Introduction: Overview of Kernel Methods Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS Department of Statistical Science, Graduate University for Advanced Studies

April 18, 2008 / Statistical Learning Theory II



Basic idea of kernel methods

- Kernel Methods as Nonlinear Data Analysis
- Essence of kernel methodology

- Kernel PCA: Nonlinear extension of PCA
- Ridge regression and its kernelization

Basic idea of kernel methods Kernel Methods as Nonlinear Data Analysis Essence of kernel methodology

- Kernel PCA: Nonlinear extension of PCA
- Ridge regression and its kernelization

Kernel Methods as Nonlinear Data Analysis Essence of kernel methodology

Nonlinear Data Analysis I

Classical linear methods

• Data is expressed by a matrix.

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \cdots & X_1^m \\ X_2^1 & X_2^2 & \cdots & X_2^m \\ & & \vdots \\ X_N^1 & X_N^2 & \cdots & X_N^m \end{pmatrix}$$

(m dimensional, N data)

- Linear operations (matrix operations) are used for data analysis. *e.g.*
 - Principal component analysis (PCA)
 - Canonical correlation analysis (CCA)
 - Linear regression analysis
 - Fisher discriminant analysis (FDA)
 - Logistic regression, etc.

Kernel Methods as Nonlinear Data Analysis Essence of kernel methodology

Nonlinear transform can help.

Nonlinear Data Analysis II

Are linear methods sufficient?

Example 1: classification



 $(x_1, x_2) \mapsto (z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$

 Basic idea of kernel methods
 Kernel Methods as Nonlinear Data Analysis

 Some examples of kernel methods
 Essence of kernel methodology

Example 2: dependence of two data

Correlation

$$\rho_{XY} = \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X])^2]E[(Y - E[Y])^2]}}$$



 Transforming data to incorporate higher moments seems attractive.

Basic idea of kernel methods Kernel Methods as Nonlinear Data Analysis Essence of kernel methodology

- Kernel PCA: Nonlinear extension of PCA
- Ridge regression and its kernelization

Kernel Methods as Nonlinear Data Analysis Essence of kernel methodology

Feature space for transforming data I

• Kernel methodology = a systematic way of analyzing data by transforming them into a high-dimensional feature space.



Apply linear methods on the feature space.

- Which type of space serves as a feature space?
 - The space should incorporate various nonlinear information of the original data.
 - The inner product of the feature space is essential for data analysis (seen in the next subsection).

Computational problem of inner product

• For example, how about this?

 $(X,Y,Z)\mapsto (X,Y,Z,X^2,Y^2,Z^2,XY,YZ,ZX,\ldots).$

 But, for high-dimensional data, the above expansion makes the feature space very huge!

e.g. If X is 100 dimensional and the moments up to the third order are used, the dimensionality of feature space is

 $_{100}C_1 + _{100}C_2 + _{100}C_3 = 166750.$

 This causes a serious computational problem in working on the inner product of the feature space.
 We need a cleverer way of computing it. ⇒ Kernel method.

Inner product by positive definite kernel

• A positive definite kernel gives a painless computation of the inner product:

With special choice of the feature space, we have a function k(x,y) such that

 $\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j),$ positive definite kernel

where

$$\mathcal{X} \ni x \quad \mapsto \quad \Phi(x) \in \mathcal{H} \quad (\text{feature space}).$$

• Many linear methods use only the inner product without necessity of the explicit form of the vector $\Phi(X)$.

Basic idea of kernel methods

- Kernel Methods as Nonlinear Data Analysis
- Essence of kernel methodology

- Kernel PCA: Nonlinear extension of PCA
- Ridge regression and its kernelization

Review of PCA I

 X_1, \ldots, X_N : *m*-dimensional data.

Principal Component Analysis (PCA)

- Find *d*-directions to maximize the variance.
- Purpose: represent the structure of the data in a low dimensional space.



Review of PCA II

Sample variance in the direction of a (||a|| = 1) is $Var[a^T X] = a^T Va$, where V is the variance-covariance matrix:

$$V = \frac{1}{N} \sum_{i=1}^{N} (X_i - \frac{1}{N} \sum_{j=1}^{N} X_j) (X_i - \frac{1}{N} \sum_{j=1}^{N} X_j)^T.$$

Eigenvectors u_1, \ldots, u_m (with eigenvalues in descending order). The *p*-th principal axis = u_p . The *p*-th principal component of $X_i = u_p^T X_i$

Observation: PCA can be done if we can compute the inner product

- covariance matrix V,
- inner product between the unit eigenvector and the data.

Kernel PCA I

 X_1, \ldots, X_N : *m*-dimensional data.

Transform the data by a feature map Φ into a feature space \mathcal{H} :

```
X_1,\ldots,X_N \quad \mapsto \Phi(X_1),\ldots,\Phi(X_N)
```

Assume that the feature space has the inner product \langle , \rangle .

Apply PCA to the transformed data:

• Maximize the variance of the projection onto the unit vector f.

$$\max_{\|f\|=1} \operatorname{Var}[\langle f, \Phi(X) \rangle] = \max_{\|f\|=1} \frac{1}{N} \sum_{i=1}^{N} \left(\langle f, \Phi(X_i) \rangle - \frac{1}{N} \sum_{j=1}^{N} \langle f, \Phi(X_j) \rangle \right)^2$$

• Note: it suffices to use $f = \sum_{i=1}^{n} a_i \tilde{\Phi}(X_i)$, where

$$\tilde{\Phi}(X_i) = \Phi(X_i) - \frac{1}{N} \sum_{j=1}^{N} \Phi(X_j).$$

The direction orthogonal to Span{ $\tilde{\Phi}(X_1), \ldots, \tilde{\Phi}(X_N)$ } does not contribute.

Kernel PCA II

• The PCA solution:

 $\max a^T \tilde{K}^2 a \qquad \text{subject to} \quad a^T \tilde{K} a = 1,$

where \tilde{K} is $N \times N$ matrix with $\tilde{K}_{ij} = \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$. Note:

$$\frac{1}{N}\sum_{i=1}^{N}\langle f, \tilde{\Phi}(X_i)\rangle^2 = \frac{1}{N}\sum_{i=1}^{N}\langle \sum_{j=1}^{N}a_j\tilde{\Phi}(X_j), \tilde{\Phi}(X_i)\rangle^2 = \frac{1}{N}a^T\tilde{K}^2a,$$
$$\|f\|^2 = \langle \sum_{i=1}^{n}a_i\tilde{\Phi}(X_i), \sum_{i=1}^{n}a_i\tilde{\Phi}(X_i)\rangle = a^T\tilde{K}a.$$

• The first principal component of the data X_i is

$$\langle \tilde{\Phi}(X_i), \hat{f} \rangle = \sum_{i=1}^N \sqrt{\lambda_1} u_i^1,$$

where $\tilde{K} = \sum_{i=1}^{N} \lambda_i u^i u^{iT}$ is the eigen decomposition.

Kernel PCA III

Observation:

• PCA in the feature space can be done if we can compute $\langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$ or

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j).$$

• The principal direction is obtained in the form $f = \sum_i a_i \tilde{\Phi}(X_i)$, i.e., in the linear hull of the data.

Note:

$$\begin{split} \tilde{K}_{ij} &= \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle \\ &= \langle \Phi(X_i), \Phi(X_j) \rangle - \frac{1}{N} \sum_{b=1}^N \langle \Phi(X_i), \Phi(X_b) \rangle \\ &- \frac{1}{N} \sum_{a=1}^N \langle \Phi(X_a), \Phi(X_j) \rangle + \frac{1}{N^2} \sum_{a=1}^N \langle \Phi(X_a), \Phi(X_b) \rangle \\ &= k(X_i, X_j) - \frac{1}{N} \sum_{b=1}^N k(X_i, X_b) - \frac{1}{N} \sum_{a=1}^N k(X_a, X_j) + \frac{1}{N^2} \sum_{a=1}^N k(X_a, X_b) \end{split}$$

Basic idea of kernel methods

- Kernel Methods as Nonlinear Data Analysis
- Essence of kernel methodology

- Kernel PCA: Nonlinear extension of PCA
- Ridge regression and its kernelization

Review: Linear Regression I

Linear regression

- Data: $(X_1, Y_1), ..., (X_N, Y_N)$: data
 - X_i: explanatory variable, covariate (m-dimensional)
 - Y_i: response variable, (1 dimensional)
- Regression model: find the best linear relation

$$Y_i = a^T X_i + \varepsilon_i$$



Kernel PCA: Nonlinear extension of PCA Ridge regression and its kernelization

Review: Linear Regression II

Least square method:

$$\min_{a} \sum_{i=1}^{N} \|Y_i - a^T X_i\|^2.$$

Matrix expression

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \cdots & X_1^m \\ X_2^1 & X_2^2 & \cdots & X_2^m \\ & & \vdots \\ X_N^1 & X_N^2 & \cdots & X_N^m \end{pmatrix}, \qquad Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}.$$

• Solution:

$$\widehat{a} = (X^T X)^{-1} X^T Y$$

For a general x,

$$\widehat{y} = \widehat{a}^T x = Y^T X (X^T X)^{-1} x.$$

Ridge Regression

Ridge regression:

Find a linear relation by

$$\min_{a} \sum_{i=1}^{N} \|Y_i - a^T X_i\|^2 + \lambda \|a\|^2.$$

 λ : regularization coefficient.

Solution

$$\widehat{a} = (X^T X + \lambda I_N)^{-1} X^T Y$$

For a general x,

$$\widehat{y}(x) = \widehat{a}^T x = Y^T X (X^T X + \lambda I_N)^{-1} x.$$

• Ridge regression is useful when $(X^T X)^{-1}$ does not exist, or inversion is numerically unstable.

Kernelization of Ridge Regression I

 $(X_1, Y_1) \dots, (X_N, Y_N)$ (Y_i : 1-dimensional)

Transform X_i by a feature map Φ into a feature space \mathcal{H} :

$$X_1,\ldots,X_N \quad \mapsto \Phi(X_1),\ldots,\Phi(X_N)$$

Assume that the feature space has the inner product \langle , \rangle .

Apply ridge regression to the transformed data:

• Find the vector *f* such that

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{N} |Y_i - \langle f, \Phi(X_i) \rangle_{\mathcal{H}}|^2 + \lambda ||f||_{\mathcal{H}}^2.$$

• Similarly to kernel PCA, we can assume $f = \sum_{j=1}^{n} c_j \Phi(X_j)$.

$$\min_{c} \sum_{i=1}^{N} |Y_{i} - \langle \sum_{j=1}^{N} c_{j} \Phi(X_{j}), \Phi(X_{i}) \rangle_{\mathcal{H}}|^{2} + \lambda \| \sum_{j=1}^{N} c_{j} \Phi(X_{j}) \|_{\mathcal{H}}^{2}$$

Kernel PCA: Nonlinear extension of PCA Ridge regression and its kernelization

Kernelization of Ridge Regression II

Solution:

 $\widehat{c} = (K + \lambda I_N)^{-1} Y,$ where $K_{ij} = \langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}} = k(X_i, X_j).$

For a general x,

$$\widehat{y}(x) = \langle \widehat{f}, \Phi(x) \rangle_{\mathcal{H}} = \langle \sum_{j} \widehat{c}_{j} \Phi(X_{j}), \Phi(x) \rangle_{\mathcal{H}} = Y^{T} (K + \lambda I_{N})^{-1} \mathbf{k},$$

where

$$\mathbf{k} = \begin{pmatrix} \langle \Phi(X_1), \Phi(x) \rangle \\ \vdots \\ \langle \Phi(X_N), \Phi(x) \rangle \end{pmatrix} = \begin{pmatrix} k(X_1, x) \\ \vdots \\ k(X_N, x) \end{pmatrix}.$$

<ロ> < (回) < (回) < (回) < (回) < (回) < (回) < (0) < (0) </p>
22/24

Kernel PCA: Nonlinear extension of PCA Ridge regression and its kernelization

Kernelization of Ridge Regression III

Proof.

Matrix expression gives

$$\sum_{i=1}^{N} |Y_i - \langle \sum_{j=1}^{N} c_j \Phi(X_j), \Phi(X_i) \rangle_{\mathcal{H}}|^2 + \lambda \| \sum_{j=1}^{N} c_j \Phi(X_j) \|_{\mathcal{H}}^2$$
$$= (Y - Kc)^T (Y - Kc) + \lambda c^T Kc$$
$$= c^T (K^2 + \lambda K) c - 2Y^T Kc + Y^T Y.$$

It follows that the optimal c is given by

$$\widehat{c} = (K + \lambda I_N)^{-1} Y.$$

Inserting this to $\widehat{y}(x) = \langle \sum_j \widehat{c}_j \Phi(X_j), \Phi(x) \rangle_{\mathcal{H}}$, we have the claim.

Kernel PCA: Nonlinear extension of PCA Ridge regression and its kernelization

Kernelization of Ridge Regression IV

Observation:

• Ridge regression in the feature space can be done if we can compute the inner product

 $\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j).$

• The resulting coefficient is of the form $f = \sum_i c_i \Phi(X_i)$, i.e., in the linear hull of the data. The orthogonal directions do not contribute to the objective function.