Methods with Kernels

Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

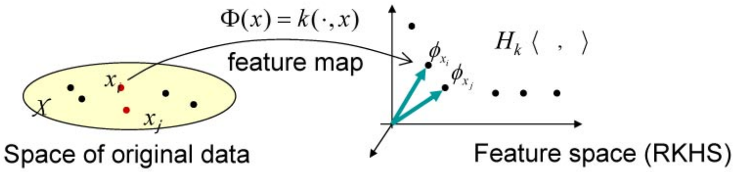
Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

May 2, 2008 / Statistical Learning Theory II

Outline

Kernel methodology: feature space by RKHS

- Kernel methodology = Data analysis by transforming data into a high-dimensional feature space given by RKHS.



Apply linear methods on RKHS.

The computation of the inner product is cheap.

Higher-order statistics by positive definite kernel

- A nonlinear kernel can include higher-order statistics.

  Example: Polynomial kernel on $\mathbb{R}$: $k(y,x) = (yx+1)^d$.

  - Data are transformed as $k(\cdot, X_1), \ldots, k(\cdot, X_N) \in \mathcal{H}_k$.
  - Regarding $k(\cdot, X) = k(y, X)$ as a function of y,

  $$k(y,X) = X^d y^d + a_{d-1} X^{d-1} y^{d-1} + \cdots + a_1 X y + a_0 \qquad (a_i \neq 0).$$

  - $\{1, y, y^2, \ldots, y^d\}$ is a basis of $\mathcal{H}_k$.
  - With respect to this basis, the component of the feature vector $k(\cdot, X)$ is

  $$(X^d, a_{d-1}X^{d-1}, \ldots, a_1 X, a_0)^T.$$

  This includes the statistics $(X, X^2, \ldots, X^d)$.

- Similar nonlinear statistics appear in other kernels such as Gaussian, Lapacian, etc.

Kernel PCA I

- $X_1, \ldots, X_N$: data on $\mathcal{X}$.
- $k : \mathcal{X} \times \mathcal{X}$ positive definite kernel, $\mathcal{H}_k$: RKHS.
- Transform the data into $\mathcal{H}_k$ by $\Phi(x) = k(\cdot, x)$:

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

Kernel PCA ([SSM98]): Apply PCA on $\mathcal{H}_k$:

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

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

- 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 $\{\tilde{\Phi}(X_1), \ldots, \tilde{\Phi}(X_N)\}$ 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 $\tilde{K}_{ij} = \langle \tilde{\Phi}(X_i), \tilde{\Phi}(X_j) \rangle$.

$$\tilde{K} = k(X_i, X_j) - \frac{1}{N}\sum_{b=1}^{N} k(X_i, X_b) - \frac{1}{N}\sum_{a=1}^{N} k(X_a, X_j) + \frac{1}{N^2}\sum_{a,b=1}^{N} k(X_a, X_b).$$

$\tilde{K}$ is called a centered Gram matrix.

Note:

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

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

Kernel PCA III

- The p-th principal direction $f^{(p)} = \sum_{i=1}^{N} \alpha_i^{(p)} \tilde{\Phi}(X_i)$ is given by

$$\max \alpha^{(p)T} \tilde{K}^2 \alpha^{(p)} \quad \text{subj. to} \quad \begin{cases} \alpha^{(p)} \tilde{K} \alpha^{(p)} = 1 \\ \alpha^{(p)} \tilde{K} \alpha^{(a)} = 0 \quad (a = 1, \ldots, p-1). \end{cases}$$

Principal component of kernel PCA

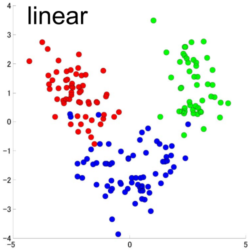
Let $\tilde{K} = \sum_{p=1}^{N} \lambda_p u^{(p)} u^{(p)T}$ is the eigen decomposition $(\lambda_1 \geq \cdots \geq \lambda_N \geq 0)$.
The p-th principal component of the data X_i is

$$\langle \tilde{\Phi}(X_i), \sum_{j=1}^{N} \alpha_j^{(p)} \Phi(\tilde{X}_j) \rangle = \sum_{j=1}^{N} \sqrt{\lambda_1} u_i^{(p)},$$

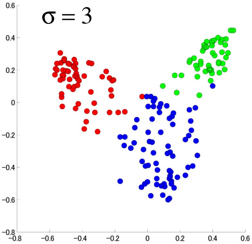
Kernel PCA: numerical examples

- Wine data (from UCI repository [MA94]).
- 178 data of 13 dimension. They represents chemical measurements of different wine.
- There are three classes, which correspond to types of wine.
- The classes are shown in different colors, but not used for the analysis.

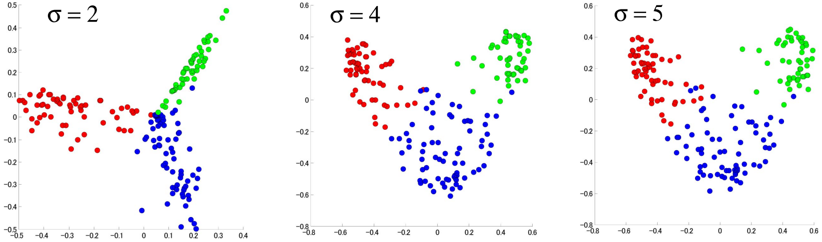
Linear PCA



KPCA Gaussian kernel



KPCA with Gaussian kernels. $k(x,y) = \exp\{-\frac{1}{2\sigma^2}\|x-y\|^2\}$.



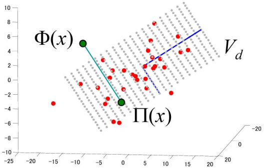
Application to noise reduction

Noise reduction by kernel PCA.

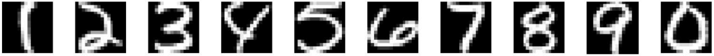
- $X_1, \ldots, X_N$: data, $\mapsto \Phi(X_1), \ldots, \Phi(X_N)$: data in RKHS.
- V_d: subspace of $\mathcal{H}_k$ spanned by $f^{(1)}, \ldots, f^{(d)}$ (d major principle directions).
- $\Pi(x)$ ($\in \mathcal{H}_k$): orthogonal projection of $\Phi(x)$ onto V_d.
- Find a point y in the original space such that

$$y = \arg\min_{y \in \mathcal{X}} \|\Phi(y) - \Pi(x)\|_{\mathcal{H}_k}.$$

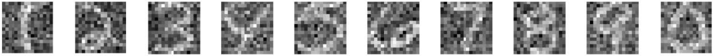
Note: $\Pi(x)$ is not necessarily in the image of embedding Φ.



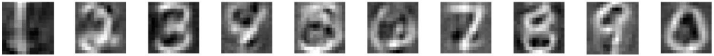
USPS hand-written digits data:
7191 images of hand-written digits of 16×16 pixels.



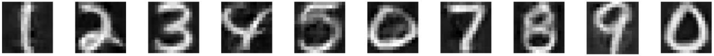
Sample of original images (not used for experiments)



Sample of noisy images



Sample of denoised images (linear PCA)



Sample of denoised images (kernel PCA, Gaussian kernel)

Properties of kernel PCA

- Nonlinear features can be considered.
- The results depend on the choice of kernel and kernel parameters. Interpreting the results may be difficult.
- Can be used for a preprocessing of other analysis like classification. (Dimension reduction / feature extraction)
- How to choose a kernel and kernel parameter?
  - Cross-validation may be possible, in general.
  - If it is a preprocessing, the performance of the final analysis should be maximized.

Fisher discriminant analysis I

Fisher's linear discriminant analysis

- Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: explanatory variable, covariate (m-dimensional)
  - $Y_i \in \{+1, -1\}$ binary,
- Linear discriminant function

$$f(x) = sgn\big(w^T x + b\big)$$

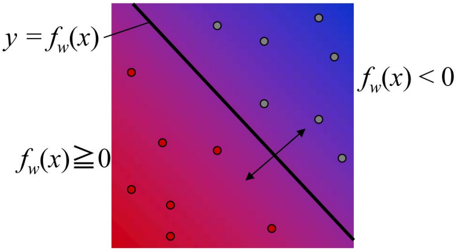
- Criterion: maximize the quotient

$$J(w) = \frac{\text{Between-class scatter along } w}{\text{Within-class scatter along } w} = \frac{w^T S_B w}{w^T S_W w},$$

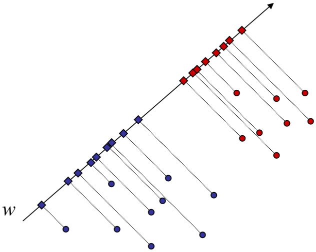
  - $S_B = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T$,
  - $S_W =$ $\sum_{i:Y_i=+1}(X_i - \mu_+)(X_i - \mu_+)^T + \sum_{j:Y_j=-1}(X_j - \mu_-)(X_j - \mu_-)^T$.
  - $\mu_+ = \frac{1}{N_+}\sum_{i:Y_i=+1} X_i$, $\mu_- = \frac{1}{N_-}\sum_{j:Y_j=-1} X_j$.

Fisher discriminant analysis II

Linear discriminant function



Fisher LDA



Fisher discriminant analysis III

- The maximization

$$\max_{w \neq 0} \frac{w^T S_B w}{w^T S_W w}$$

can be solved as a generalized eigenproblem.

- If the discriminant function is needed, a possible choice of b is

$$f(\mu_+) = -f(\mu_-) \quad \Rightarrow \quad b = \frac{1}{2} w^T (\mu_+ + \mu_-).$$

Kernel Fisher Discriminant Analysis I

Kernel Fisher discriminant analysis (kernel FDA, [MRW+99])

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: arbitrary covariate
  - $Y_i \in \{+1, -1\}$ binary,
- Embedding: $X_1, \ldots, X_N \quad \mapsto \quad \Phi(X_1), \ldots, \Phi(X_N) \in \mathcal{H}_k$, where $\Phi(x) = k(\cdot, x)$.
- Linear discriminant function on RKHS

$$f(x) = sgn\big(\langle h, \Phi(x)\rangle + b\big) = sgn(h(x) + b).$$

- Criterion:

$$J^{\Phi}(h) = \frac{\text{Between-class scatter along } h \text{ in } \mathcal{H}_k}{\text{Within-class scatter along } h \text{ in } \mathcal{H}_k}.$$

Kernel Fisher Discriminant Analysis II

- Between-class scatter:

$$\langle h, \mu_+^\Phi - \mu_-^\Phi \rangle^2$$

- Within-class scatter:

$$\sum_{i:Y_i=+1} \langle h, \Phi(X_i) - \mu_+^\Phi \rangle^2 + \sum_{j:Y_j=-1} \langle h, \Phi(X_j) - \mu_-^\Phi \rangle^2,$$

  where μ_+^Φ, μ_-^Φ are the mean of $\{\Phi(X_i) \mid Y_i = +1\}$, $\{\Phi(X_j) \mid Y_j = -1\}$, i.e.,

$$\mu_+^\Phi = \frac{1}{N_+} \sum_{i:Y_i=+1} \Phi(X_i), \quad \mu_-^\Phi = \frac{1}{N_-} \sum_{j:Y_j=-1} \Phi(X_j).$$

- It suffices to assume $h = \sum_{i=1}^N \alpha_i \Phi(X_i)$, because the orthogonal direction does not contribute to $J^\Phi(h)$.

Kernel Fisher Discriminant Analysis III

- Between-class:

$$\langle h, \mu_\pm^\Phi \rangle = \frac{1}{N_\pm} \sum_{t=1}^N \sum_{i:Y_i=\pm1} \alpha_t \langle \Phi(X_t), \Phi(X_i) \rangle = m_\pm^T \alpha,$$

where $\quad (m_\pm)_t = \frac{1}{N_\pm} \sum_{i:Y_i=\pm1} k(X_i, X_t) \quad \in \mathbb{R}^N.$

Scatter: $\langle h, \mu_+^\Phi - \mu_-^\Phi \rangle^2 = \alpha^T S_B^\Phi \alpha,$

$$S_B^\Phi := (m_+ - m_-)(m_+ - m_-)^T \quad (N \times N \text{ matrix}).$$

- Within-class:

$$\sum_{i:Y_i=\pm1} \langle h, \Phi(X_i) - \mu_\pm^\Phi \rangle = (I_{N_\pm} - \tfrac{1}{N_\pm} J_{N_\pm}) K_\pm \alpha,$$

where $\quad (K_\pm)_{it} = k(X_i, X_t) \quad (i : Y_i = \pm1, t = 1, \ldots, N),$

I_N is the unit matrix, and J_N is the $N \times N$ matrix with all entries 1.

Scatter: $\quad \alpha^T S_W^\Phi \alpha,$

$$S_W^\Phi = K_+^T (I_{N_+} - \tfrac{1}{N_+} J_{N_+}) K_+ + K_-^T (I_{N_-} - \tfrac{1}{N_-} J_{N_+}) K_- \qquad (N \times N \text{ matrix}).$$

Kernel Fisher Discriminant Analysis IV

- Objective function

$$J^{\Phi}(\alpha) = \frac{\alpha^T S_B^{\Phi} \alpha}{\alpha^T S_W^{\Phi} \alpha}.$$

- Regularization:
    - Maximizing $J^{\Phi}(\alpha)$ is ill-posed. The matrix S_W^{Φ} is of low rank!.
    - Use Tikhonov-type regularization

$$\tilde{J}^{\Phi}(\alpha) := \frac{\alpha^T S_B^{\Phi} \alpha}{\alpha(S_W^{\Phi} + \lambda I_N)\alpha}$$

(λ: regularization coefficient.)

- $\max_{\alpha} \tilde{J}^{\Phi}(\alpha)$ is solved as a generalized eigenproblem.
- The discriminant function is given by

$$f(x) = sgn\big(\langle h, \Phi(x)\rangle + b\big) = sgn\big(\textstyle\sum_t k(x, X_t)\alpha_t + b\big).$$

Large margin classifier in $\mathbb{R}^m$

Linear support vector machine (in $\mathbb{R}^m$)

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: explanatory variable (m-dimensional)
  - $Y_i \in \{+1, -1\}$ binary,
- Linear classifier

$$f(x) = sgn\left(w^T x + b\right)$$

- Large margin criterion:
  Assumption: the data is linearly separable.

  Among infinite number of separating hyperplanes, choose the one to give the largest margin.
  - Margin = distance of two classes measured along the direction of w.
  - The classifying hyperplane is the middle of the margin.

Large margin classifier in $\mathbb{R}^m$ II

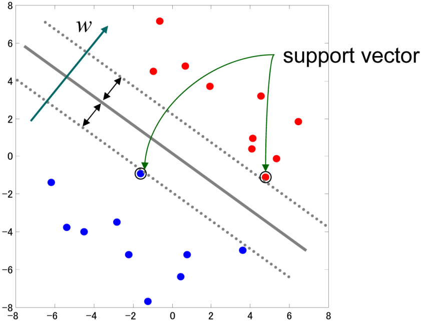
To fix a scale, assume

$$\begin{cases} \min(w^T X_i + b) = 1 & i : Y_i = +1 \\ \max(w^T X_i + b) = -1 & i : Y_i = -1 \end{cases}$$

Then,

$$\text{Margin} = \frac{2}{\|w\|}$$

The vectors attaining the minimum and maximum are called support vectors.



Large margin classifier in $\mathbb{R}^m$ III

- Large margin linear classifier

$$\max \frac{1}{\|w\|} \qquad \text{subj. to} \begin{cases} w^T X_i + b \geq 1 & \text{if } Y_i = +1, \\ w^T X_i + b \leq -1 & \text{if } Y_i = -1. \end{cases}$$

Equivalently,

Linear support vector machine (hard margin)

$$\min_{w,b} \|w\|^2 \qquad \text{subject to} \qquad Y_i(w^T X_i + b) \geq 1 \quad (\forall i).$$

- Quadratic objective function with linear constraints $\Longrightarrow$ free from local minima!
- This optimization can be numerically solved with the standard quadratic programming (QP, discussed later). Software packages are available.

SVM with soft margin

Relax the separability assumption. The linear separability is too restrictive in practice.

- Hard constraint: $Y_i(w^T X_i + b) \geq 1$
- Soft constraint: $Y_i(w^T X_i + b) \geq 1 - \xi_i \quad (\xi_i \geq 0)$

Linear support vector machine (soft margin)

$$\min_{w,b,\xi_i} \|w\|^2 + C\sum_{i=1}^N \xi_i \qquad \text{subj. to} \qquad \begin{cases} Y_i(w^T X_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- The optimization is still QP.
- C is a hyper-parameter, which we have to decide.

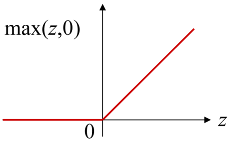
Soft margin as regularization

- Soft margin linear SVM is equivalent to the following regularization problem ($\lambda = 1/C$):

$$\min_{w,b} \sum_{i=1}^{N} \left(1 - Y_i(w^T X_i + b)\right)_+ + \lambda \|w\|^2$$

where

$$(z)_+ = \max(z, 0)$$



- $\ell(f(x), y) = (1 - yf(x))_+$ is called the soft margin loss function.

Tikhonov Regularization

General theory of regularization

- When the solution of the optimization

$$\min_{\alpha \in A} \Omega(\alpha)$$

  ($A \subset \mathcal{H}$) is not unique or stable, a regularization technique is often used.
- Tikhonov regularization: add a regularization term (or penalty term), e.g.,

$$\min_{\alpha \in A} \quad \Omega(\alpha) + \lambda \|\alpha\|^2.$$

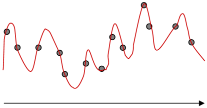
  $\lambda > 0$: regularization coefficient.
- The solution is often unique and stable.
- Other regularization terms, such as $\|\alpha\|$, are also possible, but differentiability may be lost.

Tikhonov Regularization II

- Example
  - Ill-posed problem:

$$\min_f (Y_i - f(X_i))^2.$$

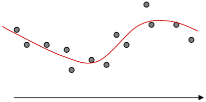
    Many f give zero error, if f is taken from a large space.



  - Regularized objective function

$$\min_f (Y_i - f(X_i))^2 + \lambda \|f\|^2 \qquad \text{(ridge regression)}$$

    finds a unique solution, which is often smoother.



SVM with kernels I

Kernelization of linear SVM

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: arbitrary covariate taking values in $\mathcal{X}$,
  - $Y_i \in \{+1, -1\}$ binary,
- k: positive definite kernel on $\mathcal{X}$. $\mathcal{H}$: associated RKHS.
- $\Phi(X_i) = k(\cdot, X_i)$: transformed data in $\mathcal{H}$.
- Large margin linear classifier on RKHS

$$f(x) = sgn\big(\langle h, \Phi(x)\rangle_{\mathcal{H}} + b\big) = sgn(h(x) + b).$$

Objective function (soft margin):

$$\min_{h,b,\xi_i} \|h\|^2_{\mathcal{H}} + C\sum_{i=1}^N \xi_i \quad \text{subj. to} \quad \begin{cases} Y_i(\langle h, \Phi(X_i)\rangle + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{cases}$$

or equivalently

$$\min_{h,b} \sum_{i=1}^N \big(1 - Y_i(\langle h, \Phi(X_i)\rangle + b)\big)_+ + \lambda\|h\|^2$$

SVM with kernels II

- It suffices to assume $h = \sum_{i=1}^{N} c_i \Phi(X_i)$, because the orthogonal direction only increases the regularization term without changing the first term of

$$\min_{h,b} \sum_{i=1}^{N} \big(1 - Y_i(\langle h, \Phi(X_i)\rangle + b)\big)_+ + \lambda \|h\|^2.$$

- In this case,

$$\|h\|^2 = \sum_{i,j=1}^{N} c_i c_j k(X_i, X_j),$$

$$\langle h, \Phi(X_i)\rangle = \sum_{j=1}^{N} c_j k(X_i, X_j).$$

SVM with kernels III

In summary,

SVM with kernel

$$\min_{c_i, b, \xi_i} \sum_{i,j=1}^{N} c_i c_j k(X_i, X_j) + C \sum_{i=1}^{N} \xi_i,$$

$$\text{subj. to} \quad \begin{cases} Y_i(\sum_{j=1}^{N} k(X_i, X_j) c_j + b) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}$$

- The optimization is numerically solved with QP.
- The dual form is simpler to solve (discussed later.)
- The parameter C and the kernel are often chosen by cross-validation.

Demonstration of SVM

Webpages for SVM Java applet

- `http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml`
- `http://www.eee.metu.edu.tr/~alatan/Courses/Demo/AppletSVM.html`

Mini-summary on SVM

- Kernel trick (a common property of kernel methods):
  - linear classifier on RKHS.
  - The computation of inner product is easy.
- Large margin criterion
  - May not be the Bayes optimal, but causes other good properties.
- Quadratic programming:
  - The objective function is solved by the standard quadratic programming.
- Sparse representation:
  - The classifier is represented by a small number of support vectors (discussed later).
- Regularization:
  - The soft margin objective function is equivalent to the margin loss with regularization.

Canonical correlation analysis I

Canonical correlation analysis (CCA)

- Linear dependence of two multivariate.
  - Data $(X_1, Y_1), \ldots, (X_N, Y_N)$
  - X_i: m-dimensional, Y_i: ℓ-dimensional.
- Find the directions a and b so that the correlation between the projections of X onto a and that of Y onto b is maximized:

$$\rho = \max_{a \in \mathbb{R}^m, b \in \mathbb{R}^\ell} \frac{\mathrm{Cov}[a^T X, b^T Y]}{\sqrt{\mathrm{Var}[a^T X]\mathrm{Var}[b^T Y]}} = \max_{a \in \mathbb{R}^m, b \in \mathbb{R}^\ell} \frac{a^T \widehat{V}_{XY} b}{\sqrt{a^T \widehat{V}_{XX} a}\sqrt{b^T \widehat{V}_{YY} b}},$$

where $\widehat{V}_{XX}$, $\widehat{V}_{YY}$, and $\widehat{V}_{XY}$ are the sample variance (covariance) matrices.

Canonical correlation analysis I

- Optimization:

$$\max a^T \widehat{V}_{XY} b \quad \text{subject to } a^T \widehat{V}_{XX} a = b^T \widehat{V}_{YY} b = 1.$$

- Lagrange multiplier:

$$\max a^T \widehat{V}_{XY} b + \frac{\mu}{2}(a^T \widehat{V}_{XX} a - 1) + \frac{\nu}{2}(b^T \widehat{V}_{YY} b - 1).$$

(μ, ν: Lagrange multiplier).

- Solution is obtained by the generalized eigenproblem:

$$\begin{pmatrix} O & \widehat{V}_{XY} \\ \widehat{V}_{YX} & O \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho \begin{pmatrix} \widehat{V}_{XX} & O \\ O & \widehat{V}_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

($\mu = \nu$ is derived. Set $\rho = -\mu = -\nu$.)

Kernel CCA I

Kernel CCA: kernelization of CCA ([Aka01, MRB01, BJ02]).

- Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$.
  - X_i, Y_i: arbitrary variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ (resp.).
- Embedding: prepare kernels $k_{\mathcal{X}}$ on $\mathcal{X}$ and $k_{\mathcal{Y}}$ on $\mathcal{Y}$.
  $$X_1, \ldots, X_N \quad \mapsto \quad \Phi_{\mathcal{X}}(X_1), \ldots, \Phi_{\mathcal{X}}(X_N) \in \mathcal{H}_{k_{\mathcal{X}}}.$$
  $$Y_1, \ldots, Y_N \quad \mapsto \quad \Phi_{\mathcal{Y}}(Y_1), \ldots, \Phi_{\mathcal{Y}}(Y_N) \in \mathcal{H}_{k_{\mathcal{Y}}}.$$
- Apply CCA on $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$.

$$\max_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}}{\sqrt{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}}^2} \sqrt{\sum_{i=1}^{N} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}^2}}$$

where

$$\tilde{\Phi}_{\mathcal{X}}(X_i) = \Phi_{\mathcal{X}}(X_i) - \tfrac{1}{N} \textstyle\sum_{j=1}^{N} \Phi_{\mathcal{X}}(X_j), \quad \text{and } \tilde{\Phi}_{\mathcal{Y}}(Y_i) \text{ similar.}$$

Kernel CCA II

- We can assume $f = \sum_{i=1}^{N} \alpha_i \tilde{\Phi}_{\mathcal{X}}(X_i)$ and $g = \sum_{i=1}^{N} \beta_i \tilde{\Phi}_{\mathcal{Y}}(Y_i)$.

$$\rho = \max_{\alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \tilde{K}_X^2 \alpha} \sqrt{\beta^T \tilde{K}_Y^2 \beta}},$$

  $\tilde{K}_X$ and $\tilde{K}_Y$ are the centered Gram matrices.

- Regularization:
  Canonical correlation in N dimensional space with N data is ill-posed with correlation 1.

$$\max_{f \in \mathcal{H}_{\mathcal{X}}, g \in \mathcal{H}_{\mathcal{Y}}} \frac{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}}{\sqrt{\sum_{i=1}^{N} \langle f, \tilde{\Phi}_{\mathcal{X}}(X_i) \rangle_{\mathcal{H}_{\mathcal{X}}}^2 + \epsilon_N \|f\|^2} \sqrt{\sum_{i=1}^{N} \langle g, \tilde{\Phi}_{\mathcal{Y}}(Y_i) \rangle_{\mathcal{H}_{\mathcal{Y}}}^2 + \epsilon_N \|g\|^2}}$$

Kernel CCA III

- Kernel CCA

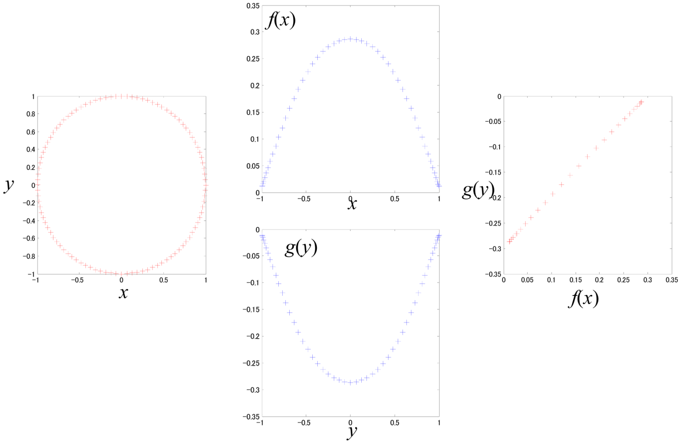
$$\begin{pmatrix} O & \tilde{K}_X\tilde{K}_Y \\ \tilde{K}_Y\tilde{K}_X & O \end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \rho\begin{pmatrix} \tilde{K}_X^2 + \epsilon_N K_X & O \\ O & \tilde{K}_Y^2 + \epsilon_N K_y \end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$

  The Solution is obtained as a generalized eigenproblem.

- The multiple feature vectors (second, third, eigenvectors) can be also obtained.
- Remark:
  - The results of kernel CCA depends on the kernels and ε_N.
  - The consistency is known if ε_N decreases sufficiently slowly as $N \to \infty$.

Toy exapmle of Kernel CCA

X, Y: one-dimensional. Gaussian RBF kernels are used.



Application of Kernel CCA

Application of kernel CCA to image retrieval ([HSST04]).

- Idea: use d eigenvectors $f_1, \ldots, f_d$ and $g_1, \ldots, g_d$ as the feature spaces which contain the dependence between X and Y.
- X_i: image, Y_i: text (extracted from webpages).
- Compute the feature vectors $f_1, \ldots, f_d$ and $g_1, \ldots, g_d$ by kernel CCA.
- Compute the projections $\xi_i = (\langle \Phi_{\mathcal{X}}(X_i), f_a \rangle_{\mathcal{H}_\mathcal{X}})_{a=1}^d \in \mathbb{R}^d$ for all images.
- For a new text Y_{new}, compute the projection $\zeta = (\langle \Phi_{\mathcal{Y}}(Y_{new}), g_a \rangle_{\mathcal{H}_\mathcal{Y}})_{a=1}^d \in \mathbb{R}^d$, and output the image

$$\arg\max_i = \xi_i^T \zeta.$$

Representer theorem I

- Minimization problems on RKHS

$$\min_{f\in\mathcal{H}_k} (Y_i - f(X_i))^2 + \lambda\|f\|^2 \qquad \text{(ridge regression)},$$

$$\min_{f\in\mathcal{H}_k, b} \textstyle\sum_{i=1}^N \big(1 - (Y_i f(X_i) + b)\big)_+ + \lambda\|f\|^2 \qquad \text{(SVM)}.$$

The solution can be taken from $f = \sum_{i=1}^N \alpha_i k(\cdot, X_i)$.

Representer theorem II

- General problem:
  - $\mathcal{H}_k$: RKHS with associated with a positive definite kernel k.
  - $X_1, \ldots, X_N, Y_1, \ldots, Y_N$: data.
  - $h_1(x), \ldots, h_m(x)$: fixed functions.
  - $\Psi : [0\ \infty) \to \mathbb{R}$: non-decreasing function (regularization term).

  Minimization

$$\min_{f \in \mathcal{H}, c \in \mathbb{R}^m} L\Big(\{X_i\}_{i=1}^N, \{Y_i\}_{i=1}^N, \{f(X_i) + \textstyle\sum_{a=1}^m c_a h_a(X_i)\}_{i=1}^N\Big) + \Psi(\|f\|).$$

Representer theorem

The solution of the above minimization is achieved by a function of the form

$$f = \textstyle\sum_{i=1}^N \alpha_i k(\cdot, X_i).$$

- The optimization in an high (or infinite) dimensional space can be reduced to the optimization in a subspace of N dimension (sample size).

Proof of the representer theorem

- Decomposition:
$$\mathcal{H}_k = H_0 \oplus H_0^{\perp},$$
$H_0 = \text{span}\{k(\cdot, X_1), \ldots, k(\cdot, X_N)\}$, $H_0^{\perp}$: orthogonal complement.
Decompose
$$f = f_0 + f^{\perp}$$
accordingly.
- Because
$$\langle f^{\perp}, k(\cdot, X_i)\rangle = 0,$$
the loss function L does not change by replacing f with f_0.
- The second term:
$$\|f_0\| \leq \|f\| \qquad \Longrightarrow \qquad \Psi(\|f_0\|) \leq \Psi(\|f\|).$$
- Thus, the optimum f can be in the space H_0.

Other kernel methods

- Kernel PLS (partial least square)
- Support vector regression (SVR)
- Kernel logistic regression
- Other variants of SVM (ν-SVM, one-class SVM etc., discussed later).

Summary of Chapter 3

- Various classical linear methods of data analysis can be kernelized – linear algorithms on RKHS.

  Kernel PCA, SVM, kernel CCA, kernel FDA, etc.

- The solution often has the form

$$f = \sum_{i=1}^{N} \alpha_i k(\cdot, X_i)$$

  (representer theorem).

- The problem is reduced to operations on Gram matrices of the sample size N.
- The kernel methods can be applied to any type of data including non-vectorial (structured) data, such as graphs, strings, etc, if a positive definite kernel is provided.

References I

Shotaro Akaho.
A kernel method for canonical correlation analysis.
In *Proceedings of International Meeting on Psychometric Society (IMPS2001)*, 2001.

Francis R. Bach and Michael I. Jordan.
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48, 2002.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor.
Canonical correlation analysis: An overview with application to learning methods.
Neural Computation, 16:2639–2664, 2004.

References II

Patrick M. Murphy and David W. Aha.
UCI repository of machine learning databases.
Technical report, University of California, Irvine, Department of Information and Computer Science.
http://www.ics.uci.edu/~mlearn/MLRepository.html, 1994.

Thomas Melzer, Michael Reiter, and Horst Bischof.
Nonlinear feature extraction using generalized canonical correlation analysis.
In *Proceedings of International Conference on Artificial Neural Networks (ICANN)*, pages 353–360, 2001.

References III

Sebastian Mika, Gunnar Rätsch, Jason Weston, Bernhard Schölkopf, and Klaus-Robert Müller.
Fisher discriminant analysis with kernels.
In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing*, volume IX, pages 41–48. IEEE, 1999.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller.
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319, 1998.