Manifold Learning for Signal and Image Analysis
Lecture 5: A Brief Introduction to Kernel Methods

Radu Horaud
INRIA Grenoble Rhone-Alpes, France
Radu.Horaud@inria.fr
http://perception.inrialpes.fr/
Outline of Lecture 5

- Linear regression in "feature space"
- Kernel construction and characterization of the feature space.
- The kernel (Gram) matrix
- The covariance matrix in feature space
- Feature-space computations
- Kernel PCA
Material for This Lecture

- C. Bishop. Pattern Analysis and Machine Learning (chapters 6 and 12).
Consider a data set: \( X = [x_1, \ldots, x_n] \in \mathbb{R}^D \)

**Definition of a kernel function**: consider a nonlinear feature space mapping: \( \phi : x \rightarrow \phi(x) \), with \( \phi(x) \in \mathbb{R}^M \). A kernel satisfies:

\[
\kappa(x_i, x_j) = \phi(x_i)^\top \phi(x_j) = \langle \phi(x_i), \phi(x_j) \rangle
\]

The main principle of *kernel methods* is to interpret the kernel function as an *inner product* in feature space and to design algorithms without making explicit the function \( \phi \).

This extends many algorithms by making use of the *kernel trick* or *kernel substitution*.

For example, we can extend basic algorithms, such as PCA and LDA in feature space, namely *kernel PCA* and *kernel Fisher discriminant*, etc.
Replace the standard regression problem with:

\[ y = \sum_{m=1}^{M} w_m \phi_m(x) = w^\top \phi(x) \]

The parameters \( w_1, \ldots, w_M \) can be estimated from a training set of pairs \((y_j, x_j)\) by minimizing the following criterion:

\[
J(w) = \frac{1}{2} \left( \sum_{j=1}^{n} (w^\top \phi(x_j) - y_j)^2 + \lambda w^\top w \right)
\]
Least-square Solution

- By taking the derivatives of $J$ with respect to $w$ and setting them to zero, we obtain the following solution:

$$w = -\frac{1}{\lambda} \sum_{j=1}^{n} (w^\top \phi(x_j) - y_j) \phi(x_j)$$

- Let $a_j = -\frac{1}{\lambda} (w^\top \phi(x_j) - y_j)$ be the $j$-th entry of a vector $a \in \mathbb{R}^n$.
- Let $\Phi = [\phi(x_1) \ldots \phi(x_j) \ldots \phi(x_n)]$ be a $M \times n$ data matrix in feature space.
- Hence:

$$w = \Phi a$$

- We will use $a$ instead of $w$. 

Radu Horaud
Manifold Learning for Signal and Image Analysis; Lecture 5
Dual representation

- Substitute $w = \Phi a$ in $J(w)$. We obtain:

$$J(a) = \frac{1}{2} \left( a^\top \Phi^\top \Phi \Phi^\top \Phi a - a^\top \Phi^\top \Phi y + \lambda a^\top \Phi^\top \Phi a \right)$$

- The $n \times n$ matrix:

$$K = \Phi^\top \Phi$$

is a Gram matrix in feature space (it will be referred to as a kernel matrix), with entries:

$$\kappa_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = \kappa(x_i, x_j)$$
Solution in Feature Space

- We obtain a new expression for $J(w)$ as a function of vector $a$ and of the Gram matrix:

  $$J(a) = \frac{1}{2} \left( a^\top K^\top K a - a^\top K y + \lambda a^\top K a \right)$$

- The solution is obtained by setting the gradient of $J$ with respect to $a$ to zero:

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

  which always has an inverse.
The linear regression model allows to predict the output $y$ from a new input $x$:

$$y = \sum_{m=1}^{M} w_m \phi_m(x) = w^\top \phi(x)$$

By substitution this becomes:

$$y = a^\top \Phi^\top \phi(x) = \sum_{j=1}^{n} a_j \kappa(x_j, x)$$

The mapping $\phi$ is not required!
Discussion

- The dual representation allows the solution to be expressed entirely in terms of the kernel function;
- Inversion of a $M \times M$ matrix (dimension of the feature space) is replaced by inversion of a $n \times n$ matrix (number of points in the training set).
- It avoids computations in feature space when $M$ is very large.
- The feature-space is a vector space equipped with an inner-product – metric space;
- This means that there is a strong similarity between feature-space methods and MDS (only the pairwise inner-product between data points are needed to construct algorithms).
Constructing Kernels

- A valid kernel function is such that the associated Gram matrix is symmetric positive semidefinite.
- The simplest kernel corresponds to $\phi(x) = x$, or
  $$\kappa(x, x') = x^\top x'$$
- Valid kernels:
  - $c\kappa(x, x')$ with $c > 0$;
  - $f(x)\kappa(x, x')f(x')$;
  - $\exp(\kappa(x, x'))$;
  - $\kappa_1(x, x') + \kappa_2(x, x')$;
  - $\kappa_1(x, x')\kappa_2(x, x')$;
  - $\kappa(\phi(x), \phi(x'))$;
  - $x^\top A x'$ with $A \succeq 0$. 
Kernel Normalization

• $x \rightarrow \phi(x)/\|\phi(x)\|$ which yields:

$$\hat{\kappa}(x, x') = \frac{\kappa(x, x')}{\sqrt{\kappa(x, x)\kappa(x', x')}}$$

$$= \kappa(x, x)^{-1/2}\kappa(x, x')\kappa(x', x')^{-1/2}$$
The Gaussian Kernel

\[ \kappa(x, x') = \exp \left( -\frac{\|x - x'\|^2}{2\sigma^2} \right) \]

- \[ \|x - x'\|^2 = x^\top x + x'^\top x' - 2x^\top x' \]
- Let \[ f(x) = \exp(-x^\top x/2\sigma^2) = \frac{1}{\sqrt{\exp(x^\top x/\sigma^2)}} \]
- The Gaussian kernel writes:

\[
\kappa(x, x') = f(x) \exp(x^\top x'/\sigma^2) f(x') = \frac{\exp(x^\top x'/\sigma^2)}{\sqrt{\exp(x^\top x/\sigma^2) \exp(x'^\top x'/\sigma^2)}}
\]

- This is also known as the basis radial function (BRF) kernel.
Let $X$ be a compact subset of $\mathbb{R}^D$. Suppose that $k$ is a continuous and symmetric function such that the integral operator is positive

$$\int_{X \times X} \kappa(x, x') f(x) f(x') dx dx' \geq 0$$

for all $f \in L_2(X)$. An $L_2$ function is a function that is square integrable.

We can expand $\kappa(x, x')$ in a uniformly convergent series in terms of functions $\{\phi_i\}_{i=1}^\infty$ satisfying $\langle \phi_i, \phi_j \rangle = \delta_{ij}$

$$\kappa(x, x') = \sum_{i=1}^\infty \phi_i(x) \phi_i(x')$$
Inner-product Space

- A vector space is an inner-product space if there exists a real-valued symmetric bilinear map that satisfies:

  \[ \langle x, x \rangle \geq 0 \]

- The inner product is strict if: \( \langle x, x \rangle = 0 \) iff \( x = 0 \).

- A strict inner product allows to define a norm of a vector \( \|x\|_2 = \sqrt{\langle x, x \rangle} \) and an associated metric or distance \( \|x - x'\|_2 \).

- A vector space with a metric is known as a metric space.

- The feature space is a metric space, equipped with the strict inner product.
The Gram/Kernel Matrix

- The $n \times n$ matrix:

$$K = \Phi^\top \Phi$$

is a Gram matrix in feature space, with entries:

$$\kappa_{ij} = \langle \phi(x_i), \phi(x_j) \rangle = \kappa(x_i, x_j)$$

- We remind that $\Phi = [\phi(x_1) \ldots \phi(x_j) \ldots \phi(x_n)]$ is the $M \times n$ data matrix.

- It is symmetric, positive, semidefinite:

$$x^\top K x = x^\top \Phi^\top \Phi x = \|\Phi x\|_2^2$$

- This matrix was studied in Lecture #1 within the context of MDS. Here we have a generalization because each entry is a kernel function which is more general than the dot-product of MDS.
Spectral Decomposition of the Kernel Matrix

Let \((\lambda_1, v_1), \ldots, (\lambda_n, v_n)\) be the eigenvalue-eigenvector pairs of a Kernel matrix. It can be written as:

\[
K = \sum_{k=1}^{n} \lambda_k v_k v_k^\top
\]

Each matrix entry can be written as:

\[
\kappa_{ij} = \kappa(x_i, x_j) = \sum_{k=1}^{n} \lambda_k v_{ik} v_{jk} = \langle \phi(x_i), \phi(x_j) \rangle
\]

with \(\phi(x_i) = (\sqrt{\lambda_1} v_{i1}, \ldots, \sqrt{\lambda_k} v_{ik}, \ldots, \sqrt{\lambda_n} v_{in})^\top\).

Therefore, we can think of the eigenvectors as defining a feature space.
Feature-space Computations

- The norm of a feature-space vector:
  \[ \| \phi(x) \|_2^2 = \langle \phi(x), \phi(x) \rangle = \kappa(x,x) \]

- The norm of a linear combination:
  \[ \left\| \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|_2^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \kappa(x_i,x_j) \]

- Distance between two feature-space vectors:
  \[ \| \phi(x_i) - \phi(x_j) \|_2^2 = \kappa(x_i,x_i) + \kappa(x_j,x_j) - 2\kappa(x_i,x_j) \]
Center of Mass

- Notation: $\overline{\phi} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$
- There is no explicit dual representation for this point. Moreover, it is not the image of a "valid" data point.
- Norm, distance from a point, and expected distance:

\[
\|\overline{\phi}\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{n} \frac{1}{n} \kappa(x_i, x_j) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j)
\]

\[
\|\phi(x) - \overline{\phi}\|^2 = \langle \phi(x), \phi(x) \rangle + \langle \overline{\phi}, \overline{\phi} \rangle - 2 \langle \phi(x), \overline{\phi} \rangle
\]

\[
= \kappa(x, x) + \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa(x_i, x_j) - \frac{2}{n} \sum_{i=1}^{n} \kappa(x, x_i)
\]

\[
\frac{1}{n} \sum_{k=1}^{n} \|\phi(x_k) - \overline{\phi}\|^2 = \frac{1}{n} \sum_{k=1}^{n} \kappa(x_k, x_k) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(x_i, x_j)
\]
The Kernel Matrix of Centered Data

- In feature-space the centered data writes:

\[ \hat{\phi}(x) = \phi(x) - \bar{\phi} \]

- The corresponding entry of the associated kernel matrix writes:

\[ \hat{\kappa}(x, x') = \langle \phi(x) - \bar{\phi}, \phi(x') - \bar{\phi} \rangle \]

\[ = \kappa(x, x') - \frac{1}{n} \sum_{i=1}^{n} (\kappa(x, x_i) - \kappa(x', x_i)) + \frac{1}{n^2} \sum_{i,j=1}^{n} \kappa(x_i, x_j) \]

- In matrix form:

\[ \hat{\mathbf{K}} = \mathbf{K} - \frac{1}{n} \sum_{i=1}^{n} \left( \mathbf{1} \mathbf{1}^\top \mathbf{K} + \mathbf{K} \mathbf{1} \mathbf{1}^\top \right) + \frac{1}{n^2} (\mathbf{1}^\top \mathbf{1} \mathbf{K} \mathbf{1}) \mathbf{1} \mathbf{1}^\top \]
The Spread of the Data

- The $M \times n$ data matrix in feature space:
  $$\Phi = [\phi(x_1) \ldots \phi(x_j) \ldots \phi(x_n)]$$

- The covariance matrix for centered data is an $M \times M$ matrix:
  $$C = \frac{1}{n} \Phi \Phi^\top$$

- Each entry of this matrix is:
  $$c_{st} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)_s \phi(x_i)_t$$
The Projected Variance

- For centered data, the variance along a vector $\mathbf{v}$ writes:

$$\sigma^2_{\mathbf{v}} = \frac{1}{n} \mathbf{v}^\top \Phi \Phi^\top \mathbf{v}$$

- If the data are not centered:

$$\sigma^2_{\mathbf{v}} = \frac{1}{n} \mathbf{v}^\top \Phi \Phi^\top \mathbf{v} - \left(\frac{1}{n} \mathbf{v}^\top \Phi \mathbf{1}\right)^2$$
Let’s write $v$ as a combination of the feature-space points:

$$v = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \Phi \alpha.$$  

By substitution in the formula of the projected variance, we obtain:

$$\sigma_v^2 = \frac{1}{n} \alpha^\top \Phi^\top \Phi \Phi^\top \Phi \alpha - \left( \frac{1}{n} \alpha^\top \Phi^\top \Phi \mathbb{1} \right)^2$$

$$= \frac{1}{n} \alpha^\top K^2 \alpha - \left( \frac{1}{n} \alpha^\top K \mathbb{1} \right)^2$$
Eigendecomposition of Covariance and Kernel Matrices

- For centred data we have:
  \[
  C = \frac{1}{n} \Phi \Phi^\top
  \]
  with \( \{(\mu_i, u_i)\}_{i=1}^M \)

  \[
  K = \Phi^\top \Phi
  \]
  with \( \{(\lambda_i, v_i)\}_{i=1}^n \)

- By premultiplication of \( \Phi \Phi^\top u = n \mu u \) with \( \Phi^\top \) we obtain:
  \[
  v = \Phi^\top u \quad \text{and} \quad \lambda = n \mu
  \]

- From which we obtain: \( \|v\|^2 = u^\top \Phi \Phi^\top u = n \mu = \lambda \)

- The normalized eigenvector of the kernel matrix is:
  \[
  v = \lambda^{-1/2} \Phi^\top u
  \]

- There is a similar dual expression:
  \[
  u = \lambda^{-1/2} \Phi v
  \]
Traces

- The traces are related by:
  \[
  \text{tr}(C) = \frac{1}{n} \text{tr}(K)
  \]

- The trace of the kernel matrix:
  \[
  \text{tr}(K) = \sum_{i=1}^{n} \kappa(x_i, x_i)
  \]

- The total variance in feature-space:
  \[
  \sum_{i=1}^{M} \mu_i = \frac{1}{n} \sum_{i=1}^{n} \kappa(x_i, x_i)
  \]

- This can be used to estimate the dimension \( m \ll M \) of the reduced feature space.
Covariance Eigenvectors in Feature-space

- The eigenvectors of the covariance matrix:

\[
U = \begin{bmatrix}
\lambda_1^{-1/2} \Phi v_1 & \ldots & \lambda_k^{-1/2} \Phi v_k & \ldots \\
\end{bmatrix} = \Phi V \Lambda^{-1/2}
\]

- Each eigenvector:

\[
u_k = \lambda_k^{-1/2} \Phi v_k = \lambda_k^{-1/2} \sum_{i=1}^{n} v_{ik} \phi(x_i)
\]

- Let: \( \beta_k = \lambda_k^{-1/2} v_k = (\lambda_k^{-1/2} v_{1k} \ldots \lambda_k^{-1/2} v_{ik} \ldots \lambda_k^{-1/2} v_{nk}) \)

- Hence:

\[
u_k = \sum_{i=1}^{n} \beta_{ik} \phi(x_i)
\]
Let’s project a data point in feature space $\phi(x)$ onto an eigenvector of the covariance matrix:

$$u_k^\top \phi(x) = \langle u_k, \phi(x) \rangle$$

$$= \sum_{i=1}^{n} \beta_{ik} \langle \phi(x_i), \phi(x) \rangle$$

$$= \sum_{i=1}^{n} \beta_{ik} \kappa(x_i, x)$$

Let $U$ be the $M \times m$ matrix formed with $m \ll M$ eigenvectors of $C$. A feature point can be mapped in the eigenspace of $C$ with:

$$\tilde{\phi}(x) = U^\top \phi(x)$$
The Kernel PCA method

- Build the centered kernel matrix associated with a data set $X$ and a kernel $\kappa(x, x')$:

$$\hat{K} = K - \frac{1}{n} \sum_{i=1}^{n} \left( 1 1^\top K + K 1 1^\top \right) + \frac{1}{n^2} (1 1^\top K 1) 1 1^\top$$

- Compute the eigen-decomposition of this matrix and retain the $K$ largest eigenvalue-eigenvector pairs:

$$\hat{K} = V \Lambda V^\top$$

- Compute the vectors: $\beta_k = \lambda_k^{-1/2} v_k, \ k = 1 \ldots K$

- Project the feature-space data onto the space spanned by the eigenvectors of the feature-space covariance:

$$\tilde{\phi}(x) = U^\top \phi(x) \text{ with } u_k^\top \phi(x) = \sum_{i=1}^{n} \beta_{ik} \kappa(x_i, x)$$
Additional Topics of Interest

- Kernel K-means clustering
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.79.2967&rep=rep1&type=pdf

- Kernel Fisher discriminant analysis
  http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=788121

- Diffusion and exponential kernels (next lecture)