CS:E4830 Kernel Methods in Machine Learning

Lecture 2 : Reproducing Kernel Hilbert Space

Rohit Babbar

10th March, 2021

Couple of Announcements

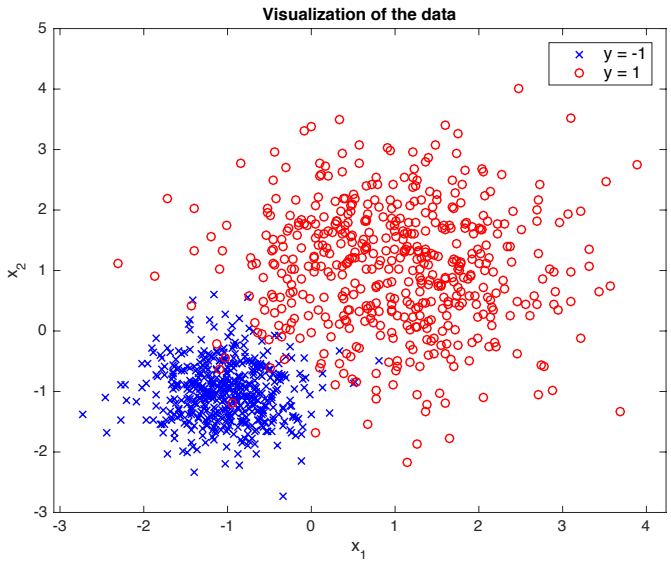
- Python refresher with Petrus Mikkola and Tejas Kulkarni
  - Tomorrow (11th March) at 16:15
  - Zoom session, details to follow on Mycourses
- Assignment 1 will be released early next week

Outline

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 $\langle \phi(x) - c_-, \phi(x) - c_- \rangle$

In the previous lecture...

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

In the previous lecture...

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

In the previous lecture...

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 $\phi(.)$ and the feature space $\mathcal{H}$

In the previous lecture...

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 $\phi(.)$ and the feature space $\mathcal{H}$
2. Checking if the function can be constructed by using elementary operation such as multiplication by a positive scalar, conic sum and product of kernels

In the previous lecture...

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 $\phi(.)$ and the feature space $\mathcal{H}$
2. Checking if the function can be constructed by using elementary operation such as multiplication by a positive scalar, conic sum and product of kernels
   - Similar to proving convexity of a function

Next - **Positive-definiteness**

Positive Definiteness

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 necessarily a kernel (meaning the feature space & feature map definition from Lecture 1)

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 necessarily a kernel (meaning the feature space & feature map definition from Lecture 1)

However, we will prove the following :

- All kernels are positive definite functions (meaning functions that satisfy the feature space/map definition are indeed positive definite)

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 necessarily a kernel (meaning the feature space & feature map definition from Lecture 1)

However, we will prove the following :

- All kernels are positive definite functions (meaning functions that satisfy the feature space/map definition are indeed positive definite)
- Conversely, all positive definite functions are kernels (meaning that given a symmetric function which is positive definite, there exists a feature map and Hilbert space such that ...)

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 forward 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 forward direction

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

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

$$= \langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \rangle_{\mathcal{H}} \text{ (linearity of dot product)}$$

$$= || \sum_{i=1}^{n} a_i \phi(x_i) ||^2_{\mathcal{H}} \geq 0$$

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 forward direction

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

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

$$= \langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \rangle_{\mathcal{H}} \text{ (linearity of dot product)}$$

$$= || \sum_{i=1}^{n} a_i \phi(x_i) ||^2_{\mathcal{H}} \geq 0$$

Proof in the backward direction at the end of the lecture

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

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 :

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

- 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)

Kernel Matrix is positive semi-definite

The kernel matrix $K_{n\times n}$ 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

Kernel Matrix is positive semi-definite

The kernel matrix $K_{n\times n}$ 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 (Same steps in proving positive definiteness of kernel function)

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

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

Reproducing Kernel Hilbert Space

RKHS - Definition I

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathbb{R}^{\mathcal{X}})$ is defined to be a **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

RKHS - Definition I

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathbb{R}^{\mathcal{X}})$ is defined to be a **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

1. For every element x in the input space $\mathcal{X}$ the space $\mathcal{H}$ contains all functions of the form $k(., x)$, i.e.,

$$\forall x \in \mathcal{X}, k(x, .) \in \mathcal{H}$$

RKHS - Definition I

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathbb{R}^{\mathcal{X}})$ is defined to be a **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

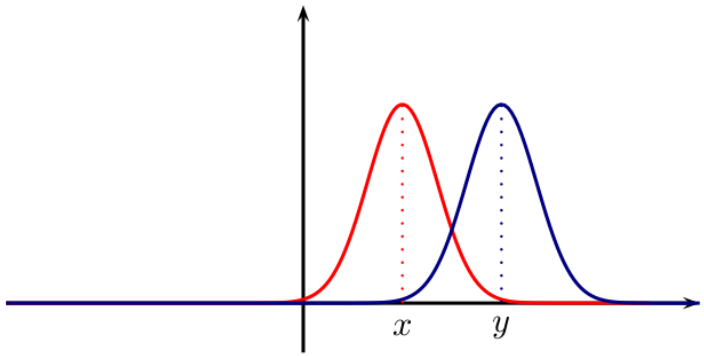
1. For every element x in the input space $\mathcal{X}$ the space $\mathcal{H}$ contains all functions of the form $k(., x)$, i.e.,
$$\forall x \in \mathcal{X}, k(x, .) \in \mathcal{H}$$

2. For every $x \in \mathcal{X}$ and for every $f \in \mathcal{H}$, the following property holds
$$f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}} \text{ called the reproducing property}$$

What is $k(x, .)$

- **it is a (non-linear) function in the input space $\mathcal{X}$** which takes an element in $\mathcal{X}$, and gives a real number $\mathbb{R}$.
- For example -

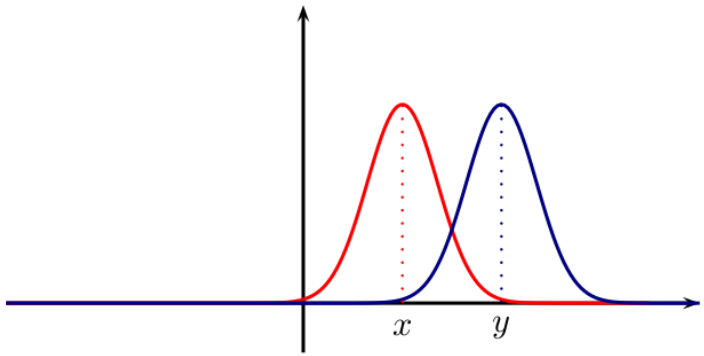
$$k(x, .) : t \mapsto \exp^{-\frac{1}{2\sigma^2}(x-t)^2}$$



What is $k(x, .)$

- **it is a (non-linear) function in the input space $\mathcal{X}$** which takes an element in $\mathcal{X}$, and gives a real number $\mathbb{R}$.
- For example -

$$k(x, .) : t \mapsto \exp^{-\frac{1}{2\sigma^2}(x-t)^2}$$



- Recall the non-linear classification setup in the 1st lecture

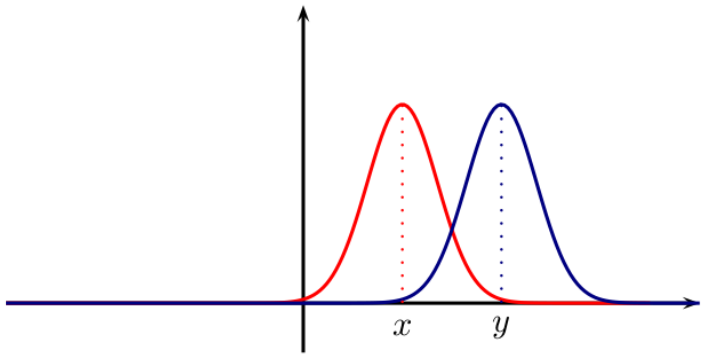
$$f(x) = w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + w^{(3)}x^{(3)} + w^{(4)}x^{(1)}x^{(2)} + w^{(5)}x^{(2)}x^{(3)} + w^{(6)}x^{(1)}x^{(3)}$$

such that $f(x) = \langle f, \phi(x) \rangle_{\mathbb{R}^6}$.

What is $k(x, .)$

- **it is a (non-linear) function in the input space $\mathcal{X}$** which takes an element in $\mathcal{X}$, and gives a real number $\mathbb{R}$.
- For example -

$$k(x, .) : t \mapsto \exp^{-\frac{1}{2\sigma^2}(x-t)^2}$$



- Recall the non-linear classification setup in the 1st lecture

$$f(x) = w^{(1)}x^{(1)} + w^{(2)}x^{(2)} + w^{(3)}x^{(3)} + w^{(4)}x^{(1)}x^{(2)} + w^{(5)}x^{(2)}x^{(3)} + w^{(6)}x^{(1)}x^{(3)}$$

  such that $f(x) = \langle f, \phi(x) \rangle_{\mathbb{R}^6}$.
- In view of the reproducing property of the function $k(., .)$,

$$f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}}$$

  $f(x)$ can be written as an inner product of f (the first argument above) and $k(x, .)$

Reproducing kernel is a kernel

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

Proof.

recall the reproducing property of the kernel (from two slides ago) :

$$f(x) = \langle f, k(x,.)\rangle_{\mathcal{H}}$$

Consider the special case in the definition of RKHS, where we replace the function $f(.)$ - with the function $k(.,.)$ itself. We get the following

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

□

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

RKHS - Definition II

Another definition of RKHS without mention of a kernel

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathcal{R}^{\mathcal{X}})$ is defined to be an RKHS if and only if, for any element $x \in \mathcal{X}$, the following function F_x, which takes a function f from the Hilbert Space $\mathcal{H}$, and maps it to its value $f(x) \in \mathbb{R}$

$$\begin{aligned} F_x : \quad \mathcal{H} &\mapsto \quad \mathbb{R} \\ f &\mapsto \quad f(x) \end{aligned}$$

is continuous

Such a function - which takes as input a function and returns a real number is called a functional. In the above case, where F_x above takes as input a function f and gives its value at $x \in \mathcal{X}$ (i.e. $f(x)$) is called the evaluational functional (represented by δ_x)

RKHS - Definition II

Another definition of RKHS without mention of a kernel

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of real-valued **functions** on the input $\mathcal{X}$. Then $\mathcal{H}(\subset \mathcal{R}^{\mathcal{X}})$ is defined to be an RKHS if and only if, for any element $x \in \mathcal{X}$, the following function F_x, which takes a function f from the Hilbert Space $\mathcal{H}$, and maps it to its value $f(x) \in \mathbb{R}$

$$\begin{aligned} F_x : \quad \mathcal{H} &\mapsto \quad \mathbb{R} \\ f &\mapsto \quad f(x) \end{aligned}$$

is continuous

Such a function - which takes as input a function and returns a real number is called a functional. In the above case, where F_x above takes as input a function f and gives its value at $x \in \mathcal{X}$ (i.e. $f(x)$) is called the evaluational functional (represented by δ_x)

- Example of functional ? Loss functions from Supervised Machine Learning course

RKHS - Equivalence of Two Definitions (Forward direction)

Reproducing property $\implies$ Evaluation functionals are continuous

Proof.

$$
\begin{aligned}
|f(x)| &= |\langle f, k(.,x)\rangle_{\mathcal{H}}| && \text{(reproducing property applied to } f\text{)} \\
&\leq ||k(.,x)||_{\mathcal{H}}||f||_{\mathcal{H}} && \text{(by Cauchy-Schwarz inequality)} \\
&= \sqrt{k(x,x)} \times ||f||_{\mathcal{H}} && \text{(because } ||k(x,.)||^2_{\mathcal{H}} = \langle k(.,x), k(.,x)\rangle = k(x,x)\text{)}
\end{aligned}
$$

□

Therefore, since $||f|| \to 0 \implies f(x) \to 0$. Being a linear Functional, it is sufficient to show the continuity at 0 for the mapping $f \in \mathcal{H} \mapsto f(x) \in \mathbb{R}$, δ_x is continuous.

RKHS - Equivalence of Two Definitions (Backward direction)

Continuous evaluation functionals $\implies$ Reproducing property

Conversely, assuming that for any $x \in \mathcal{X}$, the (linear) evaluation functional $f \in \mathcal{H} \mapsto f(x)$ is continuous. We need to show that the reproducing property

- A result from functional analysis on the property of Hilbert spaces known as Reisz representation theorem states for continuous linear functionals can be written in the form of inner products

Riesz representation Theorem

If $\mathcal{T}$ is a continuous linear functional on a Hilbert space $\mathcal{H}$, then there exists some $g \in \mathcal{H}$ such that for every $f \in \mathcal{H}$, we have

$$\mathcal{T}(f) = \langle f, g \rangle_{\mathcal{H}}$$

In our case, $\mathcal{T}(f)$ is $\delta_x(f) = f(x)$. Therefore, we have $f(x) = \langle f, g_x \rangle_{\mathcal{H}}$. This means that the function given by $g(x, .)$ is the reproducing kernel of $\mathcal{H}$

Linear SVM

Is this at all related to something we might have seen already ?

Soft-margin SVM as a regularised learning problem

- We can rewrite the soft-margin SVM problem

$$\begin{aligned} \textit{Minimize} \quad & \tfrac{1}{2}||\mathbf{w}||^2 + \tfrac{C}{m}\sum_{i=1}^{m}\xi_i \\ \text{Subject to} \quad & \xi_i \geq \max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0) \\ & \text{for all } i = 1, \ldots, N. \\ & \xi_i \geq 0 \end{aligned}$$

  equivalently in terms of Hinge loss as

$$\min_{\mathbf{w}} \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_{Hinge}(\mathbf{w}^T\mathbf{x}_i, y_i) + \frac{\lambda}{2}||\mathbf{w}||^2$$

- This is a so called **regularized learning problem**
  - First term minimizes a loss function on training data
  - Second term, called the **regularizer**, controls the complexity of the model
  - The parameter $\lambda = \frac{1}{C}$ controls the balance between the two terms

Figure: Linear SVM from Machine Learning Supervised Methods Lecture 5 (2020)

RKHS norm controls smoothness of Functions

From basic machine learning, we know that smooth functions obtained via regularization are preferred over spiky and fast changing functions.

RKHS norm and smoothness

$$
\begin{aligned}
|f(x) - f(x')| &= |\langle f, k(x,.)\rangle - \langle f, k(x',.)\rangle| \quad \text{(reproducing property applied to } f\text{)} \\
&= |\langle f, k(x,.) - k(x',.)\rangle| \quad \text{(linearity of dot product)} \\
&\leq ||k(.,x) - k(.,x')||_{\mathcal{H}} ||f||_{\mathcal{H}} \quad \text{(by Cauchy-Schwarz inequality)}
\end{aligned}
$$

- RKHS norm $||f||_{\mathcal{H}}$ controls how much the values at two points x and x' differ compared to their distance $k(.,x) - k(.,x')$
- Larger value of $||f||_{\mathcal{H}}$ allows higher variations (potentially non-smooth functions)

Smaller RKHS norm $\implies$ Smooth functions

Why RKHS are nice function spaces?

Two functions which are close in the RKHS norm, are close point-wise :

- This follows from

$$|f(x) - g(x)| \leq \sqrt{k(x,x)} \times ||f - g||_{\mathcal{H}}, \forall f, g \in \mathcal{H}$$

- As an example, for the linear kernel that we saw in Linear SVM case, $||f - g||_{\mathcal{H}} = ||w_1 - w_2||_2$ for two linear decision boundaries w_1 and w_2 respectively.
- Since $\sqrt{k(x,x)}$ is bounded, if $||f - g||_{\mathcal{H}} = 0$, then $f(x) = g(x), \forall x \in \mathcal{X}$
- Furthermore, when we are allow our search (via learning algorithm) to find functions over arbitrary function classes (which may not be RKHS), then closeness in norm **does not** imply identical pointwise evaluations.

What if we are not in RKHS?

- If we are not in RKHS (meaning we are not using the RKHS norm to measure similarity or dis-similarity), then norm convergence does not imply pointwise convergence.

- Let $\mathcal{F} = L_2([0,1])$, i.e. it represents class of functions for which $\int_0^1 |f(x)|^2 dx < \infty$, which is not an RKHS
- Suppose, we measure distance between functions as :

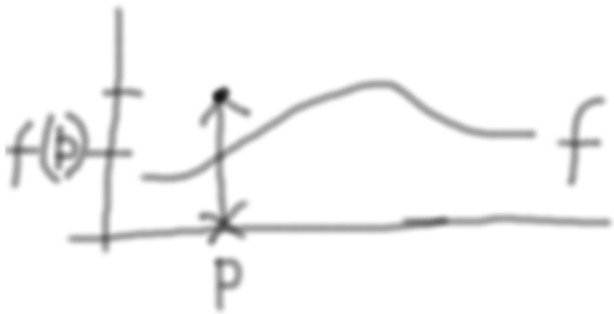


Figure: f is plotted above, g is same as f on all points except at p

$$||f-g||_{L_2([0,1])} = \left(\int_0^1 |f(x) - g(x)|^2 dx \right)^{1/2}$$

- Under this measure of distance between two functions, a function which is zero for all inputs, and one which is non-zero at finitely many points has distance 0.

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- Typically, in a machine learning setup, we do not have access to the true underlying data distribution, and instead we have access to a fixed training set $(x_i, y_i)_{i=1}^n$
- Assume that the data lies in $[0, 1]$, i.e., $x \in \mathcal{X} = [0, 1]$
  - Input x is chosen uniformly at random on $\mathcal{X}$,
  - the label y is chosen in a deterministic way as follows :

$$y = \begin{cases} -1 & \text{if } x < 0.5 \\ +1 & \text{otherwise} \end{cases}$$

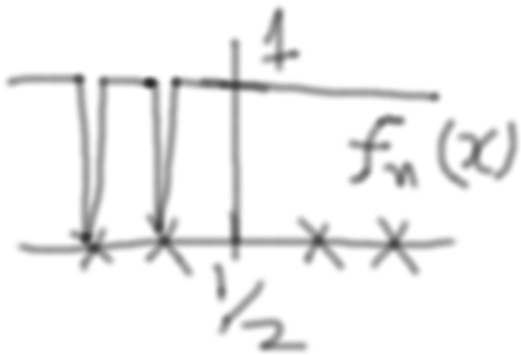
- Consider, a potential classifier based on n training samples given as follows :

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1 \dots n \\ +1 & \text{otherwise} \end{cases}$$

Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

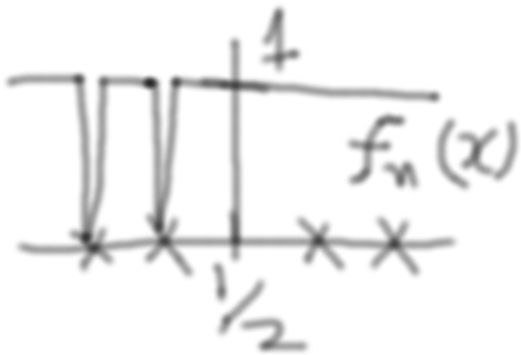
- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?



Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

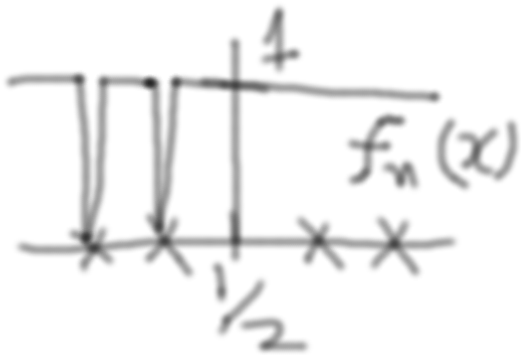
- What is it error on the training set?
  - training error $= 0$ (minimum possible)
  - Has it learnt anything?
- What is its test error?
- Why does overfitting happen?



Relevance of an Appropriate Function class

An Empirical Risk Minimization Example

- What is it error on the training set?
  - training error = 0 (minimum possible)
  - Has it learnt anything?
- What is its test error?



- Why does overfitting happen?
  - Because we allow any function (could be highly non-smooth) in our function space
- In order to generalize, we need to **restrict our function class**
  - Controlling the RKHS norm of the function $||f||_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$ does exactly that
  - Also, called regularization

Uniqueness of reproducing kernel

Theorem

If $\mathcal{H}$ is an RKHS, then its reproducing kernel is unique. Conversely, a symmetric function $k(.,.)$ can be the reproducing kernel of no more than one RKHS.

Proof of the first claim

Let $k(.,.)$ and $k'(.,.)$ be two reproducing kernels for an RKHS $\mathcal{H}$. Then for any $x \in \mathcal{H}$

$$\begin{aligned} ||k(x,.) - k'(x,.)||^2_{\mathcal{H}} &= \langle k(x,.) - k'(x,.), k(x,.) - k'(x,.) \rangle_{\mathcal{H}} \\ &= \langle k(x,.) - k'(x,.), k(x,.) \rangle_{\mathcal{H}} - \langle k(x,.) - k'(x,.), k'(x,.) \rangle_{\mathcal{H}} \\ &= k(x,x) - k'(x,x) - k(x,x) + k'(x,x) \\ &= 0 \end{aligned} \tag{1}$$

Above implies that $k(.,.)$ and $k'(.,.)$ have to be the same kernels.

Construction of RKHS

Moore-Aronszajn Theorem - Backward direction

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

Construction of RKHS

Moore-Aronszajn Theorem - Backward direction

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

Proof outline

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

- **The** feature space will be an RKHS and feature map will be the canonical feature map (Caution : We cannot start with the assumption of Hilbert space, we need to prove it by showing existence or contructing it!)
- **Key aspect** : Our proof will be a proof by construction. In fact, it will be a RKHS, i.e. the unique one in which the reproducing property is satisfied

Construction of RKHS

Moore-Aronszajn Theorem - Backward direction

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

Proof outline

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

- **The** feature space will be an RKHS and feature map will be the canonical feature map (Caution : We cannot start with the assumption of Hilbert space, we need to prove it by showing existence or contructing it!)
- **Key aspect** : Our proof will be a proof by construction. In fact, it will be a RKHS, i.e. the unique one in which the reproducing property is satisfied
- 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 $g(.)$ look like in the input space?

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

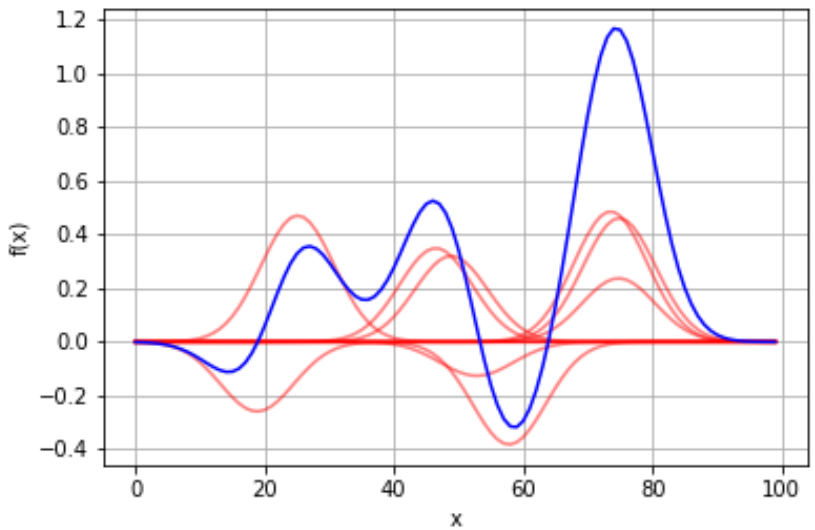


Figure: Pictorial depiction of a function $f(.)$ (in blue) as a linear combination of the positive definite function $k(.,.)$ evaluated at x_is for Gaussian function evaluated at 9 points

Defining the Feature space of functions

Proof (1/3) - Defining the function Space

- A candidate for the function space is given by the set of function given by $\mathcal{H}$

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

Defining Inner product on the space

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

Let $f, g \in \mathcal{H}$ be given by $f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .)$ and $g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$

- The (vector) space of function $\mathcal{H}$ satisfies the requirements of closure under scalar multiplication and addition
  - For $f \in \mathcal{H}, \gamma \in \mathbb{R} \implies \gamma f \in \mathcal{H}$
  - $f, g \in \mathcal{H} \implies (f + g) \in \mathcal{H}$

Defining Inner product on the space

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

Let $f, g \in \mathcal{H}$ be given by $f(.) = \sum_{i=1}^{n} \alpha_i k(x_i, .)$ and $g(.) = \sum_{j=1}^{m} \beta_j k(y_j, .)$

- The (vector) space of function $\mathcal{H}$ 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 :
  - The inner product between f and g is defined by the following

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{m} \beta_j f(y_j) \tag{2}$$

From the 2nd and 3rd equality, the inner product is well-defined i.e., it does not depend on particular expansion of $f(.)$ and $g(.)$. Also, it shows that the inner product satisfies the symmetry and linearity properties.

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

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

- The **reproducing property** follows from equation (2) is obtained by taking $g = k(x, .)$,

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

- By Cauchy-Schwarz inequality, $\forall x \in \mathcal{X}$

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

Therefore, $||f||_{\mathcal{H}} = 0 \implies f = 0$

The above construction (almost) gives us the desired Hilbert space with reproducing property, i.e. an RKHS

Conclusion

- Parzen Window Classifier
- Positive definiteness of kernel functions and kernel matrix
- Moore-Aronszajn Theorem
- Reproducing Kernel Hilbert Space
  - Two Definitions of RKHS
  - Equivalence between the two definitions

References

For more details, please refer

- For proof of Moore-Aronsajn Theorem
  - Kernel Methods for Pattern Analysis (Chapter 3) - Shawe-Taylor and Christianini
  - Learning with Kernels (Chapter 2) - Schoelkopf and Smola
- Detailed notes by Arthur Gretton
  - `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/RKHS_Notes1.pdf`

References

Books for further study

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