Μηχανική Μαθηση

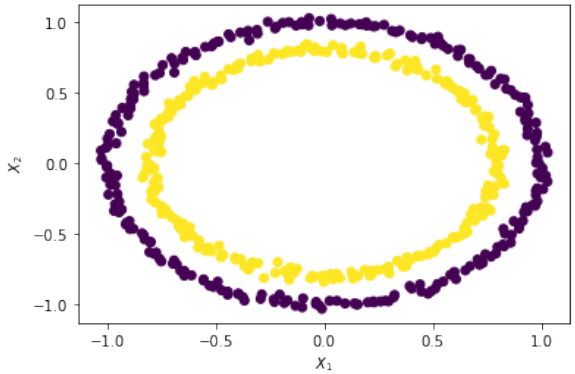
Εισαγωγή

Α. Ν. Γιαννακόπουλος

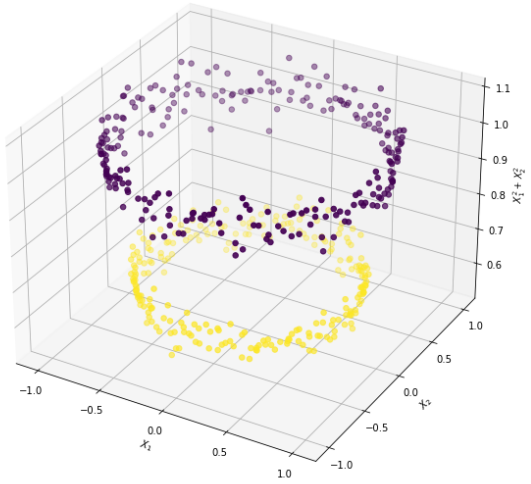
Ο.Π.Α

Χειμερινό Εξάμηνο 2025-2026

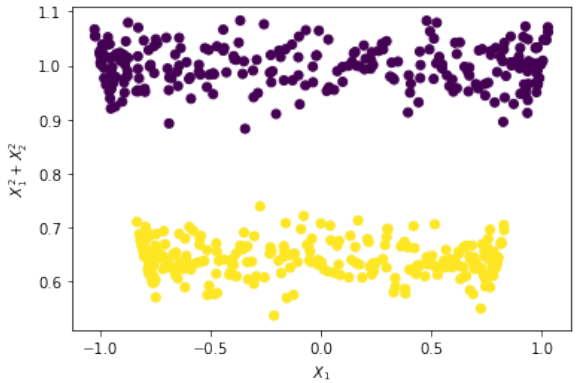
Example 1: Concentric data



(a) Original feature space

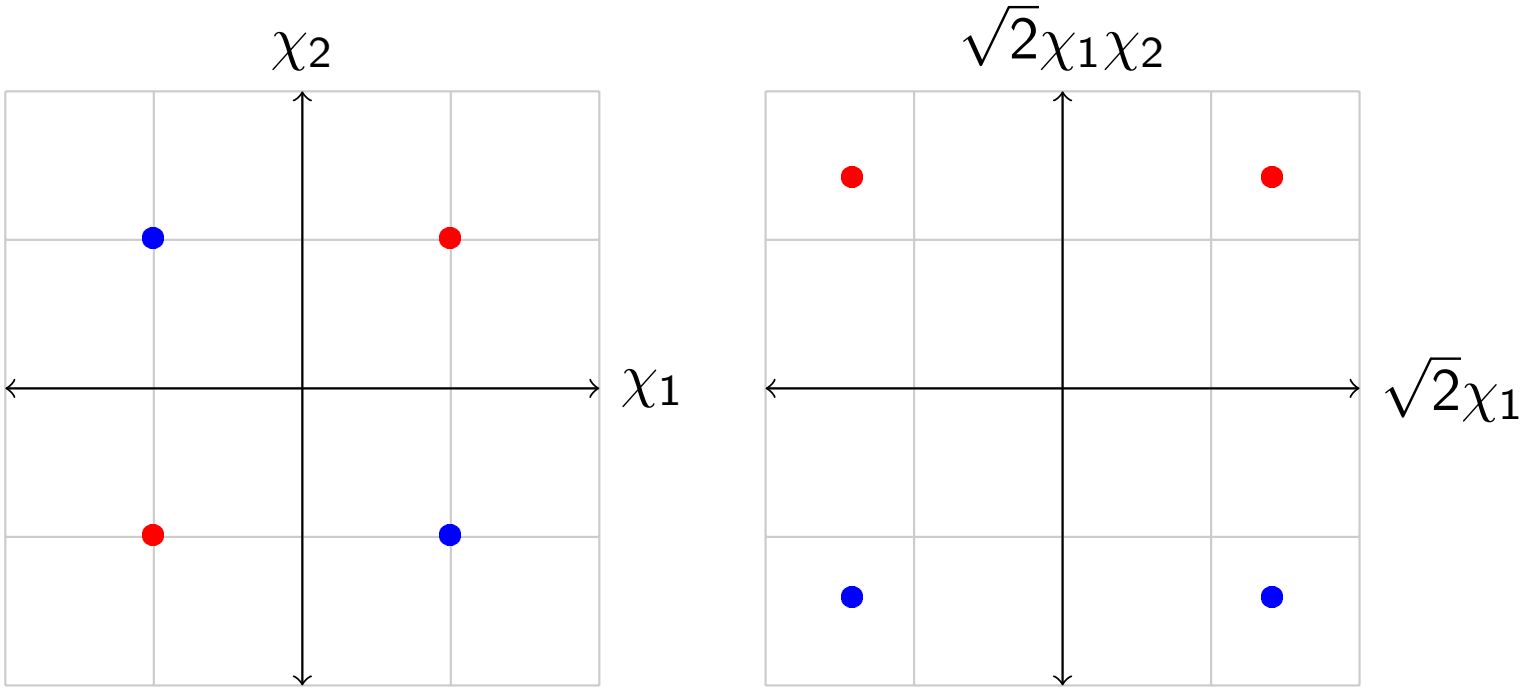


(b) Transformed feature space



(c) Transformed feature

XOR data



(a) Original feature space (b) Transformed feature space

Figure: XOR data and their nonlinear transformation

$X = \{\underline{X}_1, \cdots, \underline{X}_N\} \subset \mathbb{R}^n \mapsto Z = \{\underline{Z}_1, \cdots, \underline{Z}_N\} \subset \mathbb{R}^m$, in terms of the nonlinear coordinate transform Φ, by $\underline{Z}_i = \Phi(\underline{X}_i)$, $i = 1, \cdots, N$, we obtain a new data set Z in the higher dimensional space $\mathbb{R}^m$.

We will then treat this new data set instead of the original data set X, hoping that in the new higher dimensional space the data will be better represented and questions such as for example classification in terms of linear models will be more easily treated in the new representation

Kernel methods, exploit this observation, while at the same time considering the question of how computationally demanding would be the need to perform the nonlinear coordinate change explicitly by calculating all the new coordinates $(Z_1, \cdots, Z_m)$ $(m > n)$ for each one the data points $\underline{X}_i$ (N in total).

- $\mathbb{X}$ original feature space $dim(\mathbb{X}) = n$ (low dimensional) — Possibly difficulties in representation.
- $\mathbb{H}$ new feature space $dim(\mathbb{H}) = m$ (high dimensional - possibly infinite dimensional) — Easier (linear) representation
- $\Phi : \mathbb{X} \to \mathbb{H}$ feature map, transformation $\underline{X} \to \underline{Z} = \Phi(\underline{X})$ leads to better representation.

The important quesion is

What is this map Φ going to be?

Luckily, the theory works fine most of the times for some predescribed mappings Φ, that are constructed in terms of kernel functions.

Definition

We will call a kernel function K, a function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, a function satisfying the following properties:

1. K is symmetric, i.e. $K(x, x') = K(x', x)$ for any $x, x' \in \mathbb{X}$,
2. The Gram matrix generated by K is positive semidefinite, i.e. for any set $\{x_1, \cdots, x_N\} \subset \mathbb{X}$ and any $z = (z_1, \cdots, z_N) \in \mathbb{R}^N$ it holds that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) z_i z_j \geq 0.$$

The matrix $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^{N}$ is called the Gram matrix for the given set.

It is important to note the following:

- Given any mapping $\Phi : \mathbb{X} \to \mathbb{H}$, then $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, defined by $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}}$ is a kernel function in the above sense.
- A kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ can be used to generate a mapping $\Phi : \mathbb{X} \to \mathbb{H}$ that can be used for better data representation (additional assumptions required).

From feature map $\Phi : \mathbb{X} \to \mathbb{H}$ to kernel functions
We first consider how we can obtain kernel functions from feature maps:

Theorem

Given a feature map $\Phi : \mathbb{X} \to \mathbb{H}$, we can obtain the kernel function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ defined by

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}}, \quad \forall\, x, x' \in \mathbb{X}, \tag{1}$$

where by $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ we denote the inner product in the Hilbert space $\mathbb{H}$.

Of course one can construct new kernel functions from existing ones as follows:

Theorem (New kernels from old ones)

The following hold:

(i) *If K_1, K_2 are kernel functions and $\lambda_1, \lambda_2 \geq 0$, then so is $\lambda_1 K_1 + \lambda_2 K_2$.*

(ii) *If K_1, K_2 are kernel functions then so is $K_1 K_2$.*

(iii) *If $(K_n)_{n\in\mathbb{N}}$ is a sequence of kernel functions and the limit exists then $K := \lim_n K_n$ is a kernel function.*

(iv) *If K is a kernel function and $f : \mathbb{X} \to \mathbb{R}$ is any function then the new function $K_f : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, defined by*

$$K_f(x, x') = f(x) f(x') K(x, x'), \ \ \forall\, x, x' \in \mathbb{X},$$

is a kernel function.

Finally for kernels that can be expressed in terms of differences

$$K(x, x') = \phi(x - x'), \ \ \forall\ x, x' \in \mathbb{X}, \tag{2}$$

a general characterization can be obtained in terms of the celebrated Bochner theorem.

Theorem (Bochner)

A function defined as in (2) *with* ϕ *continuous, is positive definite if and only if* ϕ *is the characteristic function of a random variable* $Y : \Omega \to \mathbb{X}$*, where* $(\Omega, \mathcal{F}, \mu)$ *is a probability space. That means that* ϕ *can be expressed in terms of the Fourier transform*

$$\phi(x) = \int \exp(ix \cdot y) d\mu(y),$$

or assuming that μ *has density function* g*,*

$$\phi(x) = \int \exp(ix \cdot y) g(y) dy.$$

If ϕ *is an even function then* K *is a kernel function.*

Using Bochner's theorem we may easily construct kernel functions. Moreover, it allows us to draw useful and intuitive connections with probability theory.

From kernels to feature maps

Definition (Reproducing Kernel Hilbert Spaces (RKHS))

A Hilbert space H is called a reproducing Hilbert space if for any function $f : \mathbb{X} \to \mathbb{R}$, $f \in H$, the mapping $x \mapsto f(x)$ is a continuous map (considered as a map $L : H \to \mathbb{R}$).

Definition (RKHS II)

A Hilbert space H consisting of functions $f : \mathbb{X} \to \mathbb{R}$ is called a reproducing Hilbert space (RKHS) with reproducing kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, if

1. For each $x \in \mathbb{X}$, $K(x, \cdot) \in H$ (by $K(x, \cdot)$ we denote the mapping $x \to K(x, \cdot) =: \varphi_x(\cdot)$, such that $\varphi_x(x') = K(x, x')$ for all $x' \in \mathbb{X}$).
2. For each $f \in H$, and $x \in \mathbb{X}$,

$$f(x) = \langle f, K(x, \cdot) \rangle_H = \langle f, \varphi_x \rangle_H, \quad \text{(Reproducing property)}. \tag{3}$$

What type of functions are included in a RKHS? By construction, it is rather easy to characterize the elements of this space.

Any element $f \in H$ is of the form

$$f(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot), \tag{4}$$

for some set $\{x_1, \cdots, x_N\} \subset \mathbb{X}$, and some N, and $a_i \in \mathbb{R}$, or the limit as $N \to \infty$ of such a function.

Theorem

For every kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, *there exists a Hilbert space* $\mathbb{H} = H_K$ *(a RKHS with kernel* K*) and a mapping* $\Phi : \mathbb{X} \to \mathbb{H}$ *such that*

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}}, \quad \forall\, x, x' \in \mathbb{X}.$$

The space H_K *can be characterized as the functions of the form* (4) *and their limits in terms of the norm* $\|\cdot\|$.

The reproducing property allows us to construct the mapping $\Phi : \mathbb{X} \to \mathbb{H}$ from the knowledge of the kernel.

Indeed, consider the mapping Φ defined for any $x \in \mathbb{X}$ by $\Phi(x) := K(x, \cdot) \in \mathbb{H}$,

$$\Phi : \mathbb{X} \to \mathbb{H} = H_K, \;\; \mathbb{X} \ni x \mapsto K(x, \cdot) =: \Phi(x) \in \mathbb{H} = H_K.$$

Then applying the reproducing property (3) to the function $f(\cdot) = K(x', \cdot)$ we have that

$$f(x) \overset{(3)}{=} \langle f, K(x, \dot{)} \rangle_{\mathbb{H}} = \langle K(x', \cdot), K(x, \dot{)} \rangle_{\mathbb{H}} = K(x', x),$$

where for the last equality we used once more the reproducing property (3), but now for the element $K(x', \cdot) \in \mathbb{H} = H_K$.

The above, using the symmetry of the inner product and of the kernel function implies that

$$K(x, x') = \langle K(x, \cdot), K(x', \cdot) \rangle_{\mathbb{H}} = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}},$$

where in the last equality we used the definition of Φ.

The kernel trick

Consider the transformation $\mathbb{X} \ni x \rightarrow z \in \mathbb{H}$, where $\mathbb{X}$ is the original feature space and $\mathbb{H}$ is the new feature space where we hope that the data are better and more easily represented.

We will work with the new transformed data $z = \Phi(x)$, and for most data science applications we will require to calculate $\langle z, z' \rangle_{\mathbb{H}}$, which can be considered as a measure of " affinity " of the transformed data $z = \Phi(x)$ and $z' = \Phi(x')$.

On account of the high dimensionality of the new feature space, this calculation (needed for the calculation of the Gram matrix in the new feature space) is very expensive.

However, if the transformation Φ is related to a RKHS, i.e. $\mathbb{H} = H_K$, then,

$$\langle z, z' \rangle_{H_K} = \langle \Phi(x), \Phi(x') \rangle_{H_K} = K(x, x'),$$

so the inner products in $\mathbb{H}$ can be calculated directly in terms of the calculation of the kernel function K on the data points x, x' in the original (low dimensional!) feature space $\mathbb{X}$.

This is clearly a lot cheaper than performing the inner product $\langle z, z' \rangle_{H_K}$ directly.

This is one of the reasons why out of all possible data transformations we favour those that are compatible with a RKHS structure!

For such cases the Gram matrix of the transformed data (in $\mathbb{H} = H_K$),

$$\{x_1, \cdots, x_N\} \mapsto \{z_1, \cdots z_N\} = \{\Phi(x_1), \cdots, \Phi(x_N)\},$$

is expressed in terms of the Gram matrix

$$\mathbf{K} = (K_{ij})_{i,j=1}^{N}, \quad K_{ij} = K(x_i, x_j).$$

This matrix (which can be very easily calculated in terms of the original data) is essentially all we require for most of our kernel algorithms calculations. Most of the time we will not ever need to look at the data in the new space! (eventhough the calculations and the methodology is made possible because we work in this new space).

The linear kernel

The linear kernel $K(x, x') = x^T x' = \langle x, x \rangle_{\mathbb{R}^n}$, corresponding to the feature mapping $\Phi(x) = z = x$ (the identity map). This can be seen in two ways.

(a) By directly expressing the kernel function K in terms of $K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle x, x' \rangle = x^T x'$, where $\Phi(x) = x$ for every $x \in \mathbb{X}$.

(b) By considering the functions $K(x, \cdots)$ defined by $K(x, x') = x^T x'$ for any $x' \in \mathbb{X}$.

To fully specify this function $K(x, \cdot)$ for any $x \in OFS$ we need to specify only x (since then the function is created by the inner product $K(x, \cdot) = \langle x, \cdot \rangle$ so we can identify this function with x.

In this sense $\Phi(x) = K(x, \cdot) \simeq x$, i.e. Φ can be identified with the identity mapping.

Polynomial kernels

The polynomial kernels $K(x, x') = (c_1 + c_2 x^T x')^d = (c_1 + c_2 \langle x, x' \rangle)^d$ for $c_1 \geq 0$, $c_2, d \in \mathbb{N}$ corresponding to a higher dimensional polynomial map.

For example in the case where $c_1, c_2 = 1$ and $d = 2$ for $n = 2$ we obtain the mapping $x = (x_1, x_2) \mapsto z = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2) \in \mathbb{R}^6$. This can be seen again in two ways.

(a) By directly noting that

$$K(x, x') = (1 + x^T x')^2 = 1 + 2x^T x' + (x^T x')^2 = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^6}$$

for $\Phi(x) = \Phi(x_1, x_2) = (1, \sqrt{2}x_1, \sqrt{2}x_2, , x_1^2, x_2^2, \sqrt{2}x_1 x_2) \in \mathbb{R}^6$, so that Φ can be considered as a mapping $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$.

(b) By considering for any $x \in \mathbb{X} = \mathbb{R}^2$ the functions $K(x, \cdot)$ defined by

$$x' \mapsto K(x, x') = (1 + x^T x')^2 = 1 + 2x^T x' + (x^T x')^2$$
$$= 1 + 2x_1 x_1' + 2x_2 x_2' + x_1^2 (x_1')^2 + x_2^2 (x_2')^2 + 2x_1 x_2 x_1' x_2',$$

which is second order polynomial in $x' = (x_1', x_2')$.

This polynomial can be uniquely determined by 6 coefficients depending on $x = (x_1, x_2)$ and in particular by the vector
$z = (1, \sqrt{2}x_1, \sqrt{2}x_2, , x_1^2, x_2^2, \sqrt{2}x_1 x_2) \in \mathbb{R}^6$.

Since the function $K(x, \cdot)$ is uniquely determined by the vector z as defined above, we can identify $\Phi(x) = K(x, \cdot) \simeq z$.

The Gaussian kernel

The Gaussian kernel $K(x, x') = \exp(-\gamma \|x - x'\|^2)$, for $\gamma > 0$.

The fact that this kernel is expressed in terms of inner products in the space $\mathbb{X}$ can be seen by expressing $\|x - x'\|^2 = x^T x' - 2x^T x' + (x')^T x'$.

This kernel produces a map to an infinite dimensional feature space $\mathbb{H}$.

Mercer's theorem

For a particular class of kernels we can easily construct explicitly the corresponding RKHS and the feature map.

This is covered by the celebrated Mercer's theorem.

Theorem (Mercer)

Let $(\mathbb{X}, \mathcal{F}, \mu)$ be a measure space, with $\mathbb{X}$ compact, set $K : \mathbb{X} \to \mathbb{X} \to \mathbb{R}$ be a continuous positive definite kernel function such that

$$\int_{\mathbb{X}} \int_{\mathbb{X}} |K(x, x')|^2 d\mu(x) d\mu(x') < \infty,$$

and consider the corresponding integral operator $T_K : L^2(\mathbb{X}, \mathcal{F}, \mu) \to L^2(\mathbb{X}, \mathcal{F}, \mu)$ defined by

$$f \mapsto T_K f =: g, \;\; g(x) := \int_{\mathbb{X}} K(x, x') f(x') d\mu(x').$$

(i) The operator T_K is a compact operator that admits a denumerable set $\{(\lambda_j, \phi_j), \; j \in \mathbb{N}\}$ of eigenvalues and eigenfunctions, i.e. solutions to the problem

$$T_K \phi_j = \lambda_j \phi_j, \;\; j \in \mathbb{N},$$

with $\lambda_j \geq 0$. Using a standard Gram-Schmitd orthogonalization procedure, we can turn this set into an orthonormal set (assume from now on).

Theorem (Mercer continued ...)

(ii) The sequence can be turned into a basis of the corresponding Hilbert space and in particular we can use the expansion

$$K(x,x') = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(x'), \ \ \forall\ x, x' \in \mathbb{X},$$

wth the above sum being uniformly convergent.
(iii) In this case we may construct the corresponding RKHS $\mathbb{H} = H_K$ explicitly in terms of

$$H_K := \{\sum_{j=1}^{\infty} c_j\phi_j, \ :\ z_j \in \mathbb{R}, \ \text{such that} \ \sum_{j=1}^{\infty} \frac{|c_j|^2}{\lambda_j} < \infty\},$$

$$\|z\|^2_{H_K} = \sum_{j=1}^{\infty} \frac{|c_j|^2}{\lambda_j}, \ \ \forall\ z = \sum_{j=1}^{\infty} c_j\phi_j \in H_K,$$

where $z \in H_K$ is to be understood as a function $z \equiv f : \mathbb{X} \to \mathbb{R}$. In terms of this notation (and assuming that we have performed the orthonalization step) the inner product can be expressed as

$$\langle f_1, f_2\rangle_{H_K} = \sum_{j=1}^{\infty} \frac{1}{\lambda_j}\langle f_1, \phi_j\rangle\langle f_2, \phi_j\rangle,$$

$$\langle f_i, \phi_j\rangle = \int_{\mathbb{X}} f_i(x)\phi_j(x)d\mu(x), \ \ i = 1, 2.$$

The representer theorem.

Theorem (Representer theorem)

Consider the set of data $\{(x_i, y_i),\ i = 1, \cdots, N\} \subset \mathbb{X} \times \mathbb{R}$, *a kernel function* $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ *and the corresponding RKHS* $\mathbb{H} = H_K$. *Consider also the loss function*

$$\mathcal{E}(f) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_i(y_i, f(x_i)) + R(\|f\|_{H_K}),$$

where $R(\|f\|_{H_K})$ *is a regularization term, and* $R : [0, \infty) \to \mathbb{R}$ *is a strictly increasing function, and* $\mathcal{E}_i$ *is an arbitrary loss function.*
Then, the minimizer of $\mathcal{E}(f)$ *in* H_K *admits a representation of the form*

$$f^*(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot), \quad a_i \in \mathbb{R},$$

with $a_i \in \mathbb{R}$, *to be obtained numerically in terms of an optimization problem in* $\mathbb{R}^N$.

The representer theorem is important for various reasons:
(i) It guarantees that a suitable machine learning model for our data is a nonlinear model of the form

$$f^*(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot) = \sum_{i=1}^{N} a_i \psi_i(\cdot), \quad a_i \in \mathbb{R},$$

i.e. in terms of the set of "basis functions" $\{\psi_i,\ i = 1, \cdots, N\}$ that are determined in terms of the kernel function K. Importantly, the basis functions ψ_i are determined (shaped) by the data themselves since $\psi_i(\cdot) := K(x_i, \cdot)$. In this respect, for each data set we may obtain different looking basis functions, depending on the training data, even though the kernel function K is always the same.

(ii) Naturally, by changing the RKHS H_K in which we consider our model, we will obtain different models (i.e. we will employ different kernel functions K). Intimately connected with that is the regularization term, the nature of which is determined in terms of the choice of norm for the RKHS, $\|\cdot\|_{H_K}$. Different norms impose different qualitative features on the chosen model, on properties such as e.g. smoothness of the model etc.

(iii) Another important consequence of the representer theorem is that it greatly simplifies the nature of the learning problem, in terms of the optimization problem involved. The original problem of choosing $f^* \in \arg\min_{f \in \mathbb{H}} \mathcal{E}(f)$ is an infinite dimensional problem, that present both theoretical (mathematical) as well as computational difficulties. The mathematical difficulty is that the generalization of the Weierstrass maximum theorem for infinite dimensional spaces is tricky, as the characterization of compact sets in such spaces is not a straighforward (although not impossible!) task.

"Kernelized models"

A kernelized model is looking for an ML model in an appropriate RKHS $f \in H_K$, and using as a regularization term a term of the form $R(\|f\|_K)$,

$$\min_{f \in H_K} \mathcal{E}(f) := \min_{f \in H_K} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_i(y_i, f(x_i)) + R(\|f\|_{H_K}) \right).$$

Then use the representer theorem to transform the problem to the finite dimensional problem

$$\min_{(a_1, \cdots, a_N) \in \mathbb{R}^N} \mathcal{E}(a_1, \cdots, a_N),$$

which is obtained upon substituting $f(x) = \sum_{i=1}^{N} a_i K(x_i, x)$ in the original problem.

Then, this is solved using a standard optimization method, and the $a_1, \cdots, a_N$ are used to recover the model.

Kernel ridge regression

The kernel ridge regression model corresponds to the loss function

$$\mathcal{E}(f) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda\|f\|_{H_K}^2$$

This is the kernelized generalization of the standard linear ridge regression model.

The choice of K (equiv. H_K) is related to the smoothness of the functional model f.

By the representer theorem the solution of

$$\min_{f \in H_K} \mathcal{E}(f),$$

has to be of the form

$$f(x) = \sum_{i=1}^{N} a_i K(x_i, x),$$

for suitable choice of $(a_1, \cdots, a_N) \in \mathbb{R}^N$.

We will use this representation to transform the original functional (infinite dimensional) problem to a finite dimensional one with respect to the coefficients a_i, $i = 1, \cdots, N$.

We perform some necessary calculations first.

Fidelity term:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}(y_i - \sum_{j=1}^{N} a_j K(x_j, x_i))^2$$

$$= \sum_{i=1}^{N}(y_i - (Ka)_i)^2 = \|y - Ka\|_2^2,$$

where

$$K = (K_{ij})_{i,j=1}^{N}, \quad K_{ij} = k(x_i, x_j), \quad \text{Gram matrix}, \quad K = K^T,$$

$$(Ka)_i, \quad \text{i-th component of vector } Ka$$

Note the similarity with the standard linear regression problem – Here the linear design matrix is replaced by the nonlinear Gram matrix.

Regularization term:

$$\|f\|_{H_K}^2 = \langle f, f \rangle_{H_K} = \langle \sum_{i=1}^{N} a_i K(x_i, \cdot), \sum_{j=1}^{N} a_j K(x_j, \cdot) \rangle_{H_K}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{H_K}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(x_i, x_j) = \langle a, Ka \rangle_{\mathbb{R}^N}.$$

In the above we used the fundamental RKHS property that

$$\langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{H_K} = K(x_i, x_j).$$

Note the similarity with the standard linear regression problem – Here the ridge regularization term $\|a\|_2^2$ replaced by the quadratic term $\langle a, Ka \rangle_{\mathbb{R}^N}$, a new norm weighted by the nonlinear Gram matrix.

The original problem

$$\min_{f \in H_K} \mathcal{E}(f) \iff \min_{a=(a_1,\cdots,a_N)^T \in \mathbb{R}^N} \frac{1}{N}\|y - Ka\|_2^2 + \lambda \langle a, Ka \rangle_{\mathbb{R}^N}$$

We redefine λ so as to set $N = 1$.

The first order condition becomes

$$K(y - Ka) - \lambda Ka = 0 \iff K(K + \lambda I)a = Ky$$

which can be solved if K is non singular in terms of

$$a^* = (K + \lambda I)^{-1} y$$

The chosen model is

$$f^*(x) = \sum_{i=1}^{N} a_i^* K(x_i, x).$$

Kernel SVM

Support vector machines (SVMs) are useful models for classifying binary data y_i, in terms of multidimensional features x_i.

Recall the standard linear SVM:

Given data $(x_i, y_i) \in \mathbb{R}^m \times \{-1, 1\}$, $i = 1, \cdots, N$, find a linear separation boundary given by the hyperplane $b_0 = \sum_{j=1}^{m} b_j z_j = \langle b, z \rangle_{\mathbb{R}^m}$ such that the data are separated by this boundary.

In particular, if

$$\begin{aligned} \langle b, x_i \rangle > b_0 &\implies y_i = +1, \\ \langle b, x_i \rangle < b_0 &\implies y_i = -1. \end{aligned}$$

More than one hyperplanes can separate the data: We will choose the one that is optimal in terms of maximizing a margin (related to the distance of the boundary of the two populations from this boundary).

This is often expressed as the optimization problem

$$\max_{b\in\mathbb{R}^m, b_0\in\mathbb{R}} M,$$
$$\langle b, x_i\rangle - b_0 > M \text{ if } y_i = +1,$$
$$\langle b, x_i\rangle - b_0 < -M \text{ if } y_i = -1$$

or in the equivalent form

$$\max_{b\in\mathbb{R}^m, b_0\in\mathbb{R}} M,$$
$$y_i(\langle b, x_i\rangle - b_0) > M, \ i = 1, \cdots, N$$

Using convex duality methods we can show that the solution to this problem must have the form

$$b = \sum_{i=1}^{N} a_i y_i x_i, \quad a_i \in \mathbb{R},$$

with the a_i to be obtained in terms of the solution of the quadratic optimization problem

$$\max_{(a_1,\cdots,a_N)\in\mathbb{R}_+^N} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \langle x_i, x_j \rangle_{\mathbb{R}^m} a_i a_j$$

The kernelized SVM method starts from this dual problem by replacing the original dual problem by its kernelized variant

$$\max_{(a_1,\cdots,a_N)\in\mathbb{R}_+^N} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(x_i, x_j) a_i a_j,$$

where K is a chosen kernel function.

Since

$$K(x_i, x_j) = \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{H_K},$$

this can be interpreted as looking for a separation boundary of the form

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x),$$

which is a nonlinear separation boundary.

Note that the kernelized version of the dual problem is formally the same as the non kernelized one.

Both are quadratic optimization problems of the form

$$\max_{a \in \mathbb{R}^N_+} \langle \mathbf{1}, a \rangle_{\mathbb{R}^N} - \frac{1}{2} \langle a, Ga \rangle_{\mathbb{R}^N},$$

for some positive (semi) definite matrix $G \in \mathbb{R}^{N \times N}$.

- In the standard case

$$G = (G_{ij})_{i,j=1}^N, \quad G_{ij} = y_i \langle x_i, x_j \rangle_{\mathbb{R}^m} y_j,$$

- In the kernelized case

$$G = (G_{ij})_{i,j=1}^N, \quad G_{ij} = y_i K(x_i, x_j) y_j = y_i \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_{H_K} y_j,$$