Statistical Learning on Reproducing Kernel Hilbert Spaces

Su-Yun Huang
Inst. Statistical Science, Academia Sinica

Workshop on Statistics and Machine Learning
at National Donghwa University, Feb. 5-6, 2004

Kernel-based learning algorithms

- Information technology related statistics– currently active, cross discipline research area.

- Kernel-based learning algorithms: SVM, kernel PCA, kernel ICA, kernel Fisher discriminant analysis, kernel SIR, etc.

  convenient algorithm $\xrightarrow{\text{kernelization}}$ same type algorithm on an RKHS.

- Reproducing kernels (RKs) provide a convenient framework for efficient computation.

- RKHS lays a theoretical foundation for statistical inference: sparse approximation, regularization, Gauss-Markov prediction, Bayesics, likelihood criterion, etc.

Basic properties of RKHS

RKHS: Basics -1

- Consider a linear class $\mathcal{H}$ of (real) functions $f(x)$ defined in a set E, forming a Hilbert space.

- **Definition** (Aronszajn, 1950, Trans. AMS). A real symmetric function $K(x, y)$ in $E \times E$ is called an **RK** of $\mathcal{H}$ if
  - For every $x \in E$, $K(x, \cdot) \in \mathcal{H}$.
  - For every $x \in E$ and $f \in \mathcal{H}$, we have the reproducing property

$$\langle\ f(\cdot),\ K(x, \cdot)\ \rangle_{\mathcal{H}} = f(x).$$

- All kernels considered in this talk are real symmetric.

- The space $\mathcal{H}$ is called an **RKHS**.

RKHS: Basics -2

RKHS $\rightarrow$ RK

- For the **existence** of an RK, it is necessary and sufficient that for every $y \in E$, the evaluation functional, $\ell_y : f \rightarrow f(y)$, $f \in \mathcal{H}$, is a continuous functional.

- If an RK exists, it is **unique**.

- Riesz representation theory: $\ell_y(f) = \langle f, g_y \rangle_{\mathcal{H}}$. The **RK** is given by $K(x, y) = g_y(x)$.

RKHS: Basics -3

Positive definite kernel $\rightarrow$ RKHS

- $K(x, y)$ is **positive definite** on $E \times E$ if, for all $x_1, \ldots, x_n \in E$, the quadratic form in $\xi_1, \ldots, \xi_n$: $\sum_{i,j=1}^{n} K(x_i, x_j)\xi_i\xi_j \geq 0$.

- To every positive definite kernel $K(x, y)$, there corresponds one and only one class of functions forming a Hilbert space and admitting K as an RK. (**existence and uniqueness**)

- Such a Hilbert space consists of functions of the form $\sum \alpha_i K(x, x_i)$ with norm

$$\| \sum \alpha_i K(x, x_i) \|_{\mathcal{H}}^2 = \sum_{i,j=1}^{n} K(x_i, x_j)\alpha_i\alpha_j.$$

RKHS: $\mathcal{H} =$ closure$\{\sum \alpha_i K(x, x_i)\}$

RKHS: Basics -4

- **Restriction** of an RK to $E_1 \subset E$.

  $\diamond$ $K_1(\cdot,\cdot) = K(\cdot,\cdot)|_{E_1 \times E_1}$: $\mathcal{H}_1$ with norm $\|f_1\|_{\mathcal{H}_1} = \inf_{\mathcal{F}} \|f\|_{\mathcal{H}}$,

  where $\mathcal{F} = \{f \in \mathcal{H} : f|_{E_1} = f_1\}$.

- **Sum** and **product** of RKs are still RKs.

  $\diamond$ $K_1(x,y) + K_2(x',y')$: $\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$.

  $\diamond$ $K_1(x,y)\, K_2(x',y')$: $\mathcal{H}_{K_1} \otimes \mathcal{H}_{K_2}$.

RKHS: Basics -5

- Discrete kernel **spectrum**.
  - $\diamond$ $K(x, y) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(y) =: \sum_{j=1}^{\infty} \tilde{\psi}_j(x) \tilde{\psi}_j(y)$,
  - $\diamond$ where $\|\psi_j\|^2_{L_2(E,P)} = 1$ and $\int K(x, y) \psi_j(y) dP(y) = \lambda_j \psi_j(x)$.
  - $\diamond$ Note that, for $f(x) = \sum_j f_j \psi_j(x)$, $\langle f, f \rangle_{\mathcal{H}} = \sum_j f_j^2 / \lambda_j$.
  - $\diamond$ $\{\tilde{\psi}_j = \sqrt{\lambda_j}\, \psi_j\}_{j=1}^{\infty}$: complete orthonormal basis for $\mathcal{H}$.

- If (E, P) is a finite measure space, then K has a discrete spectrum.

RKHS: Basics -6

Bounded linear functionals and operators on RKHS

- $\ell_f : \mathcal{H} \to R$, $\ell_f(h) = \langle f, h \rangle_{\mathcal{H}}$ (Riesz representation).

- $\Sigma : \mathcal{H} \to \mathcal{H}$, there corresponds a kernel on $E \times E$ given by

  $\Sigma(x,t) = \Sigma K_x(t)$, where $K_x(t) =: K(x,t)$, as a function of t.

Kernel SVM (in brief)

SVM classification on RKHS

Training data: $\{x_i, y_i\}$, $x_i \in R^n$ and $y \in \{-1, 1\}$ for $i = 1, \ldots, l$.

Goal: Look for a discriminant boundary, $f(x) = 0$, that separates the positive y's from the negative y's with "*maximum margin*".

Linear SVM: The algorithm looks for the separating hyperplane $w'x + b = 0$ with largest margin (given by $2/\|w\|_2$). That is, set $f(x) = w'x + b$, and solve the following constrained minimization problem:

$$\min_{w \in R^d} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{l} \xi_i \quad \text{subject to} \quad y_i f(x_i) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i = 1, \ldots, l.$$

From linear SVM to kernel SVM

RKHS – a foundation for theoretical properties and
– a framework for efficient computation.

- start with a linear separation algorithm (maximizing margin)
- kernelization of the underlying linear learning algorithm,
- nonlinear separation $\overset{RKHS}{- - - \longrightarrow}$ linear separation in feature space.
- sparse dual representation in an RKHS $\rightarrow$ efficient algorithm,
- equivalence among regularization, sparse approximation, Bayesics, Gauss-Markov prediction (Huang and Lee, 2003);

  likelihood-based statistical inference, etc.

SVM, linear separable case

$$\min_{w\in R^d, b\in R, \alpha_i\geq 0} \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{l} \alpha_i\{y_i(w'x_i + b) - 1\}.$$

$$\partial()/\partial b = 0 \;\;\rightarrow\;\; \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\partial()/\partial w = 0 \;\;\rightarrow\;\; w = \sum_{i=1}^{l} \alpha_i y_i x_i.$$

Dual problem:

$$\max_{\alpha_i\geq 0}(\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i\alpha_j y_i y_j x_i' x_j) \;\; \text{subject to} \;\; \sum_{i=1}^{l} \alpha_i y_i = 0.$$

SVM separating hyperplane:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i\, x_i'\, x + b,$$

with $b = -\frac{1}{2}\{\max_{j\in I_-}(\sum_{i=1}^{l} \alpha_i y_i\, x_i' x_j) + \min_{j\in I_+}(\sum_{i=1}^{l} \alpha_i y_i\, x_i' x_j)\}$.

SVM, linear non-separable case

$$\min_{w \in R^d, b \in R, \xi_i \geq 0} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{l} \xi_i \quad \text{subject to} \quad y_i f(x_i) \geq 1 - \xi_i, \forall i = 1, \ldots, l.$$

Dual problem:

$$\max_{0 \leq \alpha_i \leq C} \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i' x_j \right) \quad \text{subject to} \quad \sum_{i=1}^{l} \alpha_i y_i = 0.$$

SVM separating hyperplane:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i \, x_i' x + b,$$

with $b = -\frac{1}{2}\{\max_{j \in I_-^*}(\sum_{i=1}^{l} \alpha_i y_i \, x_i' x_j) + \min_{j \in I_+^*}(\sum_{i=1}^{l} \alpha_i y_i \, x_i' x_j)\}$,
where $*$: zero slack.

Kernel SVM

- Map the data in $\mathcal{X}$ to some high dimensional space $\mathcal{Z}$, called the feature space: $x \rightarrow \tilde{\Psi}(x) = (\tilde{\psi}_1(x), \tilde{\psi}_2(x), \ldots)'$,

- $K(x, u) = \sum_{\nu=1}^{\infty} \tilde{\psi}_\nu(x)\tilde{\psi}_\nu(u) = \sum_{\nu=1}^{\infty} \lambda_\nu \psi_\nu(x)\psi_\nu(u), \qquad \lambda_\nu = \|\tilde{\psi}_\nu\|_2^2.$

  $f(x) = \sum_\nu f_\nu \psi_\nu(x), \quad \|f\|_{\mathcal{H}_K}^2 = \sum_\nu f_\nu^2/\lambda_\nu.$

- feature mapping: $\mathcal{X} \rightarrow \mathcal{Z}$, linear separation on $\mathcal{Z}$.
  RKs make the linear separation algorithm practically working without resorting to the feature mapping Ψ.

- SVM (a regularization problem on RKHS):

  $\min_{f \in \mathcal{H}_K + b} \frac{1}{2}\|f\|_{\mathcal{H}_K}^2 + C(\sum_{i=1}^{l} \xi_i)$

  subject to $y_i f(x_i) \geq 1 - \xi_i,\ \xi_i \geq 0, \forall i = 1, \ldots, l.$

Kernel SVM, continued

Dual problem: $f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x, x_i) + b$

$$\max_{0 \le \alpha_i \le C} \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$$

$$\text{subject to} \quad \sum_{i=1}^{l} \alpha_i y_i = 0.$$

Two easy-to-understand kernels

$$
\begin{aligned}
\textit{linear spline}: K_{\mathsf{lsp}}(t,s) &= \min\{s,t\}, \quad s,t \in [0,1], \\
\textit{Gaussian kernel}: K_{\mathsf{rbf}}(t,s) &= \exp\left\{-\frac{1}{2\sigma^2}\|t-s\|^2\right\}, \quad s,t \in R^d.
\end{aligned}
$$

SVM with linear splines

$\diamond$ $K_{\mathsf{lsp}}(t,s) = \min\{s,t\}$, $s,t \in [0,1]$, is the reproducing kernel for the following RKHS:

$$\mathcal{H}_K = \{f : \text{abs. conti. on } [0{,}1],\ f(0) = 0 \text{ and } \|f\|_{\mathcal{H}_K} = \|f'\|_2 < \infty\}.$$

$\diamond$ SVM: $\min_{f\in\mathcal{H}_K+b} \frac{1}{2}\|f\|^2_{\mathcal{H}_K} + C\times(\text{data goodness of fit})$

subject to ......

$\diamond$ Regularize the first derivatives with penalty on $\|f'\|_2^2$.

SVM with Gaussian kernel

$\diamond$ $K_{\text{rbf}}(t,s) = \exp\left\{-\frac{1}{2\sigma^2}\|t-s\|^2\right\}$, $s,t \in R^d$, is the reproducing kernel for the following RKHS:

$$\mathcal{H}_K = \left\{ f \in C^\infty : \ \|f\|^2_{\mathcal{H}_K} = \sum_{k=0}^{\infty} \frac{\sigma^{2k}}{2^k\, k!} \|f^{(k)}\|_2^2 < \infty \right\}.$$

$\diamond$ SVM: $\min_{f \in \mathcal{H}_K + b} \frac{1}{2}\|f\|^2_{\mathcal{H}_K} + C \times$(data goodness of fit)

subject to ......

$\diamond$ Penalize on $\sum_{k=0}^{\infty} \frac{\sigma^{2k}}{2^k\, k!} \|f^{(k)}\|_2^2$.

Note the regularization on the k-th derivative is multiplied by σ^{2k}.

Kernel Fisher discriminant analysis

Classical Fisher linear discriminant analysis

- Input data: $\{x_j \in \mathcal{X} \subset R^n\}_{j=1}^{l}$.

- Group labels: $\{y_j = \pm 1\}_{j=1}^{l}$.

- Find a discriminant hyperplane "$w^t x + b = 0$", which separates the two groups.

- Mahalanobis distance criterion: Classify a test input x by

$$\text{sign}\{d(x, \bar{x}_2) - d(x, \bar{x}_1)\},$$

where $d(x, \bar{x}_i) = (x - \bar{x}_i)^t S^{-1} (x - \bar{x}_i)$ with S the pooled covariance matrix. (i.e., $S = \sum_{i=1}^{2} \sum_{j \in I_i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)^t / l$.)

- Maximal likelihood ratio criterion: $x_j \sim N(\mu_i, \Sigma)$, $j \in I_i$. $\log MLR$

Kernel FDA – Ideas behind kernelization

- When the data space $\mathcal{X}$ is not big enough for linear separation, or the coordinate system adopted is not feasible for linear separation, we resort to other means $\rightarrow$ kernel approach.

- Map the data in $\mathcal{X}$ to some high-dimensional Hilbert space (called the feature space) $\mathcal{Z} \subset R^q$. Often, $q = \infty$.

- Transformation:

$$z =: (\tilde{\psi}_1(x), \ldots, \tilde{\psi}_q(x))^t =: (\sqrt{\lambda}_1 \psi_1(x), \ldots, \sqrt{\lambda}_q \psi_q(x))^t,$$

  where $\{\psi_k\}_{k=1}^q$ are linear independent functions with unit L_2-length, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$.

- $K(x, u) =: z(x)^t z(u) = \sum_k \lambda_k \psi_k(x) \psi_k(u)$.

- Symbolically, "*perform*" the classical FLDA on the mapped data in $\mathcal{Z}$.

$$z \to \text{sign}\left\{ z^t S_w^{-1}(\bar{z}_1 - \bar{z}_2) - \frac{1}{2}(\bar{z}_1 + \bar{z}_2)^t S_w^{-1}(\bar{z}_1 - \bar{z}_2) \right\},$$

where $S_w = \sum_{j \in I} z_j z_j^t - \sum_{i=1}^2 l_i \bar{z}_i \bar{z}_i^t$.

- Since $S_w^{-1}(\bar{z}_1 - \bar{z}_2)$ is of form $\alpha_1 z_1 + \cdots + \alpha_l z_l$,

  the discriminant function is of form $f(x) = \sum_{j=1}^l \alpha_j K(x, x_j) + b$.

- Operate on $\{K(x_i, x_j)\}_{i,j=1}^l$ and group labels $\{y_j\}_{j=1}^l$. In practice, the kernel spectrum, given by Λ and Ψ, is not known.

Notation for KFDA

- Let I_1 be the index set of training sample for group label $y = 1$, I_2 for $y = -1$ and $I = I_1 \cup I_2$. Let $l_i = |I_i|$ be the size of I_i and $l = |I|$ be the size of I.

- Let $\mathbf{1} \in R^l$ be the vector of all ones, and let $\mathbf{1}_1, \mathbf{1}_2 \in R^l$ be as binary (0,1) vectors corresponding to their group label with 0 for non-members and 1 for members. With such definition, it leads to that $\mathbf{1}_1 + \mathbf{1}_2 = \mathbf{1}$.

- Let $Z =: (\Lambda^{1/2} \circ \Psi(x_1), \ldots, \Lambda^{1/2} \circ \Psi(x_l))^t$, which is an $l \times n$ matrix. Let $K = ZZ^t$. Then, the (i,j)-th entry of K, denoted by K_{ij}, is given by $K(x_i, x_j)$.

- Let $\bar{z}_i = \frac{1}{l_i} \sum_{j \in I_i} z_j$, $i = 1, 2$, be the group centroid in the feature space, where $z_j(x) = \Lambda^{1/2} \circ \Psi(x_j)$, and let $\bar{z} = (\sum_{j=1}^{l} z_j)/l$.

- Let $\bar{k}_i = \frac{1}{l_i}\sum_{j\in I_i} K_j$, $i = 1, 2$, be the kernelized group centroid, where K_j is the j-th column vector of matrix K.

- Let $S_b = (\bar{z}_1 - \bar{z}_2)(\bar{z}_1 - \bar{z}_2)^t$ and $S_w = \sum_{j\in I} z_j z_j^t - \sum_{i=1}^2 l_i \bar{z}_i \bar{z}_i^t$ be the between- and within-class sample covariances for data in the feature space.

- Let $M_b = (\bar{k}_1 - \bar{k}_2)(\bar{k}_1 - \bar{k}_2)^t$ and $M_w = K^2 - \sum_{i=1}^2 l_i \bar{k}_i \bar{k}_i^t$ be the between- and within-class sample covariances for kernelized data.

KFDA in the feature space

Separating boundary : $z^t S_w^{-1}(\bar{z}_1 - \bar{z}_2) - \frac{1}{2}(\bar{z}_1 + \bar{z}_2)^t S_w^{-1}(\bar{z}_1 - \bar{z}_2) = 0.$

The KFDA finds the discriminant function of the form

$$f(x) = w^t z + b = \sum_{k=1}^{d} w_k \sqrt{\lambda_k}\, \psi_k(x) + b$$

passing through the mid point of group centroids, where w is the maximizing argument in the Rayleigh coefficient

$$J_{KFDA}(w) \equiv \frac{w^t S_b w}{w^t S_w w}.$$

A regularized within-class covariance of the form $S_w + rW$ is considered and w is the solution to the following maximization problem

$$\arg\max_{w \in R^q} J_{RKFDA}(w) \equiv \arg\max_{w \in R^q} \frac{w^t S_b w}{w^t (S_w + rW) w}.$$

The extra term rW is added to

– to overcome the numerical problem caused by singular within-class covariance in a high-dimensional feature space,

– to control the smoothness and the shape of the fitted discriminant hypersurface.

The discriminant function can be re-formulated as

$$f(x) = b + \sum_{j=1}^{l} \alpha_j K(x_j, x).$$

The coefficients α_js can be obtained as the solution to the following maximization problem

$$\arg\max_{\alpha \in R^l} J_{RKFDA}(\alpha) \equiv \arg\max_{\alpha \in R^l} \frac{\alpha^t M_b \alpha}{\alpha^t (M_w + rA)\alpha}.$$

Again, the extra term rA is added to the within-class sample covariance for the same purposes as before.

In next slides we formulate the KFDA and its extension as a likelihood ratio of two Gaussians on an RKHS.

KFDA – a likelihood ratio criterion

- Classical FLDA: P_1 and P_2 Gaussian with a common covariance.

  $$\log(dP_1(x)/dP_2(x)) = x^t\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^t\Sigma^{-1}(\mu_1 - \mu_2)$$

  Plug in MLE for μ_i and Σ.

- Kernel FDA: Gaussian measures on RKHS, likelihood ratio, MLE.

Gaussian measure and covariance operator on an RKHS

Definition 1 (Gaussian measure on $\mathcal{H}$) *A probability measure $P_{\mathcal{H}}$ on $(\mathcal{H}, \mathcal{T})$ is said to be Gaussian with respect to $\{\ell_f\}_{f\in\mathcal{H}}$ if for any k and any bounded linear functionals $\ell_{f_1}, \ldots, \ell_{f_k}$ the joint distribution of $\ell_{f_1}(h)$,..., $\ell_{f_k}(h)$ is normal, where h is a random element in $\mathcal{H}$ with distribution $P_{\mathcal{H}}$.*

Definition 2 (Covariance operator) *A covariance operator, denoted by Σ, is defined to be an operator mapping from $\mathcal{H}$ into $\mathcal{H}$ which is bounded, linear, nonnegative definite, self-adjoint and trace class (i.e., of finite trace).*

Let P_1 and P_2 be two equivalent probability measures on $(\mathcal{X}, \mathcal{B})$. Consider the mapping $\gamma : x \to K(x, \cdot) =: K_x(\cdot) \in \mathcal{H}$. Let $P_{1,\mathcal{H}}$ and $P_{2,\mathcal{H}}$ denote the probability measures on $(\mathcal{H}, \mathcal{T})$ induced from P_1 and P_2 by γ. Assume that $P_{1,\mathcal{H}}$ and $P_{2,\mathcal{H}}$ are Gaussian with different mean functions

$$m_i(t) = E_{P_i} K_X(t) = \sum_{\nu} \lambda_\nu \psi_\nu(t)\, E_{P_i} \psi_\nu(X), \quad i = 1, 2,$$

and a common covariance operator

$$\Sigma_{\mathcal{H}}(s,t) = cov_{P_1}(K_X(s), K_X(t)) = cov_{P_2}(K_X(s), K_X(t)).$$

The mean functions and the covariance operator satisfy the following properties (see, for instance, Vakhania *et al.*, 1987)

$$E_{P_{i,\mathcal{H}}} \langle f, K_X \rangle_{\mathcal{H}} = \langle f, E_{P_i} K_X \rangle_{\mathcal{H}},$$

$$cov_{P_{i,\mathcal{H}}} \{ \langle f, K_X \rangle_{\mathcal{H}}, \langle g, K_X \rangle_{\mathcal{H}} \} = \langle \Sigma_{\mathcal{H}} f, g \rangle_{\mathcal{H}} = \langle f, \Sigma_{\mathcal{H}} g \rangle_{\mathcal{H}}.$$

Likelihood ratios

Theorem 1 (Grenander, 1952) *Let $P_{1,\mathcal{H}}$ and $P_{2,\mathcal{H}}$ be two equivalent Gaussian measures on $(\mathcal{H}, \mathcal{T})$ with mean $m_1(t)$ and $m_2(t)$ and a common covariance operator $\Sigma_{\mathcal{H}}$, which is assumed non-singular. Also assume that $(m_1 - m_2)$ is in the range of $\Sigma_{\mathcal{H}}$. Then the logarithm of the likelihood ratio is linear and given by*

$$\begin{aligned} & \log(dP_{1,\mathcal{H}}/dP_{2,\mathcal{H}})(K_x) \\ = \; & \langle K_x, \Sigma_{\mathcal{H}}^{-1}(m_1 - m_2)\rangle_{\mathcal{H}} - \frac{1}{2}\langle m_1 + m_2, \Sigma_{\mathcal{H}}^{-1}(m_1 - m_2)\rangle_{\mathcal{H}}, \end{aligned}$$

where $K_x(t) =: K(x, t)$, as a function of t.

– a test input $x \rightarrow$ a realization of the process $K_x(t)$,

– plug in MLE for means and covariance operator.

KFDA as a maximal likelihood ratio test

Classification for a test input x:

$$\begin{aligned}
&\operatorname{sign}\{\log(dP_{1,\mathcal{H}}/dP_{2,\mathcal{H}})(K_x)\} \\
= \;&\operatorname{sign}\left\{\Sigma_{\mathcal{H}}^{-1}(m_1 - m_2)(x) - \frac{1}{2}\langle m_1 + m_2, \Sigma_{\mathcal{H}}^{-1}(m_1 - m_2)\rangle\right\}.
\end{aligned}$$

Plug in ML estimates

$$\hat{m}_i(t) =: \frac{1}{l_i}\sum_{j\in I_i} K(x_j, t),$$

$$\hat{\Sigma}_{\mathcal{H}}(s,t) =: \frac{1}{l}\sum_{i=1}^{2}\sum_{j\in I_i}(K(x_j,s) - \hat{m}_i(s))(K(x_j,t) - \hat{m}_i(t)) + \epsilon A(s,t).$$

With some technical details, then we result in the previously discussed KFDA algorithm.

Concluding remarks on kernelization

- Kernelization of a linear algorithm, or any convenient algorithm on $\mathcal{X} \xrightarrow{\text{leads to}}$ same type of algorithm on an RKHS, but more flexible and versatile one on the original data space $\mathcal{X}$.

- RK provides a framework for efficient computation.

- RKHS lays a foundation for theory of statistical inference.

Acknowledgement

Thanks to

C.R. Hwang for helpful discussion on Gaussian measure, likelihood ratio and MLE on an RKHS,

C.Y. Chang, Y.J. Lee and H.S. Lu for general discussion.

The following is a list of selected references. Some short notes are appended based on the speaker's *subjective viewpoint* and *limited knowledge*.

References - general theory of RKHS and statistical learning/inference on RKHS

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686, 337–404. (general theory of RKHS)

Baker, C.R. (1970). Mutual information for Gaussian processes. *SIAM J. Appl. Math.*, 19, 451–458. (Gaussian processes, covariance and cross-covariance operators on RKHS)

Grenander, U. (1952). Stochastic processes and statistical inference. *Arkiv för Matematik*, 1, 195–277. (statistical inference on RKHS)

Grenander, U. (1981) *Abstract Inference*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. (statistical inference in abstract sample space)

Mendelson, S. (2002). Learnability in Hilbert spaces with reproducing kernels. *J. Complexity*, 18, 152–170.

Rao, C.R. and Varadarajan, V.S. (1963). Discrimination of Gaussian processes. *Sankhyā*, A, 25, 303–330. (statistical hypothesis testing on RKHS)

Saitoh, S. (1988). *Theory of Reproducing Kernels and Its Applications*. Pitman Research Notes in Mathematics Series 189. Longman Scientific & Technical, UK. (general theory of RKHS)

Saitoh, S. (1997). *Integral Transforms, Reproducing Kernels and Their Applications*. Pitman Research Notes in Mathematics Series 369. Longman Scientific & Technical, UK. (general theory of RKHS)

Vakhania, N.N. Tarieladze, V.I. and Chobanyan, S.A. (1987). *Probability Distributions on Banach Spaces.* Translated from the Russian by W.A. Woyczynski. Mathematics and Its Applications (Soviet Series), 14, D. Reidel Publishing Co., Dordrecht, Holland.

Wahba, G. (1990). *Spline Models for Observational Data*. Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia. (spline models on RKHS)

References - kernel FDA

Baudat, G. and Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12, 2385–2404.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B. and Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In Hu, Y.-H., Larsen, J., Wilson, E., and Douglas, S., eds, *Neural Networks for Signal Processing*, IX, 41–48, IEEE.

Mika, S., Rätsch, G. and Müller, K.-R. (2001). A mathematical programming approach to the kernel Fisher Algorithm. In T.K. Leen, T.G. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems*, 13, 591–597, MIT Press.

Mika, S., Smola, A. and Schölkopf, B. (2001). An improved training algorithm for kernel Fisher discriminants. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics*, 98–104, Morgan Kaufmann.

Xu, J., Zhang, X. and Li, Y. (2001). Kernel MSE algorithm: a unified framework for KFD, LS-SVM and KRR. *Proceedings Intern. Joint Conf. Neural Networks*, 2, 1486–1491, IEEE Press.

References - other RKHS related methods

Bach, F.R. and Jordan, M.I. (2002). Kernel independent component analysis. *J. Machine Learning Research*, 3, 1–48.

Cortes, C. and Vapnik, V.N. (1995). Support vector networks. *Machine Learning*, 20, 1–25. (original article on SVM)

Fukumizu, K., Bach, F.R. and Jordan, M.I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Machine Learning Research*, 5, 73–99. (kernel dimensionality reduction method, based on cross-covariance operator in RKHS to search for effective dimension reduction subspace)

Genton, M.G. (2001). Classes of kernels for machine learning: a statistical perspective. *J. Machine Learning Research*, 2, 299–312. (a study of various kinds of kernels)

Herbrich, R., Graepel, T. and Campbell, C. (1999). Bayesian learning in reproducing kernel Hilbert spaces. Tech. report 99-11, Technical University, Berlin. http://stat.cs.tu-berlin.de/∼ralfh/publications.html

Huang, S.Y. and Lee, Y.J. (2003). Equivalence relations between support vector machines, sparse approximation, Bayesian regularization and Gauss-Markov prediction. Technical report.

Huang, S.Y. and Lu, H.H.-S. (2001). Extended Gauss-Markov theorem for nonparametric mixed-effects models. *J. Multivariate Anal.*, 76, 249–266. (Gauss-Markov theorem on RKHS)

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28, 1570–1600.

Lin, Y., Wahba, G., Zhang, H. and Lee, Y. (2002). Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48, 115–136.

Parsen, E. (1961). An approach to time series analysis. *Ann. Math. Statist.*, 32, 951–989.

Parsen, E. (1962). Extraction adn detection problems and reproducing kernel Hilbert spaces. *J. SIAM Control Ser. A*, 1, 35–62.

Rosipal, R. and Trejo, L.J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *J. Machine Learning Research*, 2, 97–123.

Smola, A. and Schölkopf, B. (1998). On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22, 211–231.

Vijayakumar, S. and Ogawa, H. (1999). RKHS based functional analysis for exact incremental learning. *Neurocomputing: Special issue on Theoretical Analysis of Real Valued Function Classes*, 29, 85–113.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces and the randomized GAVC. *Advances in Kernel Methods: Support Vector Machines.* B. Schölkopf, C. Burges and A.J. Smola, eds., 69–88, MIT Press, Cambridge, MA.

Wabha, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings NAS*, 99 (26), 16524–16530. (RKHS provides a unified context for solving a wide variety of statistical modelling and function estimation problems. This article considers soft classification via penalized likelihood estimation and SVM hard classification, both in the context of RKHS.)

Wu, H.M. and Lu, H.H.S. (2004). Supervised motion segementation by spatial-frequential analysis and dynamic sliced inverse regression. *Statistica Sinica*, to appear. (dynamic SIR+SVM for motion segementation. This work leads some thoughts to kernelized SIR yet to be studied.)