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 \( f(x) = w^T x \) which best interpolates a data set \( S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathbb{R}^d \times \mathbb{R} \)

- If the data have been generated in the form \((x, f(x))\), the vectors \( x_i \) are linearly independent and \( m = d \) then there is a unique interpolant whose parameter \( w \) solves
  \[
  Xw = y
  \]
  where, recall, \( y = (y_1, \ldots, y_m)^T \) and \( X = [x_1, \ldots, x_m]^T \)

- 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

\[ E_\lambda(w) := \sum_{i=1}^{m} (y_i - w^T x_i)^2 + \lambda \sum_{\ell=1}^{d} w_{\ell}^2 \equiv (y - Xw)^T (y - Xw) + \lambda w^T w \]

The positive parameter \( \lambda \) defines a trade-off between the error on the data and the norm of the vector \( w \) (degree of regularization)

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

\[-2X^T (y - Xw) + 2\lambda w = 0 \quad (1)\]

whose solution (called regularized solution) is

\[ w = (X^T X + \lambda I_d)^{-1} X^T y \quad (2)\]

Singular valued decomposition (review)

Singular value decomposition (SVD) establishes that

\[ X^T X = U \Lambda_d U^T, \quad XX^T = V \Lambda_m V^T \]

where \( U \) and \( V \) are orthogonal matrices, \( U = [u_1, \ldots, u_d] \), \( V = [v_1, \ldots, v_m] \),

\[ \Lambda_d = diag(\lambda_1, \ldots, \lambda_t, 0_{d-t}), \quad \Lambda_m = diag(\lambda_1, \ldots, \lambda_t, 0_{m-t}), \]

\( t = rank(XX^T) = rank(X^T X) \) and \( \lambda_1 \geq \cdots \geq \lambda_t > 0 \), \( t \leq \min(m, d) \). Moreover, we have

\[ X^T = \sum_{i=1}^{t} \sqrt{\lambda_i} u_i v_i^T = U \Sigma V^T \]

where \( \Sigma \) is the \( d \times m \) matrix with leading diagonal entries \( \sigma_j = \sqrt{\lambda_j} \)
Generalized solution

When \( \lambda \) goes to zero \( w \) tends to the \textbf{generalized solution}

\[
    w_0 := (X^T X)^+ X^T y
\]

where \((X^T X)^+\) is the pseudoinverse of \(X^T X\)

\[
(X^T X)^+ = \sum_{i=1}^{t} \sigma_i^{-1} u_i u_i^T
\]

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

Dual representation

We show that the regularized solution can be written as

\[
    w = \sum_{i=1}^{m} \alpha_i x_i \quad \Rightarrow \quad f(x) = \sum_{i=1}^{m} \alpha_i x_i^T x
\]

where the vector of parameters \( \alpha = (\alpha_1, \ldots, \alpha_m)^T \) is given by

\[
    \alpha = (XX^T + \lambda I_m)^{-1} y
\]

- \textbf{Function representations}: we call the functional form (or representation) \( f(x) = w^T x \) the \textit{primal form} and (4) the \textit{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

\[ w = \sum_{i=1}^{m} \alpha_i x_i \quad (6) \]  
where: \[ \alpha_i = \frac{y_i - w^\top x_i}{\lambda} \quad (7) \]

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

\[ \sum_{j=1}^{m} (x_i^\top x_j + \lambda \delta_{ij}) \alpha_j = y_i, \quad \text{that is:} \quad (XX^\top + \lambda I_m)\alpha = y \]
from which eq.(5) follows.

Computational considerations

Training time:

- Solving for \( w \) (eq.(2)) requires \( O(d^3) \) operations while solving for \( \alpha \) (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 $x \in \mathbb{R}^d$ has most of its components equal to zero (e.g., 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 $x^T t$ 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 $\phi : \mathbb{R}^d \to \mathbb{R}^N$,

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

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

The non-linear regression function has the primal representation

$$f(x) = \langle w, \phi(x) \rangle := w^T \phi(x) = \sum_{j=1}^N w_j \phi_j(x)$$
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 $\phi$ explicitly; we just need to know the inner product between any pair of feature vectors!

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

---

Kernel vs. feature map

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

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

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

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

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

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

\[ E_\lambda(w) = \sum_{i=1}^{m} V(y_i, \langle w, \phi(x_i) \rangle) + \lambda \langle w, w \rangle, \quad \lambda > 0 \quad (8) \]

where \( V : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) is a loss function.

**Theorem:** If \( V \) is differentiable wrt. its second argument and \( w \) is a minimizer of \( E_\lambda \) then it has the form

\[ w = \sum_{i=1}^{m} \alpha_i \phi(x_i) \Rightarrow f(x) = \langle w, \phi(x) \rangle = \sum_{i=1}^{m} \alpha_i K(x_i, x) \]

This result is usually called the **Representer Theorem**.

---

**Representer theorem**

Setting the derivative of \( E_\lambda \) wrt. \( w \) to zero we have

\[ - \sum_{i=1}^{m} V'(y_i, \langle w, \phi(x_i) \rangle) \phi(x_i) + 2\lambda w = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i \phi(x_i) \quad (9) \]

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

\[ \alpha_i = \frac{1}{2\lambda} V'(y_i, \langle w, \phi(x_i) \rangle) \quad (10) \]

Thus we conclude that

\[ f(x) = \langle w, \phi(x) \rangle = \sum_{i=1}^{m} \alpha_i \langle \phi(x_i), \phi(x) \rangle = \sum_{i=1}^{m} \alpha_i K(x, x_i), \]
Some remarks

• Plugging eq.(9) in the rhs. of eq.(10) we obtain a set of equations for the coefficients $\alpha_i$:

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

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

• Substituting eq.(9) in eq.(8) we obtain an objective function for the $\alpha$'s:

$$\sum_{i=1}^{m} V(y_i, (K\alpha)_i) + \lambda \alpha^T K \alpha, \quad \text{where } K = (K(x_i, x_j))_{i,j=1}^{m}$$

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

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

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

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

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 $\phi$ generates $K$ so does $\tilde{\phi} = U\phi$ where $U$ in an (any!) $N \times N$ orthogonal matrix. Even the dimension of $\phi$ is not unique!

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

---

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 $K(x, t) = \langle \phi(x), \phi(t) \rangle$?
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 $(K(x_i, x_j) : i, j = 1, \ldots, m)$ is positive semidefinite for every $m \in \mathbb{N}$ and every $x_1, \ldots, x_m \in \mathbb{R}^d$

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

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

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

Positive definite kernel (cont.)

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

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

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

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 $\{K(x, \cdot) : x \in \mathbb{R}^d\}$ can be made into a Hilbert space $H_K$ with inner product induced by the definition $\langle K(x, \cdot), K(t, \cdot) \rangle := K(x, t)$. In particular, the map $\phi : \mathbb{R}^d \to H_K$ defined as $\phi(x) = K(x, \cdot)$ 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 $K(x, t) = (x^T t)^d$ is a kernel. For which class of functions $p : \mathbb{R} \rightarrow \mathbb{R}$ is $p(x^T t)$ a kernel? More generally, if $K$ is a kernel when is $p(K(x, t))$ a kernel?

General linear kernel

If $A$ is an $d \times d$ psd matrix the function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$K(x, t) = x^T At$$

is a kernel

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

Alternatively, note that:

$$\sum_{i,j} c_i c_j x_i^T Ax_j = \sum_{i,j} c_i c_j (R^T x_i)^T (R^T x_j) = \| \sum_i c_i R^T x_i \|^2 \geq 0$$
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

\[
\tilde{K}(x, t) = K(\phi(x), \phi(t))
\]

is a kernel

Proof: By hypothesis, \( K \) is a kernel and so, for every \( x_1, \ldots, x_m \in \mathbb{R}^d \) the matrix \( (K(\phi(x_i), \phi(x_j)) : i, j = 1, \ldots, m) \) is psd

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

25

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

\[
F(K_1(x, t), \ldots, K_q(x, t)), \quad x, t \in \mathbb{R}^d
\]
a kernel?

Equivalently: when for every choice of \( m \in \mathbb{N} \) and \( A_1, \ldots, A_q \) \( m \times m \) 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

26
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 $K_i(x, t) = x_i t_i$.

In particular, this implies that

- $a K_1$ 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 $K_2$

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

is a kernel.

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

$$\sum_{i,j=1}^{m} a_ia_j C_{ij} = \sum_{ij} a_ia_j \sum_{r} \sigma_r U_{ir} U_{jr} \sum_{s} \lambda_s V_{is} V_{js}$$

$$= \sum_{rs} \sigma_r \lambda_s \sum_{i} a_1 U_{ir} V_{is} \sum_{j} a_j U_{jr} V_{js}$$

$$= \sum_{rs} \sigma_r \lambda_s (\sum_{i} a_1 U_{ir} V_{is})^2 \geq 0$$
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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^N$ then the following functions are kernels on $\mathbb{R}^d$:

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

Polynomial of kernels

Let $F = p$ where $p : \mathbb{R}^q \rightarrow \mathbb{R}$ 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 $q = 1$,

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

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 \( p(x^T t), x, t \in \mathbb{R}^d \) is a kernel on \( \mathbb{R}^d \). In particular if \( a \geq 0 \) the following are valid polynomial kernels

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

31

'Infinite polynomial' kernel

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

\[
\sum_{i=0}^{\infty} \frac{a^i}{i!(x^T t)^i}
\]

converges everywhere uniformly to \( \exp(ax^T t) \) showing that this function is also a kernel.

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

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

- 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
  
  $$K(x, t) = H(x - t), \quad x, t \in \mathbb{R}^d$$
  
  where $H : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function

- **Radial** if it has the form
  
  $$K(x, t) = h(||x - t||), \quad x, t \in \mathbb{R}^d$$
  
  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

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

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

$$K(x, t) = \left(\exp(-\beta (x^T x + t^T t))\right) \exp(2\beta x^T t)$$

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

**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 $\langle \phi(x), \phi(y) \rangle$ with

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

Remark: Again the feature space is infinite-dimensional. If $f(x) = \langle w, \phi(x) \rangle$, $\|w\|$ measures the smoothness of the function.