GI01/4C55: Supervised Learning

4. Regularization / Kernels

October 24, 2005

Massimiliano Pontil

Today's Plan

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

Bibliography: These lecture notes are available at:
http://www.cs.ucl.ac.uk/staff/M.Pontil/courses/index-SL05.htm
Lectures notes are in part based on chapters 2 and 3 of Shawe-Taylor and Cristianini

Linear interpolation

Problem: We wish to find a function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ which best interpolates a data set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \subseteq \mathbb{R}^d \times \mathbb{R}$

- If the data have been generated in the form $(\mathbf{x}, f(\mathbf{x}))$, the vectors $\mathbf{x}_i$ are linearly independent and $m = d$ then there is a unique interpolant whose parameter $\mathbf{w}$ solves

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

  where, recall, $\mathbf{y} = (y_1, \ldots, y_m)^\top$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^\top$

- Otherwise, this problem is *ill-posed*

Ill-posed problems

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))

Regularization theory provides a general framework to solve ill-posed problems

Ridge regression

We minimize the regularized (penalized) empirical error

$$E_\lambda(\mathbf{w}) := \sum_{i=1}^{m}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \sum_{\ell=1}^{d} w_\ell^2 \equiv (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^\top \mathbf{w}$$

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

Setting $\nabla E_\lambda(\mathbf{w}) = 0$, we obtain the modified normal equations

$$-2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda \mathbf{w} = 0 \qquad (1)$$

whose solution (called *regularized solution*) is

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1}\mathbf{X}^\top \mathbf{y} \qquad (2)$$

Singular valued decomposition (review)

Singular value decomposition (SVD) establishes that

$$\mathbf{X}^\top \mathbf{X} = \mathbf{U}\Lambda_d \mathbf{U}^\top, \qquad \mathbf{X}\mathbf{X}^\top = \mathbf{V}\Lambda_m \mathbf{V}^\top$$

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

$$\Lambda_d = diag(\lambda_1, \ldots, \lambda_t, \mathbf{0}_{d-t}), \qquad \Lambda_m = diag(\lambda_1, \ldots, \lambda_t, \mathbf{0}_{m-t}),$$

$t = rank(\mathbf{X}\mathbf{X}^\top) = rank(\mathbf{X}^\top\mathbf{X})$ and $\lambda_1 \geq \cdots \geq \lambda_t > 0$, $t \leq \min(m, d)$. Moreover, we have

$$\mathbf{X}^\top = \sum_{i=1}^{t} \sqrt{\lambda_i}\mathbf{u}_i \mathbf{v}_i^\top = \mathbf{U}\Sigma\mathbf{V}^\top$$

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

Generalized solution

When λ goes to zero $\mathbf{w}$ tends to the **generalized solution**

$$\mathbf{w}_0 := (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{y} \tag{3}$$

where $(\mathbf{X}^\top \mathbf{X})^+$ is the pseudoinverse of $\mathbf{X}^\top \mathbf{X}$

$$(\mathbf{X}^\top \mathbf{X})^+ = \sum_{i=1}^{t} \sigma_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top$$

- If $m \geq d$, typically $\mathbf{A} := \mathbf{X}^\top \mathbf{X}$ will be full rank, so $\mathbf{A}^+ = \mathbf{A}^{-1}$
- The generalized solution is **the** function which, among those which minimize $E(\mathbf{w})$ (infinitely many if $m < d$) has the smallest norm of its coefficients

Dual representation

We show that the regularized solution can be written as

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \quad \Rightarrow \quad f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i^\top \mathbf{x} \tag{4}$$

where the vector of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top$ is given by

$$\boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_m)^{-1} \mathbf{y} \tag{5}$$

- **Function representations:** we call the functional form (or representation) $f(\mathbf{x}) = \mathbf{w}^\top x$ the *primal form* and (4) the *dual form* (or representation)

The dual form is convenient when $d > m$

Dual representation (cont.)

Proof of eqs.(4),(5): We rewrite eq.(1) as

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i \quad (6) \quad \text{where}: \ \alpha_i = \frac{y_i - \mathbf{w}^\top \mathbf{x}_i}{\lambda} \ (7)$$

Consequently, we have that $\mathbf{w}^\top \mathbf{x} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i^\top \mathbf{x}$ proving eq.(4). Plugging eq.(6) in eq.(7) we obtain

$$\sum_{j=1}^{m} (\mathbf{x}_i^\top \mathbf{x}_j + \lambda \delta_{ij}) \alpha_j = y_i, \quad \text{that is}: \ (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_m)\boldsymbol{\alpha} = \mathbf{y}$$

from which eq.(5) follows

Computational considerations

Training time:

- Solving for $\mathbf{w}$ (eq.(2)) requires $O(d^3)$ operations while solving for $\boldsymbol{\alpha}$ (eq.(5)) requires $O(dm^2 + m^3)$ operations

If $m \ll d$ it is more efficient to use the dual representation

Running (testing) time:

- Computing g in the primal form requires $O(d)$ operations, while the dual form (eq.(4)) requires $O(md)$ operations

Sparse representations

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

- If k denotes the number of nonzero components of the input then computing $\mathbf{x}^\top \mathbf{t}$ requires at most $O(k)$ operations

- If $km \ll d$ (which implies $m, k \ll d$) the dual representation requires $O(km^2 + m^3)$ computations for training and $O(mk)$ for testing

Feature map

The above ideas can naturally be generalized to nonlinear function regression

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

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

Vector $\boldsymbol{\phi}(\mathbf{x})$ is called the *feature vector* and the space $\{\boldsymbol{\phi}(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d\}$ the *feature space*

The non-linear regression function has the primal representation

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

Computational considerations

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

Key observation: in the dual representation we don't need to know ϕ explicitly; we just need to know the inner product between any pair of feature vectors!

Example: $N = d^2$, $\phi(\mathbf{x}) = (x_i x_j)_{i,j=1}^d$. In this case we have $\langle \phi(\mathbf{x}), \phi(\mathbf{t}) \rangle = (\mathbf{x}^\top \mathbf{t})^2$ which requires only $O(d)$ computations whereas $\phi(\mathbf{x})$ requires $O(d^2)$ computations

Kernel vs. feature map

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

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

- Maybe for some feature map ϕ computing $K(\mathbf{x}, \mathbf{t})$ (but not $\phi(\mathbf{x}), \phi(\mathbf{t})$) is independent of N (only dependent of d)

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

$$K(\mathbf{x}, \mathbf{t}) = (\mathbf{x}^\top \mathbf{t})^r$$

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

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 as well

$$E_\lambda(\mathbf{w}) = \sum_{i=1}^{m} V(y_i, \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle) + \lambda \langle \mathbf{w}, \mathbf{w} \rangle, \quad \lambda > 0 \tag{8}$$

where $V : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function

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

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

This result is usually called the *Representer Theorem*

Representer theorem

Setting the derivative of E_λ wrt. $\mathbf{w}$ to zero we have

$$-\sum_{i=1}^{m} V'(y_i, \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle) \phi(\mathbf{x}_i) + 2\lambda \mathbf{w} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}_i) \tag{9}$$

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

$$\alpha_i = \frac{1}{2\lambda} V'(y_i, \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle) \tag{10}$$

Thus we conclude that

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

Some remarks

- Plugging eq.(9) in the rhs. of eq.(10) we obtain a set of equations for the coefficients α_i:

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

  When V is the square loss and $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$ we retrieve the linear eq.(5)

- Substituting eq.(9) in eq.(8) we obtain an objective function for the α's:

$$\sum_{i=1}^{m} V(y_i, (\mathbf{K}\boldsymbol{\alpha})_i) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}, \quad \text{where : } \mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{m}$$

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

General feature map

We can further generalize the above idea to infinite dimensional feature maps

$$\boldsymbol{\phi} : \mathbb{R}^d \to \mathcal{W}$$

with associated kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by

$$K(\mathbf{x}, \mathbf{t}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{t}) \rangle, \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^d \tag{11}$$

Here $\mathcal{W}$ is any Hilbert space, typically either $\mathcal{W} = \mathbb{R}^N$ or the space of square summable sequences,

$$\mathcal{W} = \ell_2 = \left\{ (z_i)_{i=1}^{\infty} : \sum_{i=1}^{\infty} z_i^2 < \infty \right\}$$

Redundancy of the feature map

Warning: The feature map is not unique! If ϕ generates K so does $\widehat{\phi} = \mathbf{U}\phi$ where $\mathbf{U}$ in an (any!) $N \times N$ orthogonal matrix. Even the dimension of ϕ is not unique!

Example: If $n = 2$, $K(\mathbf{x}, \mathbf{t}) = (\mathbf{x}^\top \mathbf{t})^2$ is generated by both $\phi(\mathbf{x}) = (x_1^2, x_2^2, x_1 x_2, x_2 x_1)$ and $\widehat{\phi}(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$.

Change of perspective

- Let us start directly with a kernel K and see when K can be expressed as inner product in some feature space (eq.(11))

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

Positive semidefinite kernel

Definition: A function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is **positive semidefinite** if it is symmetric and the matrix $(K(\mathbf{x}_i, \mathbf{x}_j) : i, j = 1, \ldots, m)$ is positive semidefinite for every $m \in \mathbb{N}$ and every $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$

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

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

for some feature map $\boldsymbol{\phi} : \mathbb{R}^d \to \mathcal{W}$ and Hilbert space $\mathcal{W}$

Positive definite kernel (cont.)

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

$$\sum_{i,j=1}^{m} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \sum_{i=1}^{m} c_i \boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{m} c_j \boldsymbol{\phi}(\mathbf{x}_j) \right\rangle = \left\| \sum_{i=1}^{m} c_i \boldsymbol{\phi}(\mathbf{x}_i) \right\|^2 \geq 0$$

for every choice of $m \in \mathbb{N}$, $x_i \in \mathbb{R}^d$ and $c_i \in R$, $i = 1, \ldots, m$

Note: the proof of '$\Rightarrow$' requires the notion of reproducing kernel Hilbert spaces. Informally, one can show that the linear span of the set of functions $\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathbb{R}^d\}$ can be made into a Hilbert space H_K with inner product induced by the definition $\langle K(\mathbf{x}, \cdot), K(\mathbf{t}, \cdot) \rangle := K(\mathbf{x}, \mathbf{t})$. In particular, the map $\boldsymbol{\phi} : \mathbb{R}^d \to H_K$ defined as $\boldsymbol{\phi}(\mathbf{x}) = K(\mathbf{x}, \cdot)$ is a feature map associated with K.

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 $K(\mathbf{x}, \mathbf{t}) = (\mathbf{x}^\top \mathbf{t})^d$ is a kernel. For which class of functions $p : \mathbb{R} \to \mathbb{R}$ is $p(\mathbf{x}^\top \mathbf{t})$ 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 an $d \times d$ psd matrix the function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by

$$K(\mathbf{x}, \mathbf{t}) = \mathbf{x}^\top \mathbf{A} \mathbf{t}$$

is a kernel

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

Alternatively, note that:

$$\sum_{i,j} c_i c_j \mathbf{x}_i^\top \mathbf{A} \mathbf{x}_j = \sum_{i,j} c_i c_j (\mathbf{R}^\top \mathbf{x}_i)^\top (\mathbf{R}^\top \mathbf{x}_j) = \| \sum_i c_i \mathbf{R}^\top \mathbf{x}_i \|^2 \geq 0$$

Kernel composition

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

$$\tilde{K}(\mathbf{x}, \mathbf{t}) = K(\phi(\mathbf{x}), \phi(\mathbf{t}))$$

is a kernel

Proof: By hypothesis, K is a kernel and so, for every $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$ the matrix $(K(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)) : i, j = 1, \ldots, m)$ is psd

In particular, the above example corresponds to $K(\mathbf{x}, \mathbf{t}) = \mathbf{x}^\top \mathbf{t}$ and $\phi(\mathbf{x}) = \mathbf{R}^\top \mathbf{x}$

Kernel construction (cont.)

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

$$F(K_1(\mathbf{x}, \mathbf{t}), \ldots, K_q(\mathbf{x}, \mathbf{t})), \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^d$$

a kernel?

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

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

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

Nonnegative 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 $q = n$ and $K_i(\mathbf{x}, \mathbf{t}) = x_i t_i$.

In particular, this implies that

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

Product of kernels

The pointwise product of two kernels K_1 and K_2

$$K(\mathbf{x}, \mathbf{t}) := K_1(\mathbf{x}, \mathbf{t}) K_2(\mathbf{x}, \mathbf{t}), \quad \mathbf{x}, \mathbf{t} \in \mathbb{R}^d$$

is a kernel

Proof: We need to show that if $\mathbf{A}$ and $\mathbf{B}$ are psd matrices, so is $\mathbf{C} = (A_{ij} B_{ij} : i, j = 1, \ldots, m)$ ($\mathbf{C}$ is also called the Schur product of $\mathbf{A}$ and $\mathbf{B}$). We write $\mathbf{A}$ and $\mathbf{B}$ in their singular value form, $\mathbf{A} = \mathbf{U}\Sigma\mathbf{U}^\top$, $\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}^\top$ where $\mathbf{U}, \mathbf{V}$ are orthogonal matrices and $\Sigma = diag(\sigma_1, \ldots, \sigma_m)$, $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$, $\sigma_i, \lambda_i \geq 0$. We have

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

Summary of kernel properties

The above results can be summarized as follows:

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

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

Polynomial of kernels

Let $F = p$ where $p : \mathbb{R}^q \to \mathbb{R}$ is a polynomial in q variables with nonnegative coefficients. By properties 1,2 and 3 above we conclude that p is a valid function

In particular if $q = 1$,

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

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

Polynomial kernels

The above observation implies that if $p : \mathbb{R} \to \mathbb{R}$ is a polynomial with nonnegative coefficients then $p(\mathbf{x}^\top \mathbf{t}), \mathbf{x}, \mathbf{t} \in \mathbb{R}^d$ is a kernel on $\mathbb{R}^d$. In particular if $a \geq 0$ the following are valid polynomial kernels

- $(\mathbf{x}^\top \mathbf{t})^r$
- $(a + \mathbf{x}^\top \mathbf{t})^r$
- $\sum_{i=0}^{d} \frac{a^i}{i!} (\mathbf{x}^\top \mathbf{t})^i$

'Infinite polynomial' kernel

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

$$\sum_{i=0}^{r} \frac{a^i}{i!} (\mathbf{x}^\top \mathbf{t})^i$$

converges everywhere uniformly to $\exp(a\mathbf{x}^\top \mathbf{t})$ showing that this function is also a kernel

Assume for simplicity that $d = 1$. 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 dimensionality!

Translation invariant and radial kernels

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

- *Translation invariant* if it has the form

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

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

- *Radial* if it has the form

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

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

The Gaussian kernel

An important example of a radial kernel is the Gaussian kernel

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

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

$$K(\mathbf{x}, \mathbf{t}) = \Big(\exp(-\beta(\mathbf{x}^\top \mathbf{x} + \mathbf{t}^\top \mathbf{t}))\Big) \exp(2\beta \mathbf{x}^\top \mathbf{t})$$

(We saw before that $\exp(2\beta \mathbf{x}^\top \mathbf{t})$ is a kernel. Clearly $\exp(-\beta(\mathbf{x}^\top \mathbf{x} + \mathbf{t}^\top \mathbf{t}))$ is a kernel with one-dimensional feature map $\phi(\mathbf{x}) = \exp(-\beta \mathbf{x}^\top \mathbf{x})$)

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

Periodic kernels

These are a special case of translation invariant kernels

Take $d = 1$ 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 and H is even ($H(x) = H(-x)$), its Fourier series consists of cosines 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. If $f(x) = \langle \mathbf{w}, \boldsymbol{\phi}(x) \rangle$, $\|\mathbf{w}\|$ measures the smoothness of the function.