Useful Applications in Statistical Learning with Reproducing Kernel Hilbert Spaces

Yingjie Cao

Stockholm University

zjtxcyj@gmail.com

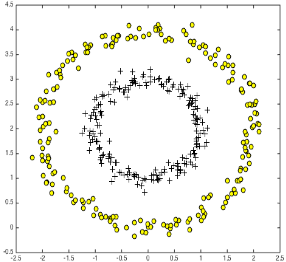
May 19, 2016

Overview

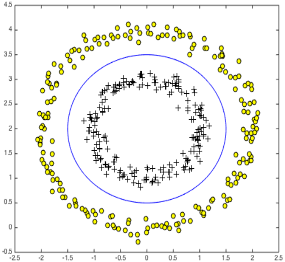
1. Nonlinear Regularization Problem
   - Nonlinear Problem
   - Regularization Problem
2. Reproducing Kernel Hilbert Spaces
   - Some Facts About RKHS
   - Apply RKHS in Regularization Problems
   - Positive Definite Kernels and Their Related RKHS
3. Applications in Different Statistical Learning Algorithms
   - Kernel PCA
   - Kernel SVM
   - Real-world Example

A Simple Example

First, let's have a look at two groups of data points (1a), how can we make a boundary between these two groups?



(a) Two Groups of Data



(b) Boundary of Groups

A Simple Example

Since this is a nonlinear problem, we might think about adding features, from previous two features X_1, X_2 to five features: $X_1, X_2, X_1^2, X_2^2, X_1X_2$. Therefore, the boundary can be written in this form:

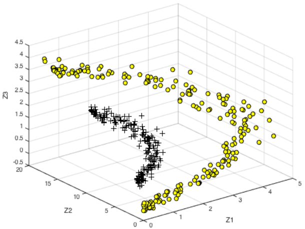
$$a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_2^2 + a_5X_1X_2 + a_6 = 0 \tag{1}$$

This idea can be approached in another way. Let $Z_1 = X_1, Z_2 = X_1^2, Z_3 = X_2, Z_4 = X_2^2, Z_5 = X_1X_2$, (1) can be written as:

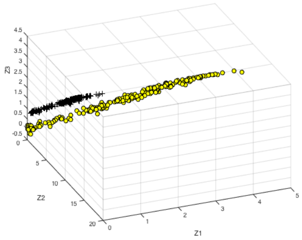
$$\sum_1^5 a_iZ_i + a_6 = 0 \tag{2}$$

A Simple Example

Given a map $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^5$, the two groups of data points can be easily divided by the five dimensional hyperplane. Here we can only use three dimensions, with $Z_1 = X_1^2, Z_2 = X_2^2, Z_3 = X_2$, the two groups of data are shown as (1c). Through rotation (1d), the data points can be easily separated by a plane.

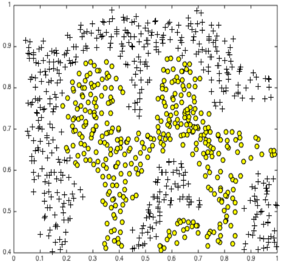


(c) AZ=-37.5° , EL=30°

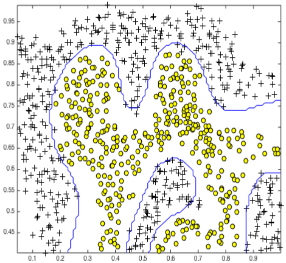


(d) AZ=22° , EL=-40°

More Complicated Case...



(e) Two Groups of Data



(f) Boundary of Groups

Conclusion

In order to move beyond linearity, a core idea is to replace the vector of inputs X with additional variables, which are transformations of X, and then use linear models in this new space of derived input features.
Denote by $h_m(X) : \mathbb{R}^p \to \mathbb{R}$ the mth transformation of $X, m = 1, \ldots, M$.
We then model

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X) \tag{3}$$

Among all functions $f(X)$, we want to find one that minimizes the penalized residual sum of squares

$$RSS(f, \lambda) = \sum_{i=1}^{N} \left[y_i - f(x_i) \right]^2 + \lambda J(f) \tag{4}$$

Regularization Problem

A general class of regularization problems has the form

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right] \quad (5)$$

where $L(y, f(x))$ is a loss function, $J(f)$ is a penalty functional, and $\mathcal{H}$ is a space of functions on which $J(f)$ is defined.

Example (Ridge Regression)

$$L(y_i, f(x_i)) = (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \quad J(f) = \sum_{j=1}^{p} \beta_j^2 \quad (6)$$

Why Regularization?

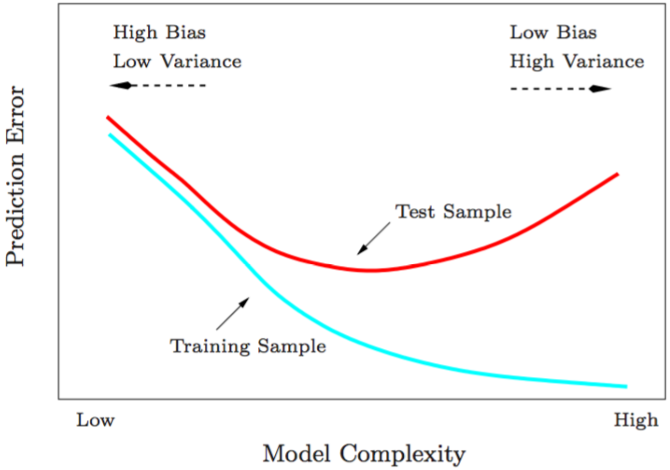
Regularization is designed to address the problem of overfitting.

- High bias or underfitting is when the form of our hypothesis maps poorly to the trend of the data. It is usually caused by a function that is too simple or uses too few features.
- At the other extreme, overfitting or high variance is caused by a hypothesis function that fits the available data but does not generalize well to predict new data.

There are two main options to address the issue of overfitting:

- Reduce the number of features.
- Regularization: keep all the features, but reduce the parameters θ_j.

Model Complexity



Before We Begin

Keywords

- reproducing kernel
- dot product kernel
- positive definite function
- reproducing kernel Hilbert space
- kernel trick
- kernel PCA
- kernel SVM

Reproducing Kernel Hilbert Space

Definition (Evaluation functional)

An evaluation functional over the Hilbert space $\mathcal{H}$ is a linear functional $\mathcal{F}_x : \mathcal{H} \to \mathbb{R}$ that evaluates each function in the space at the point x, or

$$\mathcal{F}_x[f] = f(x) \quad \text{for all } f \in \mathcal{H} \tag{7}$$

Definition (Reproducing kernel Hilbert space)

A Hilbert space $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) if the evaluation functions are bounded, i.e. if for all x there exists some $M > 0$ such that

$$|\mathcal{F}_x[f]| = |f(x)| \leq M\|f\|_{\mathcal{H}_K} \quad \text{for all } f \in \mathcal{H}_K \tag{8}$$

Reproducing Property

Definition (Reproducing kernel)

Let $\mathcal{H}_K$ be a Hilbert space of functions $f : \Omega \to \mathbb{R}$ defined on a non-empty set Ω. A function $K : \Omega \times \Omega \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}_K$ if for each $x \in \Omega$, it satisfies

- $$K_x := K(\cdot, x) \in \mathcal{H}_K \tag{9}$$
- $$\mathcal{F}_x[f] = \langle K_x, f \rangle_{\mathcal{H}_K} = f(x) \quad \text{for all } f \in \mathcal{H}_K \tag{10}$$

Definition

Definition (Dot product kernel)

A function $K : \Omega \times \Omega \to \mathbb{R}$ is called a dot product kernel on Ω if there exists a Hilbert space $\mathcal{F}$ and a map $\phi : \Omega \to \mathcal{F}$ such that $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$.

Definition (Positive definite function)

A symmetric function $K : \Omega \times \Omega \to \mathbb{R}$ is called positive definite if, for any $n \in \mathbb{N}$ and every set of real numbers $\{a_1, a_2, \ldots, a_N\}$ and $\{x_1, x_2, \ldots, x_N\}$, $x_i \in \Omega$, we have

$$\sum_{i,j=1}^{N} a_i a_j K(x_i, x_j) \geq 0 \tag{11}$$

Relationship

Every reproducing kernel is a dot product kernel

Take feature map $\phi : x \mapsto K(\cdot, x)$, then $K(x, y) = \langle K(\cdot, x), K(\cdot, y)\rangle_{\mathcal{H}}$, i.e., RKHS $\mathcal{H}$ is a feature space.

Every dot product kernel is a positive definite function

Let $K(x, y)$ be a kernel, then there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \Omega \to \mathcal{H}$ such that $K(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$. We then have

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \langle \phi(x_i), \phi(x_j)\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0\end{aligned} \tag{12}$$

So Far

reproducing kernel $\Longrightarrow$ dot product kernel $\Longrightarrow$ positive definite function

Question

Is every positive definite function a reproducing kernel for some RKHS?

Moore-Aronszajn Theorem

Theorem (Moore-Aronszajn Theorem)

To every positive definite function K on $\Omega \times \Omega$ there corresponds a unique RKHS $\mathcal{H}_K$ of real valued functions on Ω for which K is a reproducing kernel.

Proof.

Let $\mathcal{H}_0 = \text{span}\{K(x, \cdot) : x \in \Omega\}$ be endowed with the inner product:

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j K(x_i, y_j) \tag{13}$$

where $f = \sum_{i=1}^{m} a_i K(\cdot, x_i)$ and $g = \sum_{j=1}^{n} b_j K(\cdot, y_j)$.
Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ with respect to this inner product. Then $\mathcal{H}$ consists of functions of the form $f(x) = \sum_{i=1}^{\infty} a_i K_{x_i}(x)$. $\square$

Summary

reproducing kernel $\Longleftrightarrow$ dot product kernel $\Longleftrightarrow$ positive definite function

reproducing kernel $\overset{1:1}{\longleftrightarrow}$ RKHS

Apply RKHS in Regularization Problems

Review the regularization problem:

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \right] \tag{14}$$

The solution for (14) can be shown as finite-dimensional (Representer Theorem), and has the form:

$$f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i). \tag{15}$$

$$\|f\|^2_{\mathcal{H}_K} = \langle \alpha_i K(\cdot, x_i), \alpha_j K(\cdot, x_j) \rangle_{\mathcal{H}_K} = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \alpha_i \alpha_j \tag{16}$$

Therefore, the regularization problem can be simplified into a finite-dimensional criterion:

$$\min_{\boldsymbol{\alpha}} L(\boldsymbol{y}, \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha} \tag{17}$$

The Representer Theorem

Theorem (The Representer Theorem)

The minimizer over the RKHS $\mathcal{H}$, f_S^λ, of the regularized empirical functional:

$$I_S[f] + \lambda \|f\|_{\mathcal{H}}^2 \tag{18}$$

can be represented by the expression:

$$f_S^\lambda(x) = \sum_{i=1}^{n} c_i K(x_i, x) \tag{19}$$

for some n-tuple $(c_1, \dots, c_n) \in \mathbb{R}^n$.

The Representer Theorem

Proof.

Define the linear subspace of $\mathcal{H}$,

$$\mathcal{H}_0 = \{f \in \mathcal{H} : f = \sum_{i=1}^{n} \alpha_i K_{x_i}\} \tag{20}$$

Let $\mathcal{H}_0^{\perp}$ be the linear subspace of $\mathcal{H}$ orthogonal to $\mathcal{H}_0$:

$$\mathcal{H}_0 = \{g \in \mathcal{H} : \langle g, f \rangle = 0, \forall f \in \mathcal{H}_0\} \tag{21}$$

$\mathcal{H}_0$ is finite-dimensional, hence closed. So we can write $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^{\perp}$, then every $f \in \mathcal{H}$ can be decomposed into a component along $\mathcal{H}_0$, denoted by f_0, and a component perpendicular to $\mathcal{H}_0$, given by $f_0^{\perp}$:

$$f = f_0 + f_0^{\perp} \tag{22}$$

The Representer Theorem

Proof.

By orthogonality

$$\|f_0 + f_0^{\perp}\|^2 = \|f_0\|^2 + \|f_0^{\perp}\|^2 \tag{23}$$

and by the reproducing property

$$I_S[f_0 + f_0^{\perp}] = I_S[f_0] \tag{24}$$

Combining these two facts, we see that

$$\begin{aligned} I_S[f_0 + f_0^{\perp}] + \lambda\|f_0 + f_0^{\perp}\|_{\mathcal{H}}^2 &= I_S[f_0] + \lambda\|f_0\|_{\mathcal{H}}^2 + \lambda\|f_0^{\perp}\|_{\mathcal{H}}^2 \\ &\geq I_S[f_0] + \lambda\|f_0\|_{\mathcal{H}}^2 \end{aligned} \tag{25}$$

Hence the minimum $f_S^{\lambda} = f_0$ must belong to the linear space $\mathcal{H}_0$. □

Linear Kernel and Its Related RKHS

Linear kernel

$$K(x, y) = \langle x, y \rangle \tag{26}$$

RKHS of Linear Kernel

The RKHS of the linear kernel consists of function

$$x \in \mathbb{R}^d \longmapsto f(x) = \sum_i a_i \langle x_i, x \rangle = \langle \omega, x \rangle \tag{27}$$

with $\omega = \sum_i a_i x_i$.
Hence the RKHS is the set of linear forms endowed with the following inner product:

$$\langle f, g \rangle_{\mathcal{H}_K} = \langle \omega, \upsilon \rangle \tag{28}$$

when $f(x) = \langle \omega, x \rangle$ and $g(x) = \langle \upsilon, x \rangle$. So we have $\|f\|_{\mathcal{H}_K} = \|\omega\|_2$.

Polynomial Kernel and Its Related RKHS

Polynomial Kernel

$$K(x, y) = (\langle x, y\rangle + c)^d \tag{29}$$

where $c \geq 0$ be a non-negative real and $d \geq 1$ be an integer.

To prove polynomial kernel is positive definite, we can expand the sum:

$$K(x, y) = \langle x, y\rangle^d + \binom{d}{1} c\langle x, y\rangle^{d-1} + \binom{d}{2} c^2\langle x, y\rangle^{d-2} + \ldots + c^d \tag{30}$$

sum / product of p.d. kernels are p.d. kernels, thus polynomial kernel is positive definite.

RKHS of Polynomial Kernel

Let $\mathcal{H}_K$ be the RKHS corresponding the polynomial kernel, then $\mathcal{H}_K$ is $d+1$ dimensional vector space with a basis $\{1, x, x^2, \ldots, x^d\}$.

Gaussian Kernel and Its Related RKHS

Gaussian Kernel

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{31}$$

To prove gaussian kernel is positive definite, we can expand the kernel function as:

$$\begin{aligned} K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) &= \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\exp\left(\frac{\langle x,y\rangle}{\sigma^2}\right)\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\langle\phi(x),\phi(y)\rangle\exp\left(-\frac{\|y\|^2}{2\sigma^2}\right) = \langle\psi(x),\psi(y)\rangle \end{aligned} \tag{32}$$

thus gaussian kernel is positive definite.

Gaussian Kernel and Its Related RKHS

RKHS of Gaussian Kernel

The norm of gaussian kernel can be written as:

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{2\pi^d} \int \left|\hat{f}(\omega)\right|^2 \exp(\frac{\sigma^2}{2}\omega^2)\mathrm{d}\omega \tag{33}$$

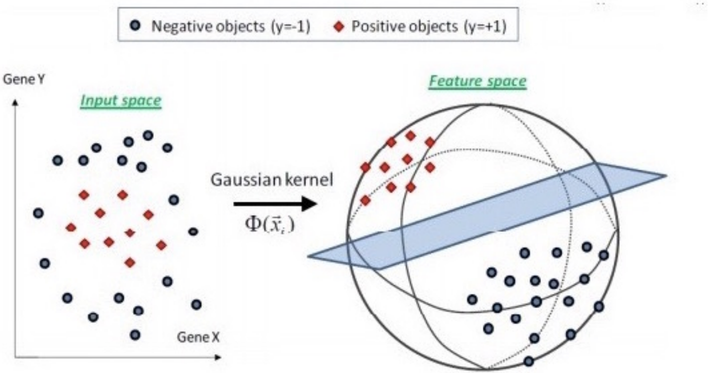
where $\hat{f}(\omega) = \mathcal{F}\{f\}(\omega) = \int_{-\infty}^{\infty} f(t)\exp(-i\omega t)\mathrm{d}t$.
So we know the RKHS of gaussian kernel is:

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}, \mathrm{d}x) : \int \left|\hat{f}(\omega)\right|^2 \exp(\frac{\sigma^2}{2}\omega^2)\mathrm{d}\omega \right\} \tag{34}$$

with inner product:

$$\langle f, g\rangle = \int \hat{f}(\omega)\hat{g}(\omega)\exp(\frac{\sigma^2}{2}\omega^2)\mathrm{d}\omega \tag{35}$$

Gaussian Kernel and Its Related RKHS



Principle Components

The principal components of a set of data in $\mathbb{R}^p$ provide a sequence of best linear approximations to that data, of all ranks $q \leq p$.
Denote the observations by $x_1, x_2, \ldots, x_N$, and consider the rank-q linear model for representing them:

$$f(\lambda) = \mu + \boldsymbol{V}_q \lambda \tag{36}$$

where μ is a location vector in $\mathbb{R}^p$, $\boldsymbol{V}_q$ is a $p \times p$ matrix with q orthogonal unit vectors as columns. Fit a model by least squares amounts:

$$\min_{\mu, \{\lambda_i\}, \boldsymbol{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \boldsymbol{V}_q \lambda_i\|^2 \tag{37}$$

Partially optimize (37) for μ and the λ_i, we obtain:

$$\hat{\mu} = \bar{x} \tag{38}$$

$$\hat{\lambda}_i = \boldsymbol{V}_q^{\mathrm{T}} (x_i - \bar{x}) \tag{39}$$

Principle Components

This leaves us to find the orthogonal matrix $\boldsymbol{V}_q$:

$$\min_{\boldsymbol{V}_q} \sum_{i=1}^{N} \left\|(x_i - \bar{x}) - \boldsymbol{V}_q \boldsymbol{V}_q^{\mathrm{T}}(x_i - \bar{x})\right\|^2 \tag{40}$$

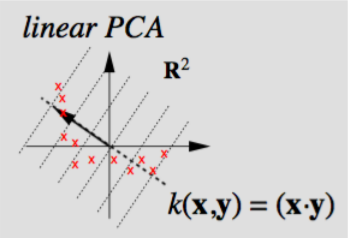
The solution can be expressed as follows. We construct the singular value decomposition of $\boldsymbol{X}$:

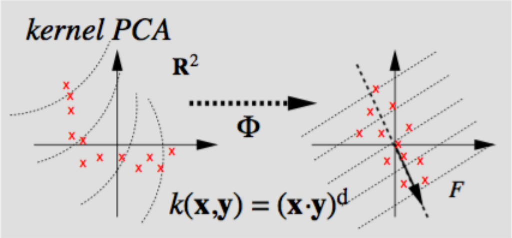
$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}} \tag{41}$$

where $\boldsymbol{U}$ is an $N \times p$ orthogonal matrix, $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix, and $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ known as the singular values.

For each rank q, the solution $\boldsymbol{V}_q$ to (40) consists of the first q columns of $\boldsymbol{V}$. The columns of $\boldsymbol{U}\boldsymbol{D}$ are called the principal components of $\boldsymbol{X}$. The N optimal $\hat{\lambda}_i$ in (39) are given by the first q principal components.

PCA v.s. Kernel PCA





Kernel PCA

The principal components variables $\boldsymbol{Z}$ of a data matrix $\boldsymbol{X}$ can be computed from the inner-product matrix $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$. We assume $\boldsymbol{X}$ is centered.

$$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\mathrm{T}} \tag{42}$$

and then $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$.
Kernel PCA simply mimics this procedure, interpreting the kernel matrix $\boldsymbol{K} = \{K(x_i, x_j)\}$ as an inner-product matrix of the implicit features $\langle \phi(x_i), \phi(x_j) \rangle$ and finding its eigenvectors. The elements of the mth component $\boldsymbol{z}_m$ can be written as $z_{im} = \sum_{j=1}^{N} \alpha_{jm} K(x_i, x_j)$, where $\alpha_{jm} = u_{jm}/d_m$. Now we will show the process of derivation.

Kernel PCA

Assume that our data mapped into feature space, $\phi(x_1), \ldots, \phi(x_N)$, is centered, i.e. $\sum_{k=1}^{N} \phi(x_k) = 0$. To do PCA for the covariance matrix:

$$C = \sum_{i=1}^{N} \phi(x_i)\phi(x_i)^{\mathrm{T}} \tag{43}$$

we have to find eigenvalues $\lambda \geq 0$ and eigenvectors $\boldsymbol{V} \in \mathcal{F} \setminus \{0\}$ satisfying $\lambda \boldsymbol{V} = C\boldsymbol{V}$. Note that all solutions $\boldsymbol{V}$ lie in the span of $\phi(x_1), \ldots, \phi(x_N)$, we may consider the equivalent system:

$$\lambda(\phi(x_k) \cdot \boldsymbol{V}) = (\phi(x_k) \cdot C\boldsymbol{V}) \quad \text{for all } k = 1, \ldots, N. \tag{44}$$

and there exist coefficients $\alpha_1, \ldots, \alpha_N$ such that:

$$\boldsymbol{V} = \sum_{i=1}^{N} \alpha_i \phi(x_i) \tag{45}$$

Kernel PCA

We arrive at:

$$N\lambda \boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{K}^2\boldsymbol{\alpha} \tag{46}$$

To find solutions of (46), we solve the eigenvalue problem:

$$N\lambda\boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha} \tag{47}$$

for nonzero eigenvalues. Clearly, all solutions of (47) do satisfy (46). We normalize the solutions $\boldsymbol{\alpha}^k$ so that $(\boldsymbol{V}^k \cdot \boldsymbol{V}^k) = 1$. We have:

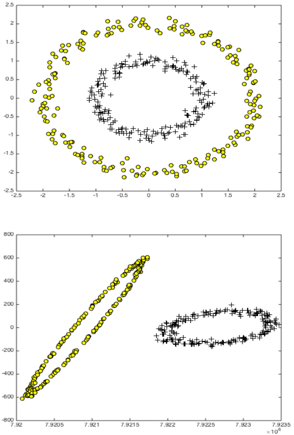
$$1 = \sum_{i,j=1}^{N} \alpha_i^k \alpha_j^k (\phi(x_i) \cdot \phi(x_j)) = (\boldsymbol{\alpha}^k \cdot \boldsymbol{K}\boldsymbol{\alpha}^k) = \lambda_k(\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k) \tag{48}$$

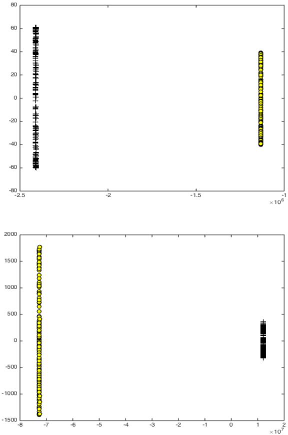
For principle component extraction, we compute projections of the image of a test point $\phi(x)$ onto the eigenvectors $\boldsymbol{V}^k$ in $\mathcal{F}$ accordance to:

$$(\boldsymbol{V}^k \cdot \phi(x)) = \sum_{i=1}^{N} \alpha_i^k (\phi(x_i) \cdot \phi(x)) \tag{49}$$

Kernel PCA

From upper left to lower right:
Input data, Gaussian Kernel, Linear Kernel, Polynomial Kernel.





Support Vector Classifier

The support vector classifier can be expressed as a regularization problem:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to } \xi_i \geq 0, y_i(x_i^{\mathrm{T}}\beta + \beta_0) \geq 1 - \xi_i, \forall i \tag{50}$$

here the linear decision boundary is $x^{\mathrm{T}}\beta + \beta_0 = 0$, which bound the maximal margin of width $2M = 2/\|\beta\|$. Some points are on the wrong side of their margin by ξ_i; points on the correct side have $\xi_i = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq \mathit{constant}$.

The Lagrange Function

The Lagrangian dual object function

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^{\mathrm{T}} x_j$$
$$\text{subject to } 0 \leq \alpha_i \leq C, \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{51}$$

the solution for β has the form

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \tag{52}$$

Kernel SVM

Transform the linear boundary to nonlinear boundary is direct, all we need to do is transform feature vectors from x_i to $h(x_i)$.
The Lagrange dual function (51) has the form

$$\begin{aligned} L_D &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle \\ &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \end{aligned} \tag{53}$$

The solution function $f(x)$ can be written as

$$f(x) = h(x)^{\mathrm{T}} \beta + \beta_0 = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + \beta_0 \tag{54}$$

Kernel SVM

With $f(x) = h(x)^{\mathrm{T}}\beta + \beta_0$, we have

$$\begin{cases} \xi_i \geq 0 \\ y_i f(x_i) \geq 1 - \xi_i \end{cases} \quad \forall i \tag{55}$$

Therefore, the optimization problem (50) can be rewritten as the form

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}[1 - y_i f(x_i)]_+ \tag{56}$$

where the subscript "+" indicates positive part. With $\lambda = 1/C$, we have

$$\min_{\beta,\beta_0} \sum_{i=1}^{N}[1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \tag{57}$$

Reproducing Kernels in SVM

Suppose the basis h arises from the eigen-expansion of a positive definite kernel K,

$$K(x, y) = \sum_{j=1}^{\infty} \gamma_j \phi_j(x) \phi_j(y) \tag{58}$$

The theory of reproducing kernel Hilbert spaces described there guarantees a finite-dimensional solution of the form

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i) \tag{59}$$

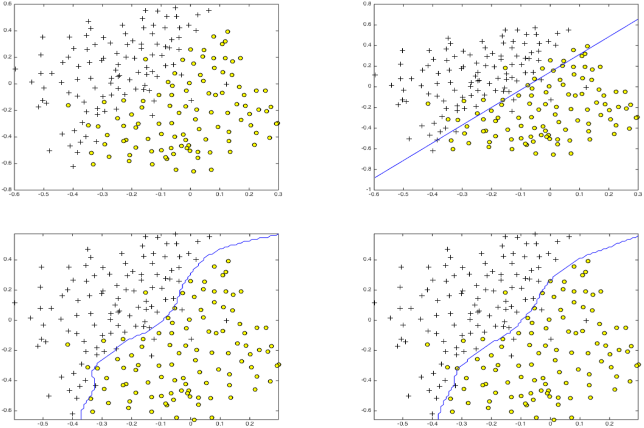
And the optimization criterion of (57) is

$$\min_{\beta_0, \theta} \sum_{i=1}^{N} (1 - y_i f(x_i))_+ + \frac{\lambda}{2} \alpha^{\mathrm{T}} \boldsymbol{K} \alpha \tag{60}$$

where $\boldsymbol{K}$ is the $N \times N$ matrix of kernel evaluations for all pairs of training features.

Training SVM with Different Kernels

From upper left to lower right:
Input data, Linear Kernel, Gaussian Kernel, Log Kernel.

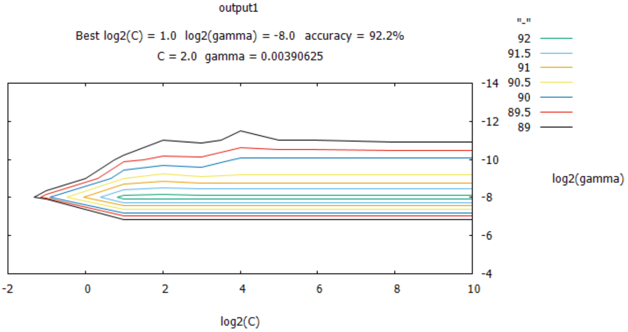


MNIST database

The MNIST database of handwritten digits has a training set of 60,000 examples, and a test set of 10,000 examples.

Preprocess the Data

```
> train_sparse = sparse(train_images_scale);
> libsvmwrite('train',train_labels,train_sparse);
> subset.py train 1000 output1 output2
> grid.py -log2c -2,10,1 -log2g -4,-14,-1 output1
```



Train the Model

```
% linear model
> model1 = svmtrain(train_labels, train_images_scale, ...
          '-s 0 -t 0 -c 2');
% gaussian model
> model2 = svmtrain(train_labels, train_images_scale,...
            '-s 0 -t 2 -g 0.0039 -c 2');
> [predicted_label, accuracy, decision_values] = ...
      svmpredict(test_labels, test_images_scale, model);
> index = find(test_labels ~= predicted_label);
> display_network(test_images(index(1:100),:)');
```

Wrongly-matched Digits

(g) No Kernel

(h) Gaussian Kernel

In conclusion, we have 5.84% of predicting error with linear SVM, and only have 1.65% of predicting error with kernel SVM.

Main References

Trevor Hastie and Robert Tibshirani and Jerome Friedman
The Elements of Statistical Learning (2009)
Springer-Verlag 167–173, 417–455, 547–550

Wahba Grace
Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV (1999)
Advances in Kernel Methods 69–88

Bernhard Schölkopf and Alexander Smola and Klaus-Robert Müller
Kernel Principal Component Analysis (2006)
Lecture Notes in Computer Science 1327(6) 583–588

Andrew Ng
Machine Learning Course, UFLDL Tutorial

The End.