CS:E4830 Kernel Methods in Machine Learning

Lecture 2 : Reproducing Kernel Hilbert Space

Rohit Babbar

16th January, 2019

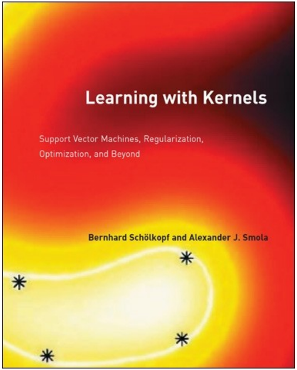
Outline

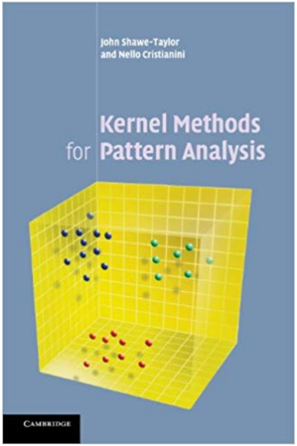
From the preface of a book

albeit with heuristic algorithms and incomplete statistical analysis. The impact of the nonlinear revolution cannot be overemphasised: entire fields such as data mining and bioinformatics were enabled by it. These nonlinear algorithms, however, were based on gradient descent or greedy heuristics and so suffered from local minima. Since their statistical behaviour was not well understood, they also frequently suffered from overfitting.

A third stage in the evolution of pattern analysis algorithms took place in the mid-1990s with the emergence of a new approach to pattern analysis known as kernel-based learning methods that finally enabled researchers to analyse nonlinear relations with the efficiency that had previously been reserved for linear algorithms. Furthermore advances in their statistical analysis made it possible to do so in high-dimensional feature spaces while avoiding the dangers of overfitting. From all points of view, computational, statistical and conceptual, the nonlinear pattern analysis algorithms developed in this third generation are as efficient and as well founded as linear ones. The problems of local minima and overfitting that were typical of neural networks and decision trees have been overcome. At the same time, Books for further study

- Learning with kernels - Schoelkopf and Smola
- Kernel Methods for Pattern Analysis - Shawe-Taylor and Christianini

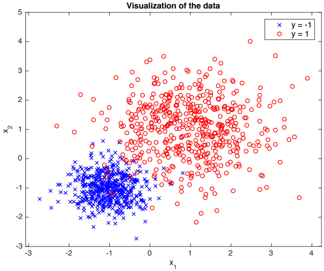




Parzen Window Classifier - Problem setup

A simple binary classification scheme

- Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}, x_i \in \mathcal{X}, y_i \in \{-1, +1\}$ be the training set that contains m_+ positive examples and m_- negative examples.
- Let $I = \{1, \ldots, m = m_+ + m_-\}$ be the indices of the training examples.
- $I^+ = \{i \in I | y_i = +1\}$ the set containing the indices of the positive training examples. Similarly $I^- = \{i \in I | y_i = -1\}$ for the negative training examples.
- $k(.,.)$ is a kernel defined on $\mathcal{X} \times \mathcal{X}$, and ϕ is a feature map associated with this kernel.
- Let $c_+ = \frac{1}{m_+} \sum_{i \in I^+} \phi(x_i)$ and $c_- = \frac{1}{m_-} \sum_{i \in I^-} \phi(x_i)$ be the means of the two classes in the feature space.



Parzen Window Classifier - Kernelization

Given a new point $x \in \mathcal{X}$ to classify, the idea of the Parzen window classifier is to assign x to the closest class in the feature space:

$$h(x) = \begin{cases} +1 \text{ if } ||\phi(x) - c_-||^2 > ||\phi(x) - c_+||^2 \\ -1 \text{ otherwise.} \end{cases}$$

The function h can be expressed using the sign function:

$$h(x) = \text{sgn}(||\phi(x) - c_-||^2 - ||\phi(x) - c_+||^2).$$

(Practice Exercise) To show that the function h can be written as:
$h(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i k(x, x_i) + b\right)$, where

$$b = \frac{1}{2m_-^2} \sum_{i,j \in I^-} k(x_i, x_j) - \frac{1}{2m_+^2} \sum_{i,j \in I^+} k(x_i, x_j),$$

$$\alpha_i = \begin{cases} \dfrac{1}{m_+} \text{ if } y_i = +1 \\ \dfrac{-1}{m_-} \text{ if } y_i = -1 \end{cases}$$

Parzen Window Classifier - Kernelization

Given a new point $x \in \mathcal{X}$ to classify, the idea of the Parzen window classifier is to assign x to the closest class in the feature space:

$$h(x) = \begin{cases} +1 \text{ if } ||\phi(x) - c_-||^2 > ||\phi(x) - c_+||^2 \\ -1 \text{ otherwise.} \end{cases}$$

The function h can be expressed using the sign function:

$$h(x) = \text{sgn}(||\phi(x) - c_-||^2 - ||\phi(x) - c_+||^2).$$

(Practice Exercise) To show that the function h can be written as:
$h(x) = \text{sgn}\left(\sum_{i=1}^{m} \alpha_i k(x, x_i) + b\right)$, where

$$b = \frac{1}{2m_-^2} \sum_{i,j \in I^-} k(x_i, x_j) - \frac{1}{2m_+^2} \sum_{i,j \in I^+} k(x_i, x_j),$$

$$\alpha_i = \begin{cases} \dfrac{1}{m_+} \text{ if } y_i = +1 \\ \dfrac{-1}{m_-} \text{ if } y_i = -1 \end{cases}$$

Hint : Write $||\phi(x) - c_-||^2$ as $\langle \phi(x) - c_-, \phi(x) - c_- \rangle$

Positive Definiteness

So Far...

We have seen two ways for checking if a symmetric function with two arguments is a kernel or not?

So Far...

We have seen two ways for checking if a symmetric function with two arguments is a kernel or not?

1. Finding a feature map $\phi(.)$, and a Hilbert space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$

So Far...

We have seen two ways for checking if a symmetric function with two arguments is a kernel or not?

1. Finding a feature map $\phi(.)$, and a Hilbert space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$
   - Not always easy to find the feature map, can be infinite dimensional as in the case of Gaussian kernel

So Far...

We have seen two ways for checking if a symmetric function with two arguments is a kernel or not?

1. Finding a feature map $\phi(.)$, and a Hilbert space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$
    - Not always easy to find the feature map, can be infinite dimensional as in the case of Gaussian kernel
2. Checking if the function can be constructed using the conic sum and product of kernels

So Far...

We have seen two ways for checking if a symmetric function with two arguments is a kernel or not?

1. Finding a feature map $\phi(.)$, and a Hilbert space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$
   - Not always easy to find the feature map, can be infinite dimensional as in the case of Gaussian kernel
2. Checking if the function can be constructed using the conic sum and product of kernels

Next - **Positive-definiteness** of the kernel function

Positive Definite Functions

Definition - Positive definite functions

A symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Caution : Here we are referring to $k(.,.)$ as a normal symmetric function with two arguments, and not a kernel

Positive Definite Functions

Definition - Positive definite functions

A symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Caution : Here we are referring to $k(.,.)$ as a normal symmetric function with two arguments, and not a kernel

- For example - Gaussian kernel is a positive definite function

Positive Definite Functions

Definition - Positive definite functions

A symmetric function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Caution : Here we are referring to $k(.,.)$ as a normal symmetric function with two arguments, and not a kernel

- For example - Gaussian kernel is a positive definite function
- In fact, as we will see next, all kernels are positive definite functions

As we will see next, all kernels are positive definite functions and **vice-versa**

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j)$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i\phi(x_i), a_j\phi(x_j)\rangle_{\mathcal{H}} = ||\sum_{i=1}^{n} a_i\phi(x_i)||_{\mathcal{H}}^2 \geq 0$$

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \langle a_i\phi(x_i), a_j\phi(x_j)\rangle_{\mathcal{H}} = ||\sum_{i=1}^{n} a_i\phi(x_i)||_{\mathcal{H}}^2 \geq 0$$

- Is that what we need to find if a function is a kernel or not ?

Characterization of a Kernel

Moore-Aronszajn Theorem

A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof in one direction

- Suppose $k(.,.)$ is a kernel. Then, surely it is symmetric. (Why?)
- Positive definiteness :

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}\langle a_i\phi(x_i), a_j\phi(x_j)\rangle_{\mathcal{H}} = ||\sum_{i=1}^{n} a_i\phi(x_i)||_{\mathcal{H}}^2 \geq 0$$

- Is that what we need to find if a function is a kernel or not ? No
- **Interesting bit** - The *converse* also holds i.e., which is what we require.
- *Converse* - A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle\phi(x), \phi(x')\rangle_{\mathcal{H}}$. (Proof after three slides)

Conic combination of Kernels

First, we see an implication of the converse, i.e., positive definite function is a valid kernel

- Earlier, we proved that conic sum $\sum_{j=1}^{K} \alpha_j k_j$ of kernels $(k_j)_{j=1}^{K}$, with positive co-efficients $(\alpha_j)_{j=1}^{K}$ is also a kernel by

Conic combination of Kernels

First, we see an implication of the converse, i.e., positive definite function is a valid kernel

- Earlier, we proved that conic sum $\sum_{j=1}^{K} \alpha_j k_j$ of kernels $(k_j)_{j=1}^{K}$, with positive co-efficients $(\alpha_j)_{j=1}^{K}$ is also a kernel by
  - Demonstrating the existence of a feature map and a feature space

Conic combination of Kernels

First, we see an implication of the converse, i.e., positive definite function is a valid kernel

- Earlier, we proved that conic sum $\sum_{j=1}^{K} \alpha_j k_j$ of kernels $(k_j)_{j=1}^{K}$, with positive co-efficients $(\alpha_j)_{j=1}^{K}$ is also a kernel by
  - Demonstrating the existence of a feature map and a feature space
- Now, we will use positive definiteness property to prove that conic sum of kernels is a kernel

Conic combination of Kernels

First, we see an implication of the converse, i.e., positive definite function is a valid kernel

- Earlier, we proved that conic sum $\sum_{j=1}^{K} \alpha_j k_j$ of kernels $(k_j)_{j=1}^{K}$, with positive co-efficients $(\alpha_j)_{j=1}^{K}$ is also a kernel by
  - Demonstrating the existence of a feature map and a feature space
- Now, we will use positive definiteness property to prove that conic sum of kernels is a kernel
- If $\alpha_1 k_1(.,.) + \ldots + \alpha_K k_K(.,.)$ is a kernel function, then it better be positive definite. Let's verify

Conic combination of Kernels

First, we see an implication of the converse, i.e., positive definite function is a valid kernel

- Earlier, we proved that conic sum $\sum_{j=1}^{K} \alpha_j k_j$ of kernels $(k_j)_{j=1}^{K}$, with positive co-efficients $(\alpha_j)_{j=1}^{K}$ is also a kernel by
  - Demonstrating the existence of a feature map and a feature space
- Now, we will use positive definiteness property to prove that conic sum of kernels is a kernel
- If $\alpha_1 k_1(.,.) + \ldots + \alpha_K k_K(.,.)$ is a kernel function, then it better be positive definite. Let's verify

Consider K kernels $k_1(.,.), \ldots, k_K(.,.)$ and $\alpha_1 \ldots \alpha_K > 0$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j [\alpha_1 k_1(x_i, x_j) + \ldots + \alpha_K k_K(x_i, x_j)]$$

$$= \alpha_1 \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k_1(x_i, x_j) + \ldots + \alpha_K \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k_2(x_i, x_j)$$

≥ 0 (Since each of the individual K terms is positive, so is the sum)

The kernel matrix

- A **kernel matrix** (also called the **Gram matrix**), is an $n \times n$ matrix of pairwise similarity values is used:

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_\ell, x_1) & k(x_\ell, x_2) & \dots & k(x_\ell, x_\ell) \end{bmatrix}$$

- Each entry is an inner product between two data points $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, where $\phi(.)$ is a feature map in vector form
- Since an inner product is symmetric, therefore K is a symmetric matrix
- In addition, K is positive definite (proof on next slide)

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

- Idea behind positive definiteness - Similar to positivity of a number but in matrix form

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

- Idea behind positive definiteness - Similar to positivity of a number but in matrix form
- The operation $v^T K v$ in the positive definiteness is a conversion of matrix in $\mathbb{R}^{n \times n}$ to a number in $\mathbb{R}$

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

- Idea behind positive definiteness - Similar to positivity of a number but in matrix form
- The operation $v^T K v$ in the positive definiteness is a conversion of matrix in $\mathbb{R}^{n \times n}$ to a number in $\mathbb{R}$
- This is done by taking dot product between v and Kv for every vector $v \in \mathbb{R}^n$

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

- Idea behind positive definiteness - Similar to positivity of a number but in matrix form
- The operation $v^T K v$ in the positive definiteness is a conversion of matrix in $\mathbb{R}^{n \times n}$ to a number in $\mathbb{R}$
- This is done by taking dot product between v and Kv for every vector $v \in \mathbb{R}^n$
- Then check if the resulting number is positive or not

The kernel matrix

Kernel Matrix is positive semi-definite

The kernel matrix K is positive semi-definite, i.e. $\forall v \in \mathbb{R}^n, v^T K v \geq 0$.

- Idea behind positive definiteness - Similar to positivity of a number but in matrix form
- The operation $v^T K v$ in the positive definiteness is a conversion of matrix in $\mathbb{R}^{n \times n}$ to a number in $\mathbb{R}$
- This is done by taking dot product between v and Kv for every vector $v \in \mathbb{R}^n$
- Then check if the resulting number is positive or not

Proof in one direction

$$v^T K v = \sum_{i,j=1}^{n} v_i K_{ij} v_j = \sum_{i,j=1}^{n} v_i \langle \phi(x_i), \phi(x_j) \rangle v_j =$$

$$= \langle \sum_{i=1}^{n} v_i \phi(x_i), \sum_{j=1}^{n} v_j \phi(x_j) \rangle = || \sum_{i=1}^{n} v_i \phi(x_i) ||^2 \geq 0$$

What implies what?

Function vs Matrix

We have seen,

- positive definiteness of a function, and
- positive definiteness of a matrix

What implies what?

Function vs Matrix

We have seen,

- positive definiteness of a function, and
- positive definiteness of a matrix

Which implies the other? In other words, which is a stronger requirement?

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space
- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**. How does that look like ?

Construction of Feature map and Feature space for a positive definite function

Moore-Aronszajn Theorem

A symmetric and positive definite function is a valid kernel, i.e. there exists a feature map $\phi(.)$ and a feature space $\mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.

Proof outline

Though, we just need to prove the existence of the feature (Hilbert) space and feature map. However, in this case, it is a proof by construction, meaning

- We construct **the** feature map and **the** feature space
- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**. How does that look like ?
- A possible candidate form of the function space is given by the set

$$\mathcal{H} = \left\{ \sum_{i=1}^{\ell} \alpha_i k(x_i, .) : \ell \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \ldots, \ell \right\}$$

Function in the candidate space of functions

How do typical functions $f(.)$ and $g(.)$ in $\mathcal{H}$ look like?

Function in the candidate space of functions

How do typical functions $f(.)$ and $g(.)$ in $\mathcal{H}$ look like?

- Consider the following :
  1. $f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .)$
  2. $g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$

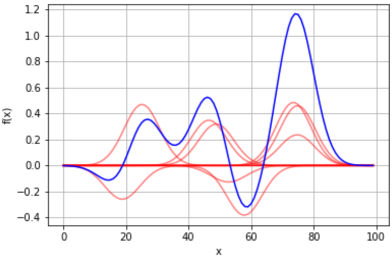


Figure: Pictorial depiction of a function $f(.)$ (in blue) as a linear combination of kernel functions $k(x_i, .)$ for Gaussian kernel evaluated at 9 points (Figure by Eric Bach)

Defining the Feature (function) space

Proof (1/3) - Defining the function Space

- **Key aspect** : The feature (Hilbert) space will be a space of functions, i.e., the points in the space are, in fact, **functions**.
- A candidate for the function space is given by the set of function given by $\mathcal{H}$

$$\mathcal{H} = \left\{ \sum_{i=1}^{\ell} \alpha_i k(x_i, .) : \ell \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, \ldots, \ell \right\}$$

Key points :

- These are non-linear functions in the input space (as shown in picture in last slide)
- Linear functions in the (possibly infinite dimensional) feature space
  - That is, these are of the form $f(x) = f_1\phi_1(x) + f_2\phi_2(x) + \ldots$ (more on this part later)

Defining Inner product on the space

Proof (2/3) - Verifying elementary properties and defining inner product

- It is a function (vector) space which satisfies the requirements of closure under scalar multiplication and addition
  - For $f \in \mathcal{H}, \gamma \in \mathbb{R} \implies \gamma f \in \mathcal{H}$
  - $f, g \in \mathcal{H} \implies (f+g) \in \mathcal{H}$
- Define Inner product on $\mathcal{H}$ as follows :
  - Let $f, g \in \mathcal{H}$ be given by

$$f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .) \text{ and } g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$$

  - The inner product between f and g is defined by the following

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(y_j) \tag{1}$$

which satisfies the symmetry and linearity properties of the inner product

Reproducing property

Proof (3/3) Using positive definiteness for positive of IP with itself

- Now, we need $\langle f, f \rangle_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}$
  - This follows from the positive definiteness of the given function

$$\langle f, f \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

- An important property that follows from equation (1) is obtained by taking $g = k(x, .)$,

$$\langle f, k(x, .) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i k(x_i, x) = f(x)$$

  - The above is called the **reproducing property** of the kernel $k(., .)$.
  - Also, $f(x)$ has the form $\langle f, k(x, .) \rangle_{\mathcal{H}}$, i.e, the evaluation of f at x is in the form of an inner product (linear function) between f and a feature map $\phi(x) = k(x, .)$, which is called the **canonical feature map** for kernel $k(., .)$

Reproducing Kernel Hilbert Space

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. $\mathcal{H}$ is defined to an **Reproducing kernel Hilbert Space (RKHS)** with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel, if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x) \rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. $\mathcal{H}$ is defined to an **Reproducing kernel Hilbert Space (RKHS)** with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel, if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x) \rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

- The reproducing kernel $k(., .)$ of the Hilbert space $\mathcal{H}$ is a kernel.

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. $\mathcal{H}$ is defined to an **Reproducing kernel Hilbert Space (RKHS)** with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel, if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x)\rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

- The reproducing kernel $k(., .)$ of the Hilbert space $\mathcal{H}$ is a kernel.
  - Above statement - definition or assertion?

RKHS - Definition I

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. $\mathcal{H}$ is defined to an **Reproducing kernel Hilbert Space (RKHS)** with $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ as the reproducing kernel, if the following conditions are satisfied

- $\forall x \in \mathcal{X}, k(., x) \in \mathcal{H}$ i.e., the space $\mathcal{H}$ contains all functions of the form $k(., x)$ for every element x in the input space $\mathcal{X}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, the following property holds :$f(x) = \langle f, k(., x)\rangle_{\mathcal{H}}$ (it is called the reproducing property of the kernel).

- The reproducing kernel $k(., .)$ of the Hilbert space $\mathcal{H}$ is a kernel.
  - Above statement - definition or assertion?

Proof.

$k(x, x') = \langle k(., x), k(., x')\rangle_{\mathcal{H}}$, from the reproducing property $\square$

$\phi(x) = k(., x)$ is called the canonical feature map.

Conclusion

- Review of the 1st lecture - Kernels - definition and various properties

Conclusion

- Review of the 1st lecture - Kernels - definition and various properties
- Positive definiteness of kernel functions and kernel matrix
- Moore-Aronszajn Theorem
- Reproducing Kernel Hilbert Space
  - Two Definitions of RKHS
  - Equivalence between the two definitions

Conclusion

- Review of the 1st lecture - Kernels - definition and various properties
- Positive definiteness of kernel functions and kernel matrix
- Moore-Aronszajn Theorem
- Reproducing Kernel Hilbert Space
    - Two Definitions of RKHS
    - Equivalence between the two definitions

For more details, please refer

- For proof of Moore-Aronsajn Theorem
  - Kernel Methods for Pattern Analysis - Shawe-Taylor and Christianini
- Detailed notes by Arthur Gretton
  - http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf

Books for further study

- Learning with kernels - Schoelkopf and Smola
- Kernel Methods for Pattern Analysis - Shawe-Taylor and Christianini