy sequences converge is called **complete**.

Example: $\mathbb{R}$ is complete, but not $\mathbb{Q}$.

Function spaces as vector spaces

Sets of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ as vector spaces - apply vector axioms pointwise.

Three functions $f, g, h \in \mathcal{F}$, $\alpha, \beta \in \mathbb{R}$,

$$(f+g)(x) := f(x) + g(x), \quad \forall x \in \mathcal{X},$$
$$(\alpha f)(x) := \alpha\, f(x), \quad \forall x \in \mathcal{X}.$$

Associativity	$\big(f(x)+g(x)\big)+h(x) = f(x) + \big(g(x)+h(x)\big)$,
Commutativity	$f(x)+g(x) = g(x)+f(x)$,
Identity (addition)	$f(x)+0 = f(x)$, $\Rightarrow$ zero function $h(x) = 0$, $\forall x \in \mathcal{X}$,
Distributivity I	$(\alpha+\beta)f(x) = \alpha f(x) + \beta f(x)$,
Distributivity II	$\alpha(f(x)+g(x)) = \alpha f(x) + \alpha g(x)$,
Compatibility	$(\alpha\beta)f(x) = \alpha(\beta f(x))$,
Identity (multiplication)	$(1f(x)) = f(x)$.

Examples of function spaces

Sets of functions as vector spaces:

- all linear functions (finite dimensional),
- all polynomials (infinite dimensional),
- given a set of functions $\{\phi_1, \dots, \phi_D\}$, they generate an D-dimensional vector space by taking all linear combinations:

$$\mathcal{F} = \mathrm{span}\{\phi_1, \dots, \phi_D\} := \left\{ \sum_{i=1}^{D} \alpha_i \phi_i \,\middle|\, \alpha_i \in \mathbb{R}, \quad i = 1, \dots, D \right\}.$$

$\Longrightarrow$ given that the functions are linearly independent,

$$\sum_{i=1}^{D} c_i \phi_i(x) = 0, \quad \forall\, x \in \mathcal{X}, \quad \Longrightarrow \quad c_i = 0, \quad i = 1, \dots, D,$$

$\{\phi_1, \dots, \phi_D\}$ is then also a basis of $\mathcal{F}$ (by definition).

Inner product spaces

Definition

A real vector space V is called an **inner product space** if there is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the following four conditions $\forall\, x, y, z \in V$ and $\forall \alpha \in \mathbb{R}$:

1. $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$,
2. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$,
3. $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$,
4. $\langle x, y \rangle = \langle y, x \rangle$.

The function $\langle \cdot, \cdot \rangle$ is called **inner product**.

Every inner product defines a norm, $\|x\| := \sqrt{\langle x, x \rangle}$, and a metric, $d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle}$.
On inner product spaces we have the **Cauchy-Schwarz inequality:**

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| \, .$$

Example - Inner product spaces

An inner product on functions: Let $f, g : \mathcal{X} \to \mathbb{R}$,

$$\langle f, g \rangle := \int_{\mathcal{X}} f(x) g(x) dx.$$

We obtain:

- $\langle f, f \rangle = \int_{\mathcal{X}} \big(f(x)\big)^2 dx \geq 0$,
- $\langle f, f \rangle = 0$ if and only if $f = 0$ (almost everywhere),
- $\langle f, g + h \rangle = \int_{\mathcal{X}} f(x)\big(g(x) + h(x)\big) dx = \int_{\mathcal{X}} f(x) g(x) dx + \int_{\mathcal{X}} f(x) h(x) dx$,
- $\langle f, \alpha g \rangle = \int_{\mathcal{X}} f(x)(\alpha\, g)(x) dx = \alpha \int_{\mathcal{X}} f(x) g(x) dx = \alpha \langle f, g \rangle$,
- $\langle f, g \rangle = \int_{\mathcal{X}} f(x) g(x) dx = \int_{\mathcal{X}} g(x) f(x) dx = \langle g, f \rangle$.

Induced norm

$$\|f\| = \sqrt{\langle f, f \rangle} = \left(\int_{\mathcal{X}} \big(f(x)\big)^2 dx \right)^{\frac{1}{2}}.$$

Hilbert Space I

Definition

A complete inner product space is called a **Hilbert space**.

The space $L_2(\mathcal{X})$ of square-integrable functions:

$$L_2(\mathcal{X}) := \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, \int_{\mathcal{X}} \left(f(x)\right)^2 dx < \infty \right\},$$

is a Hilbert space together with the inner product,

$$\langle f, g \rangle := \int_{\mathcal{X}} f(x) g(x) dx.$$

but !

- $\|f\| = 0$ if and only if f is zero almost everywhere !
- Functions which agree almost everywhere are identified (the relation equal almost everywhere defines equivalence classes of functions),
- $L_2(\mathcal{X})$ is not a space of pointwise defined functions !

Hilbert space II

Definition

If S is an orthonormal set in a Hilbert space $\mathcal{H}$ and no other orthonormal set contains S as a proper subset, then S is called an **orthonormal basis** (or a **complete orthonormal system**) for $\mathcal{H}$.

Theorem

Let $\mathcal{H}$ be a Hilbert space and $S = \{x_\alpha\}_{\alpha \in A}$ an orthonormal basis. Then for each $y \in \mathcal{H}$,

$$y = \sum_{\alpha \in A} \langle x_\alpha, y \rangle \, x_\alpha, \qquad \|y\|^2 = \sum_{\alpha \in A} | \langle x_\alpha, y \rangle |^2.$$

The coefficients $\langle x_\alpha, y \rangle$ are called the **coefficients** of y with respect to the basis $\{x_\alpha\}$.

Construction of RKHSs I

Steps for the construction of a RKHS:

- consider the set of all finite linear combinations of the kernel:

$$\mathcal{G} = \mathrm{Span}\{k(x,.) : x \in \mathcal{X}\}$$

- Let $f(x) = \sum_i a_i k(x_i, x)$ and $g(x) = \sum_j b_j k(z_j, x)$. Then

$$\left\langle \sum_i a_i k(x_i,.), \sum_j b_j k(z_j,.) \right\rangle_{\mathcal{G}} := \sum_{i,j} a_i\, b_j\, k(x_i, z_j)\,.$$

- check that $\langle \cdot, \cdot \rangle$ is well-defined.

$$\sum_{i,j} a_i\, b_j\, k(x_i, z_j) = \sum_i a_i g(x_i) = \sum_j b_j f(z_j)$$

The value of the inner product does not depend on the expansion of f or g $\Rightarrow$ (semi)-inner product with the reproducing property on $\mathcal{G}$.

Construction of RKHSs II

Steps for the construction of a RKHS:

- construct the **semi-norm** associated to this inner product,

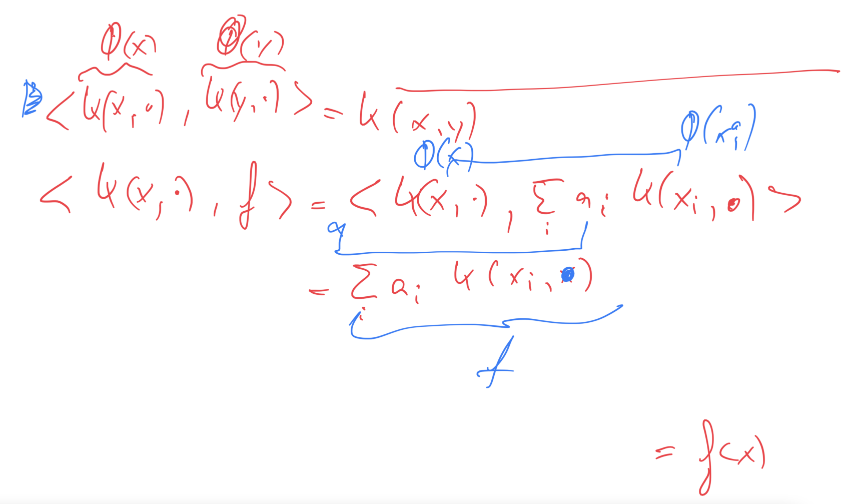
  $$\|f\|_{\mathcal{G}}^2 = \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j).$$

  The Cauchy-Schwarz inequality holds also on semi-inner product spaces,

  $$|f(x)| = |\langle f, k(x,.)\rangle_{\mathcal{G}}| \leq \|f\|_{\mathcal{G}} \|k(x,.)\|_{\mathcal{G}} = \|f\|_{\mathcal{G}} \sqrt{k(x,x)}.$$

  $\|f\|_{\mathcal{G}} = 0$ implies $f \equiv 0 \Rightarrow$ inner product on $\mathcal{G}$ and $\mathcal{G}$ is an **inner product space**.
- Standard completion by adding all limits of Cauchy sequences in $\mathcal{G}$
  - one has to check that the inner product as well as the reproducing property carries over to the limit elements.

Reproducing Kernel Hilbert Space



$$\langle \overbrace{k(x,\cdot)}^{\phi(x)}, \overbrace{k(y,\cdot)}^{\phi(y)} \rangle = k(x,y)$$

$$\langle k(x,\cdot), f \rangle = \underbrace{\langle \overbrace{k(x,\cdot)}^{\phi(x)}, \sum_i a_i}_{a} \overbrace{k(x_i,\cdot)}^{\phi(x_i)} \rangle$$

$$= \underbrace{\sum_i a_i \, k(x_i,\cdot)}_{f}$$

$$= f(x)$$

... refer to video lecture ...