Support Vector machine

Kernel Functions: Theory and Construction

Professor Abdellatif El Afia

Reminder

Polynomial transformation or **Feature map**

- $\phi_Q : X \subseteq \mathbb{R}^d \to Z$ such that $\phi_Q(x) = \begin{pmatrix} 1, \\ x_1, \dots, x_d, \\ x_1^2, x_1 x_2, \dots, x_d^2, \\ \dots \\ x_1^Q, x_1^{Q-1} x_2, \dots, x_d^Q \end{pmatrix}$

The optimization Primal problem becomes

- $(\phi_Q(x_i), y_i)$

- $C - SVCNS \begin{cases} Min & \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \\ s.t & y_i(w^T \boldsymbol{\phi_Q(x_i)} + b) \geq 1 - \xi_i, i = 1, .., n \\ & \xi_i \geq 0, w \in \mathbb{R}^d, b \in \mathbb{R} \end{cases}$

Reminder

Dual form:

$$Dual: C - SVCNS \Leftrightarrow \begin{cases} Max \quad \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_j \lambda_i y_j y_i \left(\phi_Q(x_j)^T \phi_Q(x_i) \right) \\ s.t \quad \sum_{i=1}^{n} \lambda_i y_i = 0 \\ \quad C \geq \lambda_i \geq 0 \; i = 1, \dots, n \end{cases}$$

Motivation

Consider a two-dimensional input space $X \subseteq \mathbb{R}^2$together with the Feature map $\phi: X \longrightarrow Z$ such that

$$\phi(x) = \phi(x^1, x^2) = \begin{pmatrix} (x^1)^2 \\ (x^2)^2 \\ \sqrt{2}x^1x^2 \end{pmatrix} \in Z \subseteq \mathbb{R}^3$$

With Z is the feature space

The hypothesis space of hyperplanes in Z would then be

$$h(x) = w_1(x^1)^2 + w_3(x^2)^2 + w_3\sqrt{2}x^1x^2$$

The composition of the feature map with inner product in the feature space can be evaluated as follows

- $\langle\phi(x_i), \phi(x_j)\rangle = (\phi(x_i))^T\phi(x_j) = ((x_i^1)^2, (x_i^2)^2, \sqrt{2}x_i^1x_i^2)((x_j^1)^2, (x_j^2)^2, \sqrt{2}x_j^1x_j^2)^T$
- $\Longrightarrow \langle\phi(x_i), \phi(x_j)\rangle = (x_i^1)^2(x_j^1)^2 + (x_i^2)^2(x_j^2)^2 + 2x_i^1x_i^2x_j^1x_j^2 = (x_i^1x_j^1 + x_i^2x_j^2)^2$
- $\Longrightarrow \boldsymbol{\langle\phi(x_i), \phi(x_j)\rangle = \langle x_i, x_j\rangle^2 = k(x_i, x_j)}$

Then we can compute the inner product between the projections of two points into the feature space without explicitly evaluating their coordinates

Motivation

Note that the same Function $\boldsymbol{k}$ computes the inner product corresponding to the four-dimensional feature map $\phi: X \longrightarrow Z$ such that

$$\phi(x) = \phi(x^1, x^2) = \begin{pmatrix} (x^1)^2 \\ (x^2)^2 \\ x^1 x^2 \\ x^2 x^1 \end{pmatrix} \in Z \subseteq \mathbb{R}^4$$

The composition of the feature map with inner product in the feature space can be evaluated as follows

- $\langle \phi(x_i), \phi(x_j) \rangle = (\phi(x_i))^T \phi(x_j) = ((x_i^1)^2, (x_i^2)^2, x_i^1 x_i^2, x_i^2 x_i^1)((x_j^1)^2, (x_j^2)^2, x_j^1 x_j^2, x_j^2 x_j^1)^T$
- $\Longrightarrow \langle \phi(x_i), \phi(x_j) \rangle = (x_i^1)^2 (x_j^1)^2 + (x_i^2)^2 (x_j^2)^2 + x_i^1 x_i^2 x_j^1 x_j^2 + x_i^2 x_i^1 x_j^2 x_j^1 = (x_i^1 x_j^1 + x_i^2 x_j^2)^2$
- $\Longrightarrow \langle \phi(x_i), \phi(x_j) \rangle = (x_i^1 x_j^1 + x_i^2 x_j^2)^2 = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle^{\mathbf{2}} = \boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j})$

showing that the feature space is not uniquely determined by the function $\boldsymbol{k}$: $\boldsymbol{X} \times \boldsymbol{X} \longrightarrow \mathbb{R}$

Motivation

Consider a n-dimensional input space $X \subseteq \mathbb{R}^n$together with the Feature map $\phi: X \longrightarrow Z$ such that

$$\phi(x) = \phi(x^1, \dots, x^n) = (x^t x^s)_{s,t=1}^n \in Z \subseteq \mathbb{R}^{n^2}$$

The composition of the feature map with inner product in the feature space can be evaluated as follows

- $\langle \phi(x_i), \phi(x_j) \rangle = (\phi(x_i))^T \phi(x_j) = ((x_i^t x_i^s)_{s,t=1}^n)((x_j^t x_j^s)_{s,t=1}^n)^T$
- $\Longrightarrow \langle \phi(x_i), \phi(x_j) \rangle = \sum_{t,s=1}^n x_i^t x_i^s x_j^t x_j^s = \sum_{t=1}^n x_i^t x_j^t \sum_{s=1}^n x_i^s x_j^s$
- $\Longrightarrow \boldsymbol{\langle \phi(x_i), \phi(x_j) \rangle = \langle x_i, x_j \rangle^2 = k(x_i, x_j)}$

The inner products can, however, sometimes be computed more efficiently as a direct function of the input features, with explicitly computing the mapping ϕ. In other words the feature-vector representation step can be by-passed. A function that performs this direct computation is known as

kernel Function

plan

1. **Theory of kernel functions**
   1. Reproducing kernel Hilbert spaces
   2. Caracterizing Kernel Functions
2. **Construction of kernel functions**
   1. Kernel Constructions
   2. Transforming Kernel Matrices

Theory of kernel function

Reproducing Kernel Hilbert Spaces

1. Inner Product Space
2. Hilbert Space
3. Function Spaces
4. Separable Hilbert Spaces

Inner product space

Definition : An inner product space X is a vector space with an associated inner product

$$\begin{cases} h & X \times X & \rightarrow & \mathbb{R} \\ & (x,y) & \rightarrow & h(x,y) \end{cases}$$

that satisfies :

- **Symmetry**: $h(x,y) = h(y,x)$
- **Linearity:**
  - $h(ax,y) = \text{a}h(x,y)$
  - $h(x+z, y+z) = h(x,y) + h(z,y)$
- **Positive Semi-Definiteness**(PSD): $h(x,x) \geq 0$

- **The inner product space is strict if $\boldsymbol{h(x,x) = 0 \Leftrightarrow x = 0}$**
- A strict inner product space X has a natural norm given by $\|x\|_2 = \sqrt{x^T x}$ The associated metric is $h\ (x,z) = \|x-z\|_2$
- The space $\mathbb{R}^n$ has the inner product $h(x,y) = x^T y$ which yields the Euclidean norm:

$$(\|\boldsymbol{x}-\boldsymbol{y}\|_2)^2 = \sum_{i=1}^{n} (\boldsymbol{x_i} + \boldsymbol{y_i})^2$$

Hilbert Space

Definition :

A strict inner product space X is a Hilbert space if it is:

- **Complete: Technical Condition required for potentially infinite-dimensional sets**
  Every Cauchy sequence $\{x_i \in X\}_{i=1}^{\infty}$ such that $\lim_{n\to\infty} \sup_{m>n} ||x_n - x_m|| = 0$ converges to an element $x \in X$; i.e., $\lim_{i\to\infty} x_i = x$
- **Separable: Condition required to make Hilbert space isomorphisms**
  There is a countable subset $\hat{X} = \{x_i \in X\}_{i=1}^{\infty}$ such that $\forall x \in X$ and $\varepsilon > 0$, $\exists x_i \in \hat{X}$ such that : $\|x_i - x\| < \varepsilon$

Examples:

- the interval [0, 1], the reals $\mathbb{R}$, the complex numbers C and Euclidean spaces $\mathbb{R}^n$ for $n \in \mathbb{N}$, are the Hilbert space
- The subspace ℓ^2 for which $\forall\, x \quad h(x,x) < \infty$ is a Hilbert space
- The Subspace $L_2\,(X)$ defined on X, a compact subspace of $\mathbb{R}^d$, for which $\forall\, f \in L_2\,(X)$, $h(f\,,f) = \int_x f(x)f(x)dx < \infty$ is a Hilbert space

Separable Hilbert Spaces

- Hilbert space F is isomorphic to H if there is a one-to-one linear mapping $T : F \rightarrow H$ such that for $\forall\, x, y \in F$

$$h_H(T(x), T(y)) = h_F(x, y)$$

- Every separable Hilbert space is isomorphic to :
  - $\mathbb{R}^d$ if it has a dimension d
  - l_2 if it has an infinite dimension

- Since Hilbert space F is isomorphic to $\mathbb{R}^d$ or l_2, F has an orthonormal basis $\{\phi_i\}$ and $\forall x \in F$ have a Fourier decomposition:

$$\boldsymbol{x} = \sum_{\boldsymbol{i}} \boldsymbol{h_F(\phi_i, x)\phi_i}$$

Theory of kernel functions

Caracterizing Kernel Functions

1. Kernel Termenology
2. Kernel Matrices
3. Reproducing Kernel Function
4. Kernel Functions

Kernel termenology

Definition :

A kernel, k, is a two-argument real-valued function over $X \times X$

$$
\begin{array}{llll}
k: & X \times X & \rightarrow & \mathbb{R} \\
 & (x, y) & \rightarrow & k(x, y) = h_F\big(\phi(x), \phi(y)\big) \quad (1)
\end{array}
$$

for some inner-product space F such that $\phi: X \rightarrow F$ and $\forall\, x \in X \rightarrow \phi(x) \in F$

- Kernel functions must be symmetric since inner products are symmetric
- To show that k is a valid kernel, it is sufficient to show that a mapping ϕ exists that yields (1). However, this is generally difficult to construct.
- In this rest of this chapter, we will demonstrate additional ways to construct and validate kernels

- $\phi: X \rightarrow F$ and $\forall\, x \in X \rightarrow \phi(x) \in F$
- $\phi: X = \mathbb{R}^d \rightarrow F = \mathbb{R}^q$
- $q > d$

Kernel Matrices

Definition:

A kernel matrix (or Gram matrix) K is the matrix that results from applying k to all pairs of trainning set $\{x_i\}_{i=1}^{n}$

$$K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

that is, $k_{i,j} = k(x_i, x_j)$

Kernel matrices are square and symmetric. And $\mathrm{t}r(K) = \sum_{i=1}^{n} k(x_i, x_i)$

Proposition 1:

- Since K is a symmetric $n \times n$ real-valued matrix, it can be written as
  - If $rank(K) = n$ then $K = V\Lambda V^T = \sum_{i=1}^{n} \lambda_i v_i (v_i)^T$
  - Else $(rank(K) = k < n)$, then $K = V\Lambda V^T = \sum_{i=1}^{k} \lambda_i v_i (v_i)^T$

where (λ_i, v_i) are eigen-value/vector pairs of K. This is called the spectral decomposition of K

- $\mathrm{t}r(K_1 K_2) = \mathrm{t}r(K_2 K_1)$

Kernel Matrices

Proposition 2:

Kernel matrices, which are constructed from a kernel corresponding to a strict inner product space F, are PSD.

Proof:

By definition of a kernel matrix, for all $i, j \in \{1, \dots, n\}, k_{i,j} = h_F(\phi(x_i), \phi(x_j))$

Thus, for any $v \in \mathbb{R}^n$:

- $v^T K v = \sum_{i,j}^{n} v_i k_{i,j} v_j = \sum_{i,j}^{n} v_i h_F(\phi(x_i), \phi(x_j)) v_j = h_F\left(\sum_{i=1}^{n} v_i \phi(x_i), \sum_{j=1}^{n} v_j \phi\left(x_j\right)\right)$
- $\Longrightarrow v^T K v = \|\sum_{i=1}^{n} v_i \phi(x_i)\|_F^2|| \geq 0$

Proposition 3:

- Matrix K is PSD iff there exists a real matrix B such that $K = BB^T = V\sqrt{\Lambda}\sqrt{\Lambda}V^T$

Reproducing Kernel Function

Definition (Aronszajn, 1950)

Suppose F is a **Hilbert space** of functions over X; the function $k\colon X \times X \to \mathbb{R}$ is a reproducing kernel of F if

1. $\forall x \in X$, the function $f_x(\cdot) = k(\cdot, x) \in F$.
2. **Reproducing Property**: $\forall y \in X, \forall f \in F\colon f(y) = h_F(f, k(.,y))$

Further, the space is called a Reproducing Kernel Hilbert Space (RKHS)

Remarks:

- By 1st property and closure of F, $\forall \alpha_i \in \mathbb{R}, \forall x_i \in X$ we have

$$\sum_{i=1}^{n} \alpha_i k(., x_i) \in \hat{X} = \{x_i \in X\}_{i=1}^{\infty}$$

- Applying f_x from 1st property to 2nd property, $\forall (x, y) \in X^2$, we have

$$k(x, y) = h_F(k(., x), k(., y))$$

Kernel functions

Definition (Finitely Positive Semi-definite)

A function $k: X \times X \rightarrow \mathbb{R}$ is finitely positive semi-definite (FPSD) if

- It is symmetric: i.e., $\forall x, z \in X^2 \quad k(x, z) = k(z, x) < \infty$

- The matrix K formed by applying k to any finite subset of X is positive semi-definite: $v^T K v \geq 0$

Theorem :

$k: X \times X \rightarrow \mathbb{R}$ (either continuous or with a countable domain) is FPSD iff $\exists$ Hilbert space F with feature map $\phi: X \longrightarrow F$ such that:

$$k(x, z) = h_F(\phi(x), \phi(z))$$

Kernel functions

Proof

- Case $\Leftarrow$: Follows from Proposition 2.
- Case $\Rightarrow$: Suppose k if FPSD and we construct Hilbert Space F_k with k as its reproducing kernel; i.e., F_k is the closure of functions: $f_x(.) = k(.,x)$

Thus, $\forall \alpha_i \in \mathbb{R}, \forall x_i \in X,\ \ g(.) = \sum_i \alpha_i\, k(.,x_i) \in F_k$ and by the reproducing property,

$$h_F(g,g) = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha$$

where K is the kernel matrix $\{x_i\}_{i=1}^{n}$, and thus $\alpha^T K \alpha \geq 0$ since K is PSD.

Kernel functions

- **(Completeness)** Follows from the Cauchy-Schwarz inequality, ?
- **(Separability)** Separability follows from k being continuous or having a countable domain ?.

Finally, the mapping ϕ is specified by k and $\phi(x) = k(.,x) \in F_k$

Note:

the Inner Product defined above is strict since:

$$\text{if } \|f\| = 0 \text{ then } \forall x \in X, |f(x)| \leq \|f\| \|\phi(x)\| = 0$$

Constructions of kernel function

Kernel Constructions

1. Simple Kernels
2. Closure Properties of Kernels
3. Additional Kernel Functions
4. Kernel Questions

Simple Kernels

Clearly, the linear kernel defined by

$$K_{lin}\,(x,z) = h_F(x,z) = x^T z$$

is a valid kernel function since it is an inner product in X

For any $n \times n$ matrix $B \geq 0$,

$$\mathrm{k_B(x,z)} = h_F(x, Bz) = x^T Bz$$

is a valid kernel function

Closure Properties of Kernels

Proposition 3

Suppose:

- k_1 and k_2 are kernels on X,
- $a > 0$,
- $f : X \to \mathbb{R}$,
- $\varphi : X \to \mathbb{R}^n$,
- k_3 is a kernel on $\mathbb{R}^n$.

Then these are all kernel functions on X:

1. $k(x,z) = k_1(x,z) + k_2(x,z)$
2. $k(x,z) = a \cdot k_1(x,z)$
3. $k(x,z) = k_1(x,z) \cdot k_2(x,z)$
4. $k(x,z) = f(x)f(z)$
5. $k(x,z) = k_3(\varphi(x), \varphi(z))$

Closure Properties of Kernels

Proof

Let K_1 and K_2 be the kernel matrices of k_1 and k_2 applied to any set $\{x_i\}_{i=1}^{n}$ both these matrices are PSD. Also let ϑ be any n-vector:

- $\mathrm{K} = K_1 + K_2 \Longrightarrow \vartheta^T K\vartheta = \vartheta^T K_1\vartheta + \vartheta^T K_2\vartheta \geq 0$
- $\mathrm{K} = aK_1 \ \Longrightarrow \vartheta^T K\vartheta = a\vartheta^T K_1\vartheta \geq 0$
- Since $K_1 = BB^T,\ K_2 = CC^T \Longrightarrow \mathrm{K} = BB^T CC^T \Longrightarrow\ \vartheta^T K\vartheta = tr(D_\vartheta BB^T D_\vartheta CC^T) = tr(C^T D_\vartheta BB^T D_\vartheta C) = tr((C^T D_\vartheta B)^T C^T D_\vartheta B)$

- $k(x,z) = h\big(\varphi\,(x), \varphi\,(z)\big)$ where $\varphi :\ X\ \rightarrow\ \mathbb{R}^n$ thus, k is PSD.
- Since k_3 is a kernel, applying it to any set of vectors $\{\varphi(x_i)\}_{i=1}^{N}$yields a PSD matrix.

Closure Properties of Kernels

The feature spaces for these kernels are as follows:

- For kernel $k_1(x,z) + k_2(x,z)$, the new feature map is equivalent to stacking the feature maps of k_1 and k_2:

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix}$$

- For kernel $a \cdot k_1(x,z)$, its feature space is scaled by $\sqrt{a}$
- For kernel $k_1(x,z) \times k_2(x,z)$, if ϕ_1 has dimension n_1 and ϕ_2 has dimension n_2 , ϕ has $n_1 n_2$ features given by

$$(\phi(x))_{ij} = (\phi_1(x))_i (\phi_2(x))_j$$

- It follows that the features of $k_1(x,z)^d$ are all monomials of the form

$$(\phi_1(x))_1^{d_1} (\phi_1(x))_2^{d_2} \dots . (\phi_1(x))_n^{d_n}, \qquad \sum_{i=1}^{n} d_i = 1$$

- $a \cdot k_1(x, z) = k_1(\sqrt{a}x, \sqrt{a}z)$

Additional Kernel Functions

Proposition

Suppose k_1 is a kernel on X and $p : \mathbb{R} \rightarrow \mathbb{R}$ is a polynomial with non-negative coefficients. Then, the following are kernels:

1. Polynomial Kernel:
   - $k_{poly}(x, z) = p(k_1(x, z))$
   - $k_{poly}(x, z) = (x^T z + R)^d$
2. Gaussian kernel:
   - $k(x, z) = e^{k_1(x,z)}$
   - Radial Basis function (RBF) Kernel: $k_{RBF}(x, z) = e^{-\frac{\|x-z\|_2^2}{2\sigma^2}}$

Proof

1. Constructing a polynomial kernel from base kernel k_1 proceeds directly from Proposition 3 (1, 2, 3)
2. Consider that $exp(x) = 1 + x + \frac{1}{2} x^2 + \cdots + \frac{1}{i!} x^i + \cdots$ Thus, it is a limit of polynomials and the PSD property is closed under pointwise limits.(RBF Kernel) Left as an exercise.

Kernel Questions

Which of the following functions are kernels?

- $k_1(x,z) = \sum_{i=1}^{D}(x_i + z_{i)}$
- $k_2(x,z) = \prod_{i=1}^{D} h\left(\frac{x_i - c}{a}\right) h\left(\frac{z_i - c}{a}\right)$ where $h(x) = \cos(1.75x)\, e^{-\frac{x^2}{2}}$
- $k_3(x,z) = \frac{x^T z}{||x||_2 ||z||_2}$
- $k_4(x,z) = \sqrt{||x - z||_2^2 + 1}$

Constructions of kernel function

Transforming Kernel Matrices

1. Simple Transformations
2. Centering Data
3. Normalizing Data

Simple Transformations

- Adding a non-negative constant to the Kernel Matrix: corresponds to adding a new constant feature to each training example; i.e., given the matrix Φ of features such that $K = \Phi\, \Phi^T$,
$$[\Phi \quad \mathrm{c}1] * [\Phi \quad \mathrm{c}1]^T = K + c^2\, 11^T$$
- Adding a non-negative constant to its diagonal: corresponds to adding an indicator feature for every data point
$$\begin{bmatrix} \phi(x_1) & c & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_n) & 0 & \dots & c \end{bmatrix} \begin{bmatrix} \phi(x_1) & c & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_n) & 0 & \dots & c \end{bmatrix}^T = K + c^2\, I$$

Centering Data

Suppose we want to translate the origin to the data's center of mass, this transformation can be expressed as kernel transform

$$K \leftarrow K - \frac{1}{N}\, 11^T K - \frac{1}{N} K 11^T + \frac{1K1^T}{N^2}\, 11^T$$

Normalizing Data

Suppose we want to project all data to be norm 1; i.e., $\|\hat{x}\| = 1$

This transformation can be achieved using only the information from the kernel matrix:

$$\hat{k}\ (x, z) = \frac{k\ (x, z)}{\sqrt{k\ (x, x) k(z, z)}}$$