GI13/4C60: Advanced Topics in Machine Learning

Part 1: Kernels in Machine Learning

Massimiliano Pontil

About these lectures

Theme: We introduce the elements of the course and establish our notation. We first review a simple learning method for linear regression. This method serves to illustrate the idea of feature maps and kernels in action. We then present the notion of positive definite kernel functions, characterize some of their properties and discuss polynomial and translation invariant kernels.

Math required: Calculus, elements of linear algebra and optimization, Supervised Learning (GI01/4C55).

Outline

- Ridge regression
- Feature maps
- Positive definite kernels
- Kernel construction
- Kernels on Euclidean spaces

Notation

We summarize our notation. More symbols will be introduced during the course when needed.

$\mathcal{X}, \mathcal{Z}, \mathcal{W}$ denote sets, x, z, w their elements

$\mathbb{N}$ is the set of natural numbers, $\mathbb{R}$ the set of real ones

$N_n := \{1, \ldots, n\}$ is the set of integers between 1 and up to n

$\mathbb{R}^n$ is the set of $n-$dimensional vectors. $\mathbf{x} \in \mathbb{R}^n$ is regarded as a column vector ($n \times 1$ matrix), $\mathbf{x}' = (x_1, \ldots, x_\ell) \equiv (x_i : i \in \mathbb{N}_n) \equiv (x_i)_{i=1}^n$, ($1 \times n$ matrix)

If $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathbb{R}^n$, $\mathbf{X}$ denote the $\ell \times n$ matrix whose rows are the row vectors $\mathbf{x}_j'$. Alternatively, $\mathbf{X}' = [\mathbf{x}_1, \ldots, \mathbf{x}_\ell]$

If $\mathbf{A} = (A_{ij} : i \in \mathbb{N}_n, j \in \mathbb{N}_m)$ is a $n \times m$ matrix, $\mathbf{A}'$ denote its transpose and $\mathbf{A}^+$ its pseudoinverse.

$\mathbf{I}$ (or sometimes $\mathbf{I}_n$) is the $n \times n$ identity matrix

If $\mathcal{W}$ is a Hilbert space, $\langle w, z \rangle$ denotes the inner product and $\|w\| := \sqrt{\langle w, w \rangle}$ the associated norm

Notation (cont.)

By a kernel we mean a positive semidefinite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$\mathbf{K}$ is the kernel matrix, $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j) : i, j \in \mathbb{N}_\ell,\ \ell \in \mathbb{N}$

$\boldsymbol{\phi} : \mathcal{X} \to \mathcal{W}$ is a feature map and $\{\boldsymbol{\phi}(x) : x \in \mathcal{X}\}$ the feature space.

psd means *positive semidefinite*

SVD means *singular value decomposition*

Linear regression

Problem: We wish to find a function $g(\mathbf{x}) = \mathbf{w}'\mathbf{x}$ which best interpolates a data set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\} \subseteq \mathbb{R}^n \times \mathbb{R}$

- If the data have been generated in the form $(\mathbf{x}, g(\mathbf{x}))$, the inputs are independent and $\ell = n$ then there is a unique interpolant whose parameter $\mathbf{w}$ solves the equation

  $$\mathbf{X}\mathbf{w} = \mathbf{y}$$

  where $\mathbf{y} \in \mathbb{R}^\ell$ is the vector $(\mathbf{y})_i = y_i$, $i \in \mathbb{N}_\ell$.

- Otherwise, this problem is *ill-posed*

Remark: : If we wish to learn a non-homogeneous function $g(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$ we simply add one more dimension to $\mathbf{x}$ and $\mathbf{w}$. Specifically, we replace $\mathbf{x}'$ by the vector $(\mathbf{x}', 1)$ and $\mathbf{w}'$ by $(\mathbf{w}', b)$.

Ill-posed problems

Informally, a problem is well-posed (in the sense of Hadamard) if

(1) a solution exists

(2) the solution is unique

(3) the solution depends continuously on the data

A problem is ill-posed if it is not well-posed

Learning problems are in general ill-posed (usually because of (2)). A theme of this course is that regularization is an effective mean to make the learning problem well-posed. We will focus on regularization in a Hilbert space

Ridge regression

We define the square error as

$$\mathcal{L}(\mathbf{w}; S) := \sum_{i=1}^{\ell} (y_i - \mathbf{w}'\mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w})$$

and solve the variational problem

$$\min_{\mathbf{w}} \{\mathcal{L}(\mathbf{w}, S) + \lambda \mathbf{w}'\mathbf{w}\}, \quad \lambda > 0 \tag{1}$$

The parameter λ defines a trade-off between the error on the data and the norm of the vector $\mathbf{w}$ (degree of regularization)

Setting the derivative of $\mathcal{L}_\lambda$ wrt. $\mathbf{w}$ to zero,

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w} = 0 \tag{2}$$

we obtain the *regularized solution*

$$\mathbf{w} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_n)^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

or

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_n)^{-1}\mathbf{x} \tag{4}$$

Singular valued decomposition (review)

Let $\mathbf{X}' = [\mathbf{x}_1, \cdots, \mathbf{x}_\ell]$ be the an $n \times \ell$ data matrix. Singular value decomposition (SVD) establishes that

$$\mathbf{X}'\mathbf{X} = \mathbf{U}\Lambda_n\mathbf{U}' \quad \mathbf{X}\mathbf{X}' = \mathbf{V}\Lambda_\ell\mathbf{V}'$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_\ell]$,

$$\Lambda_n = diag(\lambda_1, \ldots, \lambda_t, \mathbf{0}_{n-t}), \qquad \Lambda_\ell = diag(\lambda_1, \ldots, \lambda_t, \mathbf{0}_{\ell-t}),$$

$t = rank(\mathbf{X}\mathbf{X}') = rank(\mathbf{X}'\mathbf{X})$ and $\lambda_1 \geq \cdots \geq \lambda_t > 0$, $t \leq \min(\ell, n)$. Moreover, we have

$$\mathbf{X}' = \sum_{i=1}^{t} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i' = \mathbf{U}\Sigma\mathbf{V}'$$

where Σ is the $n \times \ell$ matrix with leading diagonal entries $\sigma_j = \sqrt{\lambda_j}$

Generalized solution

When λ goes to zero $\mathbf{w}$ tends to

$$\mathbf{w}_0 := (\mathbf{X}'\mathbf{X})^+\mathbf{X}'\mathbf{y} \tag{5}$$

where $(\mathbf{X}'\mathbf{X})^+$ is the pseudoinverse of $\mathbf{X}'\mathbf{X}$. Using SVD this is given by

$$(\mathbf{X}'\mathbf{X})^+ = \sum_{i=1}^{t} \sigma_i^{-1}\mathbf{u}_i\mathbf{u}_i'$$

- $g_0(\mathbf{x}) = \mathbf{w}_0'\mathbf{x}$ is called the *generalized solution*. It is *the* function which, among those which minimize $\mathcal{L}(\mathbf{w}; S)$, has the smallest norm of its coefficients

Dual representation

We defined the kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as $K(\mathbf{x}_i, \mathbf{x}) := \mathbf{x}_i'\mathbf{x}$. The regularized solution can be written as a linear combination of ℓ kernel functions centered at the data points

$$g(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_i : i \in \mathbb{N}_n)'$ is given by

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_\ell)^{-1}\mathbf{y} \tag{7}$$

and $\mathbf{K} := (K(\mathbf{x}_i, \mathbf{x}_j) : i, j \in \mathbb{N}_n)$

- **Function representations:** we call the functional form (or representation) $g(\mathbf{x}) = \mathbf{w}'\mathbf{x}$ the *primal form* and (1) the *dual form* (or representation)

Dual representation (cont.)

Proof of eqs.(6),(7): We rewrite eq.(2) as

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i \tag{8}$$

where

$$\alpha_i := \frac{y_i - \mathbf{w}'\mathbf{x}_i}{\lambda} \tag{9}$$

Consequently, we have that $\mathbf{w}'\mathbf{x} = \sum_{i=1}^{\ell} \alpha_i \mathbf{x}_i'\mathbf{x} = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}_i, \mathbf{x})$. This proves eq.(6). Plugging eq.(8) in eq.(9) we obtain

$$\sum_{j=1}^{\ell} (K(\mathbf{x}_i, \mathbf{x}_j) + \lambda \delta_{ij})\alpha_j = y_i, \quad \text{or, in matrix notation :} \quad (\mathbf{K} + \lambda \mathbf{I}_\ell)\boldsymbol{\alpha} = \mathbf{y}$$

from which eq.(7) follows.

Remark: in the limit of λ going to 0 $\boldsymbol{\alpha}$ goes to $\boldsymbol{\alpha}_0 = \mathbf{K}^+\mathbf{y}$

Computational considerations

Training time:

- Solving for $\mathbf{w}$ (eq.(3)) requires $O(n^3)$ operations while solving for $\boldsymbol{\alpha}$ (eq.(7)) requires $O(n\ell^2 + \ell^3)$ operations

If $\ell \ll n$ it is more efficient to use the dual representation

Running (testing) time:

- Computing g in the primal form (eq.(6)) requires $O(n)$ operations, while the dual form (eq.(4)) requires $O(\ell n)$ operations

Sparse representations

Suppose each input $\mathbf{x} \in \mathbb{R}^n$ has most of its components equal to zero (eg, we consider a dataset of images where most pixels are 'black' or a dataset of short text documents represented as 'bag of words',...)

- If k denotes the number of nonzero components of the input then computing $\mathbf{x}'\mathbf{t}$ requires at most $O(k)$ operations

- If $k\ell \ll n$ (so, also $\ell, k \ll n$) the dual representation is advantageous versus the primal representation both for training ($O(k\ell^2 + \ell^3)$ vs. $O(n^3)$) and testing ($O(\ell k)$ vs. $O(n)$)

Feature map

The above ideas can naturally be generalized to nonlinear function regression.

By a *feature map* we mean a function $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^N$,

$$\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x})), \quad \mathbf{x} \in \mathbb{R}^n$$

(typically $N \gg n$). $\boldsymbol{\phi}(\mathbf{x})$ is called the *feature vector* and the space $\{\boldsymbol{\phi}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$ the *feature space*

The non-linear regression function has the primal representation

$$g(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle := \sum_{j=1}^{N} w_j \phi_j(\mathbf{x})$$

Remark: Later we will discuss feature maps whose range is a Hilbert space (informally, an infinite dimensional space with the structure of $\mathbb{R}^N$)

Computational considerations

Again, if $\ell \ll N$ it is more efficient to work with the dual representation.

Key observation: in the dual representation we don't need to know $\boldsymbol{\phi}$ explicitly; we just need to know the inner product between any pair of feature vectors!

Example: $N = n^2$, $\boldsymbol{\phi}(\mathbf{x}) = (x_i x_j : i, j \in \mathbb{N}_n)$. In this case we have $\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{t}) \rangle = (\mathbf{x}'\mathbf{t})^2$ which requires only $O(n)$ computations whereas $\Phi(\mathbf{x})$ requires $O(n^2)$ computations.

Kernel vs. feature map

Given a feature map $\boldsymbol{\phi}$ we define its associated kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as

$$K(\mathbf{x}, \mathbf{t}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{t}) \rangle, \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^n$$

- Maybe for some feature map $\boldsymbol{\phi}$ computing $K(\mathbf{x}, \mathbf{t})$ (but not $\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{t})$) is independent on N, only depends on n

Example (cont.) If $\boldsymbol{\phi}(\mathbf{x}) = (x_{i_1} x_{i_2} \cdots x_{i_d} : i_1, i_2, \ldots, i_d = 1, \ldots, n)$ then we have that

$$K(\mathbf{x}, \mathbf{t}) = (\mathbf{x}'\mathbf{t})^d.$$

In this case $K(\mathbf{x}, \mathbf{t})$ is computed with $O(n)$ operations, which is independent of d (and essentially of N). On the other hand, computing $\boldsymbol{\phi}(\mathbf{x})$ requires $O(N)$ operations.

Other learning methods

Many other learning methods can be 'kernelized'. Like ridge regression they all require solving a problem where the data appears in the form of inner products and whose solution can be expressed as a linear combination of the kernel centered at the data, eg $g(\mathbf{x}) = \sum_{i=1}^{\ell} c_i K(\mathbf{x}_i, \mathbf{x})$. In particular this includes:

- Other regression or classification methods such as support vector machines, logistic regression, perceptron algorithm, Fisher discriminant analysis, etc.
- Unsupervised learning methods based on projections in feature space such as principal component analysis, partial least squares; $k-$means clustering, anomaly detection, etc.

We will discuss some of these methods later in the course.

Regularization-based learning algorithms

Let us open a short parenthesis and show that the dual form of ridge regression holds true for other loss functions. Let

$$\mathcal{L}_\lambda(\mathbf{w}; S) = \sum_{i=1}^{\ell} V(y_i, \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle) + \lambda\langle \mathbf{w}, \mathbf{w}\rangle, \quad \lambda > 0 \qquad (10)$$

where $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function (before $(V(a,b) = (a-b)^2)$.

Theorem: If V is differentiable wrt. its second argument and $\mathbf{w}$ is a minimizer of $\mathcal{L}_\lambda$ then it has the form

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i) \quad \Rightarrow \quad g(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle = \sum_{i=1}^{\ell} \alpha_i K(x_i, x)$$

This result is usually called the *Representer Theorem*

Representer theorem

Setting the derivative of $\mathcal{L}_\lambda$ wrt. $\mathbf{w}$ to zero we have

$$-\sum_{i=1}^{\ell} V'(y_i, \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle)\boldsymbol{\phi}(\mathbf{x}_i) + 2\lambda\mathbf{w} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i) \qquad (11)$$

where V' is the partial derivative of V wrt. its second argument and we defined

$$\alpha_i = \frac{1}{2\lambda} V'(y_i, \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle) \qquad (12)$$

Thus we conclude that

$$g(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle = \sum_{i=1}^{\ell} \alpha_i \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x})\rangle = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i),$$

Some remarks

- Plugging eq.(11) in the rhs. of eq.(12) we obtain a set of equations for the coefficients α_i:

$$\alpha_i = \frac{1}{2\lambda} V' \left(y_i, \sum_{j=1}^{\ell} K(\mathbf{x}_i, \mathbf{x}_j)\alpha_j \right) \quad , \quad i = 1, \ldots, \ell$$

  when V is the square loss we retrieve the linear eq.(7)

- Substituting the same equation in eq.(10) we obtain an objective function for the α's:

$$\sum_{i=1}^{\ell} V(y_i, (\mathbf{K}\boldsymbol{\alpha})_i) + \lambda \boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha}$$

Remark: the Representer Theorem holds true under more general conditions on V (for example V can be any convex function.)

General feature map

Let $\mathcal{W}$ be a Hilbert space and $\mathcal{X}$ a set. (**Remark:** we now denote by $x, t, \ldots$ the elements of $\mathcal{X}$. Careful! depending on the context, x_i is an element of $\mathcal{X}$ or a component of vector $\mathbf{x} \in \mathbb{R}^n$)

The above ideas can be generalized to

- functions $g : \mathcal{X} \to \mathbb{R}$
- a feature map $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{W}$ and the associated kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$K(x, t) = \langle \boldsymbol{\phi}(x), \phi(t) \rangle, \quad x, t \in \mathcal{X} \tag{13}$$

In the previous slides, $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{W} = \mathbb{R}^N$.

$\mathcal{W}$ can be any Hilbert space; a frequent choice is either $\mathcal{W} = \mathbb{R}^N$ or the space of square summable sequences,

$$\mathcal{W} = \ell_2 = \{x = (x_i : i \in \mathbb{N}) : \sum_{i=1}^{\infty} x_i^2 < \infty\}$$

This is a separable Hilbert space with the inner product $\langle x, t \rangle := \sum_{i=1}^{\infty} x_i t_i$

Redundancy of the feature map

Warning: The feature map is not unique! For ex., if $\mathcal{W} = \mathbb{R}^N$ and ϕ generates K so does $\hat{\phi} = \mathbf{U}\phi$ where $\mathbf{U}$ in an (any!) $N \times N$ orthogonal matrix. Even the dimension of ϕ is not unique!

Example: If $n = 2$, $K(\mathbf{x}, \mathbf{t}) = (\mathbf{x}'\mathbf{t})^2$ is generated by both $\phi(\mathbf{x}) = (x_1^2, x_2^2, x_1x_2, x_2x_1)$ and $\hat{\phi}(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$.

To emphasize this fact, instead of eq.(13) we sometime write

$$K(x, t) = \langle K_x, K_t \rangle \tag{14}$$

where $K_x : \mathcal{X} \to \mathcal{W}$ is a (anyone we like) feature map associated to K.

Remark: The quantity K_x is, at this point, only 'symbolic' but we will see later in the course that $K_x \equiv K(x, \cdot)$ (a function!) and $\mathcal{W}$ is the reproducing kernel Hilbert space of kernel K.

Change of perspective

- Let us start directly with a kernel K and see when K can be expressed an inner product is some feature space (eq.(13))

Question: Given a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which properties of K guarantee that there exist a Hilbert space $\mathcal{W}$ and a feature map $\phi : \mathcal{X} \to \mathcal{W}$ such that $K(x, t) = \langle \phi(x), \phi(t) \rangle$?

Positive definite kernel

Definition: A function $K : \mathcal{X} \times \mathcal{X} \to$ is **positive semidefinite** if it is symmetric and the matrix $(K(x_i, x_j) : i, j \in \mathbb{N}_\ell)$ is positive semidefinite for every $\ell \in \mathbb{N}$ and every $x_1, \ldots, x_\ell \in \mathcal{X}$

Remark: Some authors use the notation 'positive definite' to denote what we have called 'positive semidefinite'

Theorem: K is positive semidefinite if and only if

$$K(x,t) = \langle \phi(x), \phi(t) \rangle, \quad x, t \in \mathcal{X}$$

for some feature map $\phi : \mathcal{X} \to \mathcal{W}$

Positive definite kernel (cont.)

Proof of "$\Leftarrow$": If $K(x,t) = \langle \phi(x), \phi(t) \rangle$ then we have that

$$\sum_{i,j=1}^{\ell} c_i c_j K(x_i, x_j) = \left\langle \sum_{i=1}^{\ell} c_i \phi(x_i), \sum_{j=1}^{\ell} c_j \phi(x_j) \right\rangle = \left\| \sum_{i=1}^{\ell} c_i \phi(x_i) \right\|^2 \geq 0$$

for every choice of $\ell \in \mathbb{N}$, $x_i \in \mathcal{X}$ and $c_i \in R$, $i \in \mathbb{N}_\ell$.

The proof of '$\Rightarrow$' will be given later in the course where we will also discuss reproducing kernel Hilbert spaces.

Kernel construction

Which operations/combinations (eg, products, sums, composition, etc.) of a given set of kernels is still a kernel?

If we address this question we can build more interesting kernels starting from simple ones.

Example: We have already seen that $K(\mathbf{x},\mathbf{t}) = (\mathbf{x}'\mathbf{t})^d$ is a kernel. For which class of functions $p : \mathbb{R} \to \mathbb{R}$ is $p(\mathbf{x}'\mathbf{t})$ a kernel? More generally, if K is a kernel when is $p(K(\mathbf{x},\mathbf{t}))$ a kernel?

General linear kernel

If $\mathbf{A}$ is a $n \times n$ psd matrix the function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$K(\mathbf{x},\mathbf{t}) = \mathbf{x}'\mathbf{A}\mathbf{t}$$

is a kernel

Proof: Since $\mathbf{A}$ is psd we can write it in the form $\mathbf{A} = \mathbf{R}\mathbf{R}'$ for some $n \times n$ matrix $\mathbf{R}$. Thus K is represented by the feature map $\phi(\mathbf{x}) = \mathbf{R}'\mathbf{x}$.

Alternatively, note that:

$$\sum_{ij} c_i c_j \mathbf{x}_i'\mathbf{A}\mathbf{x}_j = \sum_{ij} c_i c_j (\mathbf{R}'\mathbf{x}_i)'(\mathbf{R}'\mathbf{x}_j) = \|\sum_i c_i \mathbf{R}'\mathbf{x}_i\|^2 \geq 0$$

Kernel composition

More generally, if $K : \mathbb{R}^N \times \mathbb{R}^N \to$ is a kernel and $\phi : \mathcal{X} \to \mathbb{R}^N$, then

$$\tilde{K}(x,t) = K(\phi(x), \phi(t)), \quad x,t \in \mathcal{X}$$

is a kernel on $\mathcal{X}$.

Proof: By hypothesis K is a kernel and, so, for every $x_1, \ldots, x_\ell$ the matrix $(K(\phi(x_i), \phi(x_j)) : i,j \in \mathbb{N}_\ell)$ is psd

In particular, the above example corresponds to $K(\mathbf{x}, \mathbf{t}) = \mathbf{x}'\mathbf{t}$ and $\phi(\mathbf{x}) = \mathbf{R}'\mathbf{x}$

Kernel construction (cont.)

Question: If $K_1, \ldots, K_q$ are kernels on $\mathcal{X}$ and $F : \mathbb{R}^q \to \mathbb{R}$, when is the function

$$F(K_1(x,t), \ldots, K_q(x,t)), \quad x,t \in \mathcal{X}$$

a kernel?

Equivalently: when for every choice of $\ell \in \mathbb{N}$ and $\mathbf{A}_1, \ldots, \mathbf{A}_q$ $\ell \times \ell$ psd matrices, is the following matrix psd?

$$(F(A_{1,ij}, \ldots, A_{q,ij}) : i,j = 1, \ldots \ell)$$

We discuss some examples of functions F for which the answer to these questions is YES.

Convex combination of kernels

If $\lambda_j \geq 0$, $j = 1, \ldots, q$ then $\sum_{j=1}^{q} \lambda_j K_j$ is a kernel.

This fact is immediate (a non-negative combination of psd matrices is still psd)

Example: Let $q = n$, $\mathcal{X} = \mathbb{R}^n$ and $K_i(\mathbf{x}, \mathbf{t}) = x_i t_i$.

In particular, this implies that

- aK_1 is kernels if $a \geq 0$
- $K_1 + K_2$ is kernels

Product of kernels

The pointwise product of two kernels K_1 and K_2

$$K(x,t) := K_1(x,t)K_2(x,t), \quad x,t \in \mathcal{X}$$

is a kernel.

Proof: We need to show that if $\mathbf{A}$ and $\mathbf{B}$ are psd matrices, so is $\mathbf{C} = (A_{ij}B_{ij} : i,j \in \mathbb{N}_\ell)$ ($\mathbf{C}$ is also called the Schur product of $\mathbf{A}$ and $\mathbf{B}$). We write $\mathbf{A}$ and $\mathbf{B}$ in their singular value form, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{U}'$, $\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}'$ where $\mathbf{U}, \mathbf{V}$ are orthogonal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_\ell)$, $\Lambda = diag(\lambda_1, \ldots, \lambda_\ell)$. We have

$$\begin{aligned}
\sum_{i,j=1}^{\ell} a_i a_j C_{ij} &= \sum_{ij} a_i a_j \sum_r \sigma_r U_{ir} U_{jr} \sum_s \lambda_s V_{is} V_{js} \\
&= \sum_{rs} \sigma_r \lambda_s \sum_i a_i U_{ir} V_{is} \sum_j a_j U_{jr} V_{js} \\
&= \sum_{rs} \sigma_r \lambda_s (\sum_i a_i U_{ir} V_{is})^2 \geq 0
\end{aligned}$$

Summary of kernel properties

The above results can be summarized as follows:

If K_1, K_2 are kernels on a set $\mathcal{X}$, $a \geq 0$, K a kernel on $\mathbb{R}^N$ and $\phi : \mathcal{X} \to \mathbb{R}^N$ then the following functions are kernels on $\mathcal{X}$

1. $K_1(x,t) + K(x,t)$
2. $aK_1(x,t)$
3. $K_1(x,t)K_2(x,t)$
4. $K(\phi(x), \phi(t))$

Polynomial of kernels

Let $F = p$ where $p : \mathbb{R}^q \to \mathbb{R}$ is a polynomial with nonnegative coefficients. By property 1 and 2 above we conclude that p is a valid function.

In particular if $q = 1$,

$$\sum_{i=1}^{d} a_i (K(x,t))^i$$

is kernel if $a_1, \ldots, a_d \geq 0$

Polynomial kernels

We now discuss kernels on $\mathbb{R}^n$. The above observation implies that if $p : \mathbb{R} \to \mathbb{R}$ is a univariate polynomial with nonnegative coefficients then $p(\mathbf{x}'\mathbf{t}), \mathbf{x}, \mathbf{t} \in \mathbb{R}^n$ is kernel on $\mathbb{R}^n$. In particular if $a \geq 0$ the following are valid polynomial kernels

- $(\mathbf{x}'\mathbf{t})^d$
- $(a + \mathbf{x}'\mathbf{t})^d$
- $\sum_{i=0}^{d} \frac{a^i}{i!}(\mathbf{x}'\mathbf{t})^i$

'Infinite polynomial' kernel

If in the last equation we set $d = \infty$ the series

$$\sum_{i=0}^{d} \frac{a^i}{i!}(\mathbf{x}'\mathbf{t})^i$$

converges everywhere uniform to $\exp(a\mathbf{x}'\mathbf{t})$ showing that this function is also a kernel.

Assume for simplicity that $n = 1$. A feature map corresponding to the kernel $\exp(axt)$ is

$$\phi(x) = \left(1,\ \sqrt{a}x,\ \sqrt{\frac{a}{2}}x^2,\ \sqrt{\frac{a^3}{6}}x^3, \ldots\right) = \left(\sqrt{\frac{a^i}{i!}}x^i : i \in \mathbb{N}\right)$$

- The feature space has an infinite number of dimensions!

Translation invariant and radial kernels

We say that a kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is

- *Translation invariant* if it has the form

$$K(\mathbf{x}, \mathbf{t}) = H(\mathbf{x} - \mathbf{t}), \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^n$$

  where $H : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function.

- *Radial* if it has the form

$$K(\mathbf{x}, \mathbf{t}) = h(\|\mathbf{x} - \mathbf{t}\|), \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^n$$

  where $h : [0, \infty) \to [0, \infty)$ is a differentiable function

The Gaussian kernel

An important example of radial kernel is the Gaussian

$$K(\mathbf{x}, \mathbf{t}) = \exp(-\beta \|\mathbf{x} - \mathbf{t}\|^2), \quad \beta > 0, \mathbf{x}, \mathbf{t} \in \mathbb{R}^n$$

It is a kernel because it is the product of two kernels

$$K(\mathbf{x}, \mathbf{t}) = \Big(\exp(-\beta(\mathbf{x}'\mathbf{x} + \mathbf{t}'\mathbf{t}))\Big) \exp(2\beta \mathbf{x}'\mathbf{t})$$

(We saw before that $\exp(2\beta \mathbf{x}'\mathbf{t})$ is a kernel. Clearly $\exp(-\beta(\mathbf{x}'\mathbf{x} + \mathbf{t}'\mathbf{t})$ is a kernel with one dimensional feature map $\phi(\mathbf{x}) = \exp(-\beta \mathbf{x}'\mathbf{x})$ is indeed a kernel.)

Exercise: Can you find a feature map representation for the Gaussian kernel?

Periodic kernels

These are a special case of translation invariant kernels

Take $\mathcal{X} = \mathbb{R}$ and $K(x, y) = H(x - y)$, where $H : \mathbb{R} \to \mathbb{R}$ is differentiable, even and $2\pi-$periodic. Since K is symmetric, H is even ($H(x) = H(-x)$) and its Fourier series consists of cosine's only:

$$H(x) = \sum_{n=0}^{\infty} a_n \cos(nx).$$

Then we have

$$K(x, y) = H(x - y) = a_0 + \sum_{n=1}^{\infty} a_n \sin(nx) \sin(ny) + \sum_{n=1}^{\infty} a_n \cos(nx) \cos(ny)$$

which, assuming $a_n \geq 0$, is of the form $\langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(y) \rangle$ with

$$\boldsymbol{\phi}(x) \equiv \left(\sqrt{a_0}, \sqrt{a_1} \sin(x), \sqrt{a_1} \cos(x), \sqrt{a_2} \sin(2x), \sqrt{a_2} \cos(2x), \ldots\right)$$

Remark: Again the feature space is infinite dimensional. We will se that if $g(x) = \langle \mathbf{w}, \boldsymbol{\phi}(x) \rangle$, $\|\mathbf{w}\|$ measures the smoothness of the function...

Bibliography

These lectures are based on Chapter 2 and 3 of:

- J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.