Kernel Methods in Machine Learning

Dr. Sandeep Kumar
ksandeep@iitd.ac.in

Department of Electrical Engineering
Indian Institute of Technology, Delhi

March 21, 2022

Motivation for Kernel Methods

For a learning problem with domain set χ, label set $\mathcal{Y}$ and training data

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\} \in \chi \times \mathcal{Y}$$

A hypothesis function h_S needs to be estimated based on the training dataset which generalizes well on the test data.

$$h_S : \chi \rightarrow \mathcal{Y}$$

What is meant by generalization?
Given any new sample x from the domain set the hypothesis function should predict $y \in \mathcal{Y}$ correctly.
In simple words by generalization we mean that the ordered pair (x, y) should have some sense of **SIMILARITY** with the elements of the training set S.

Kernels as Similarity Function

- Obtaining similarity between the samples of the label set $\mathcal{Y}$ is trivial. However, it is not so obvious for the samples of the domain set χ.

Similarity in the domain set

Let $k : \chi \times \chi \to \mathbb{R}$ be a similarity function that takes two elements from the domain set and the corresponding image depicts a sense of relationship between both the elements.

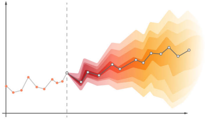
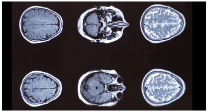
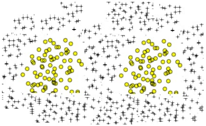
Example

Inner product function provides a similarity measure between two same dimensional vectors. Inner product can induce norms which provides a mathematical sense of distance between two vectors.

Can inner product be an obvious choice as kernel function to compare elements of the domain set?

Real World Data

Real world data have a domain set that may not be in a space where inner product is defined.

Kernel method theory extends the concept of linear learning machines for a far more complex and non-linearly separable datasets.

Kernel Functions

Let χ be any arbitrary non empty set of features. A function $k : \chi \times \chi \to \mathbb{R}$ is a kernel if there exists an hilbert space $\mathcal{H}$ and a mapping ϕ defined as

$$\phi : \chi \to \mathcal{H} \quad s.t. \forall x_1, x_2 \in \chi$$

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$$

Gram Matrix

Given a kernel function k, and elements $x_1, x_2, ..., x_m \in \chi$, the $m \times m$ matrix K such that

$$K_{ij} = k(x_i, x_j) = K_{ji}$$

The matrix K is called as gram matrix of the kernel function k w.r.t the m elements of the domain set χ.

Polynomial Kernel Function

- Let the domain set $\chi = \mathbb{R}^2$, Is the function k defined as

$$k(x, \tilde{x}) = \langle x, \tilde{x} \rangle_{\mathbb{R}^2}^2 = (x_1 \tilde{x}_1 + x_2 \tilde{x}_2)^2$$

  a valid kernel function?

If we choose a mapping function ϕ such that

$$\phi(x) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2) \in \mathbb{R}^3$$

The corresponding hilbert space is $\mathbb{R}^3$.

$$\begin{aligned} \langle \phi(x), \phi(\tilde{x}) \rangle_{\mathbb{R}^3} &= x_1^2 \tilde{x}_1^2 + 2 x_1 x_2 \tilde{x}_1 \tilde{x}_2 + x_2^2 \tilde{x}_2^2 \\ &= (x_1 \tilde{x}_1 + x_2 \tilde{x}_2)^2 \\ &= k(x, \tilde{x}) \end{aligned}$$

Hence, k is a valid kernel function.

Positive Definite Kernels

A kernel function k, which satisfies the following property is known as positive definite kernel functions.

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

Where, $\alpha_t \in \mathbb{R}$.

Linear Algebra Definition

If the gram matrix K corresponding to the kernel function k w.r.t the elements of set χ is a positive semidefinite matrix, then the kernel function is positive definite.

Algorithms which take input as the gram matrix are known as kernel methods.

Hilbert Space

- A vector space endowed with inner product is known as inner product spaces.
- Inner product spaces have induced norms associated which is defined as
$$\|x\| = \langle x, x \rangle^{\frac{1}{2}}$$
- A sequence in a metric space is termed as cauchy sequence if there exists an N for all $\zeta > 0$ such that
$$d(x_n, x_m) < \zeta \quad \forall m, n \geq N$$
- A space in which all cauchy sequences are convergent is known as a complete space.
- A complete inner product space is known as hilbert space.

Why Completeness is good?

Many convergence results from the euclidean space can be directly extended to infinite (arbitrarily large) dimensional spaces.

Functional Spaces for Kernel Methods

- For learning with kernels, the hilbert space of functions that maps the elements of the domain set χ to $\mathbb{R}$ are of practical interest.
- The motivation behind working on this specific function space is that the hypothesis function also lies in that space.
- From the definition of kernel function we know that there exists a map ϕ such that

$$\phi_k : \chi \to \mathcal{H}$$

- Since, $\mathcal{H}$ is a function space which maps each element of the domain set to a real number, each of the points in the domain set is represented by its similarity to all other points of the set.

How to Map Elements by Functions

Given k is a positive definite kernel each of the element $x_i \in \chi$ is represented as

$$x_i \in \chi \rightarrow \phi_k(x_i) := k_{x_i} := k(x_i, \cdot)$$

The kernel function is a bivariate function while the function $k(x_i, \cdot)$ is a univariate function or partial evaluation of the kernel function.

Will $k(x_i, \cdot)$ be in the set of functions mapping elements from $\chi \rightarrow \mathbb{R}$?

Yes

$$k(x_i, \cdot) : \chi \rightarrow \mathbb{R}$$

Obtaining a Feature Space of Linear Functionals

- Previously it is shown that the partial evaluation of the kernel function lies in the set of functions mapping elements from $\chi \rightarrow \mathbb{R}$.
- However, it is still unclear that all such pointwise evaluations of the kernel function lead to a hilbert space or not.

Steps to construct a hilbert space of functions

1. Turn the image of ϕ_k into a vector space.
2. Define a inner product corresponding to that space.
3. Check whether the inner product satisfies the kernel definition

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$$

Create a Vector Space

Let k be a positive definite kernel and χ be a non-empty domain set. Let $\mathbb{R}^{\chi}$ be a set of linear functionals defined over set χ.

$$\mathbb{R}^{\chi} = f : \chi \to \mathbb{R}$$

Now a mapping function

$$\phi_k : \chi \to \mathbb{R}^{\chi}$$

maps each element of χ to a linear functional.
Let G represent a vector space spanned by each of the linear functionals

$$\begin{aligned} G &= span \left\{ \phi_k(x_i) \right\}_{i=1}^{m} \\ &= \sum_{i=1}^{m} \alpha_i \phi_k(x_i) \end{aligned}$$

This completes the first step.

Define an Inner Product

Take two functions from the vector space G

$$f(\cdot) = \sum_{i=1}^{m_1} \beta_i k(x_i, \cdot)$$

$$g(\cdot) = \sum_{j=1}^{m_2^{'}} \gamma_j k(x_j^{'}, \cdot)$$

$$\langle f, g \rangle_G = \sum_{i=1}^{m_1} \beta_i k(x_i, \cdot) \sum_{j=1}^{m_2^{'}} \gamma_j k(x_j^{'}, \cdot)$$

$$= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2^{'}} \beta_i \gamma_j k(x_j^{'}, x_i)$$

By observation the inner product satisfies all the properties.
Hence, this completes the second step.

Kernel Property

By definition of the inner product defined above

$$\left\langle \phi_k(x), \phi_k(x') \right\rangle = k(x, \cdot)k(x', \cdot) = k(x, x')$$

Reproducing Property of Kernels

$$\begin{aligned}
\langle k(x, \cdot), f \rangle &= \left\langle k(x, \cdot), \sum_i \alpha_i k(x_i, \cdot) \right\rangle \\
&= \sum_i \alpha_i k(x_i, \cdot) k(x, \cdot) \\
&= \sum_i \alpha_i k(x_i, x) \\
&= f(x)
\end{aligned}$$

The linear form in hilbert space may correspond to non-linear model in χ.

Reproducing Kernel Hilbert Space

Let χ be a non-empty set and $\mathcal{H}$ be a hilbert space of linear functionals over χ. Then $\mathcal{H}$ is called an RKHS if there exists a kernel $k : \chi \times \chi \to \mathbb{R}$ such that

1. k has reproducing property i.e.,

$$f(x) = \langle k(x, \cdot), f \rangle$$

2. k spans the hilbert space $\mathcal{H}$.

Kernel Trick

If an algorithm takes only pairwise inner product of the elements of the domain set as input, the same algorithm can be potentially applied to non-vectorial data or infinite dimensional data as well by replacing the inner product with kernel evaluation.

Example

If ϕ maps elements of the domain set to an hilbert space $\mathcal{H}$, the pairwise distance can be evaluated by using kernel functions.

$$\begin{aligned}
d^2(\phi(x_1), \phi(x_2)) &= \|\phi(x_1) - \phi(x_2)\| \\
&= \langle \phi(x_1) - \phi(x_2), \phi(x_1) - \phi(x_2) \rangle \\
&= \langle \phi(x_1), \phi(x_2) \rangle - \langle \phi(x_1), \phi(x_2) \rangle \\
&- \langle \phi(x_2), \phi(x_1) \rangle + \langle \phi(x_2), \phi(x_2) \rangle \\
&= k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2)
\end{aligned}$$

Representer's Theorem (Optimization in RKHS)

Let k be a positive definite kernel with $\mathcal{H}$ being the associated RKHS, χ be the domain set with elements $x_1, x_2, ..., x_m$. If $\mathcal{L}$ is any arbitrary loss function, then minimizer of the regularized risk with strictly monotonically increasing regularization function

$$\mathcal{L}\left\{(x_i, y_i, f(x_i))\right\}_{i=1}^{m} + \omega(\|f\|)$$

can be represented as

$$f^*(x) = \sum_{i=1}^{m} \alpha_i^* k(x_i, x)$$

Maximum Margin Classifier

For a maximum margin classifier the optimization problem is expressed as

$$\min_{w,b,\zeta} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\zeta_i$$

$$\text{subject to,}$$

$$y_i(w^T x_i + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0$$

$$\forall i = 1, 2, 3, ..., m$$

The above constrained problem can be re expressed as unconstrained using hinge loss expression

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} max\left(0, 1 - [y_i(w^T x_i) + b]\right)$$

Kernel Maximum Margin Classifier (1/2)

Let ϕ be a mapping from the domain set χ to RKHS $\mathcal{H}$, then the primal form of optimization will be

$$\min_{w,b,\zeta} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\zeta_i$$

subject to,

$$y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0$$

$$\forall i = 1, 2, 3, ..., m$$

The lagrangian of the above problem will be

$$\mathcal{L}(w, b, \zeta, \lambda, \nu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\zeta_i$$

$$- \sum_{i=1}^{m}\lambda_i[y_i\left(w^T\phi(x_i) + b\right) - 1 + \zeta_i] - \sum_{i=1}^{l}\nu_i\zeta_i$$

Kernel Maximum Margin Classifier (2/2)

The dual of the previous problem will be

$$\max_{\lambda \geq 0, \nu \geq 0} \quad -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \lambda_i \lambda_j \left(\phi(x_i)^T \phi(x_j) \right) + \sum_{j=1}^{m} \lambda_j$$

subject to,

$$\sum_{i=1}^{m} \lambda_i y_i$$

$$C - \lambda_i - \nu_i = 0$$

$$\forall i = 1, 2, 3, ..., m$$

By using the kernel trick inner product term can be equivalently expressed as kernel evaluation.

$$\phi(x_i)^T \phi(x_j) = k(x_i, x_j)$$