Learning with Positive Definite Kernels: Theory, Algorithms and Applications

Bharath K. Sriperumbudur* & Dougal J. Sutherland†

*Department of Statistics, Pennsylvania State University

†Gatsby Unit, University College London

Data Science Summer School
École Polytechnique
June 2019

Course Outline

- Introduction: Positive definite kernels and RKHS (Lecture 1)
  - Feature space vs. Function space
  - Kernel trick
  - Applications: Ridge regression, Principal component analysis
- Generalization of kernel trick to probabilities (Lecture 2)
  - Hilbert space embedding of probabilities
  - Mean element and covariance operator
  - Applications: Two-sample testing, GAN
- Approximate Kernel Methods (Lecture 3)
  - Computational vs. Statistical trade-off
  - Applications: Ridge regression, Principal component analysis

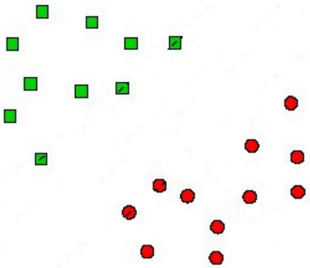
Lecture-1 Outline

- Motivating Examples
  - Nonlinear classification
  - Statistical learning
- Feature space vs. Function space
  - Kernels and properties
  - RKHS and properties
- Applications: Ridge regression, Principal component analysis
  - Kernel trick
  - Representer theorem

Motivating Example: Binary Classification

- Given: $D := \{(x_j, y_j)\}_{j=1}^{n}$, $x_j \in \mathcal{X}$, $y_j \in \{-1, +1\}$
- Goal: Learn a function $f : \mathcal{X} \to \mathbb{R}$ such that

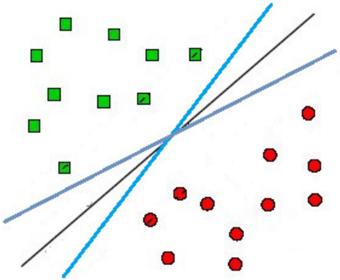
$$y_j = \mathrm{sign}(f(x_j)), \; \forall j = 1, \ldots, n.$$



Linear Classifiers

- Linear classifier: $f_{w,b}(x) = \langle w, x \rangle_2 + b$, $w, x \in \mathbb{R}^d$, $b \in \mathbb{R}$
- Find $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y_j \left(\langle w, x_j \rangle_2 + b \right) \geq 0, \; \forall j = 1, \ldots, n.$$

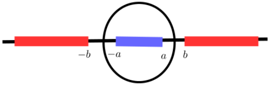


- Fisher discriminant analysis, Support vector machine, Perceptron, ...

Nonlinear Classification: 1

- There is no linear classifier that separates red and blue regions.

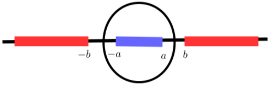
Nonlinear Classification: 1



- There is no linear classifier that separates red and blue regions.
- However, the following function perfectly separates red and blue regions

$$f(x) = x^2 - r = \left\langle \underbrace{(1, -r)}_{w}, \underbrace{(x^2, 1)}_{\Phi(x)} \right\rangle_2, \ a < r < b.$$

Nonlinear Classification: 1

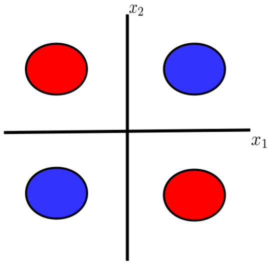


- There is no linear classifier that separates red and blue regions.
- However, the following function perfectly separates red and blue regions

$$f(x) = x^2 - r = \Big\langle \underbrace{(1, -r)}_{w}, \underbrace{(x^2, 1)}_{\Phi(x)} \Big\rangle_2 , \; a < r < b.$$

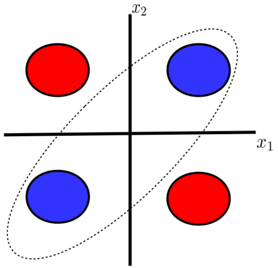
- By mapping $x \in \mathbb{R}$ to $\Phi(x) = (x^2, 1) \in \mathbb{R}^2$, the nonlinear classification problem is turned into a linear problem.
- We call Φ as the feature map (starting point of kernel trick)

Nonlinear Classification: 2



- There is no linear classifier that separates red and blue regions.

Nonlinear Classification: 2



- There is no linear classifier that separates red and blue regions.
- A conic section, however, perfectly separates them

$$
\begin{aligned}
f(x) &= ax_1^2 + bx_1x_2 + cx_2^2 + dx_1 + ex_2 + g \\
&= \left\langle \underbrace{(a, b, c, d, e, g)}_{w}, \underbrace{(x_1^2, x_1x_2, x_2^2, x_1, x_2, 1)}_{\Phi(x)} \right\rangle_2 .
\end{aligned}
$$

- $\Phi(x) \in \mathbb{R}^6$.

Motivating Example: Statistical Learning

- Given: A set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs drawn independently from an unknown probability distribution $\mathbf{P}$ on $X \times Y$.

- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.

- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).

- Loss function: $L : Y \times Y \to [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$

- One common quality measure is the average loss or expected loss of f, called the risk functional i.e.,

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x)) \, d\mathbf{P}(x, y).$$

Motivating Example: Statistical Learning

- Given: A set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs drawn independently from an unknown probability distribution $\mathbf{P}$ on $X \times Y$.
- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.
- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).
- Loss function: $L : Y \times Y \to [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$
- One common quality measure is the average loss or expected loss of f, called the risk functional i.e.,

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x))\, d\mathbf{P}(x, y).$$

Motivating Example: Statistical Learning

- Given: A set $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of input/output pairs drawn independently from an unknown probability distribution $\mathbf{P}$ on $X \times Y$.
- Goal: "Learn" a function $f : X \to Y$ such that $f(x)$ is a good approximation of the possible response y for an arbitrary x.
- We need a means to assess the quality of an estimated response $f(x)$ when the true input and output pair is (x, y).
- Loss function: $L : Y \times Y \to [0, \infty)$
  - Squared-loss: $L(y, f(x)) = (y - f(x))^2$
  - Hinge-loss: $L(y, f(x)) = \max(0, 1 - yf(x))$
- One common quality measure is the average loss or expected loss of f, called the risk functional i.e.,

$$\mathcal{R}_{L,\mathbf{P}}(f) := \int_{X \times Y} L(y, f(x)) \, d\mathbf{P}(x, y).$$

Bayes Risk and Bayes Function

- Idea: Choose f that has the smallest risk.

$$f^* := \arg \inf_{f: X \to \mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f),$$

where the infimum is taken over the set of all measurable functions.

- f^* is called the Bayes function and $\mathcal{R}_{L,\mathbf{P}}(f^*)$ is called the Bayes risk.
- If $\mathbf{P}$ is known, finding f^* is often a relatively easy task and there is nothing to learn.
  - Example: $L(y, f(x)) = (y - f(x))^2$ and $L(y, f(x)) = |y - f(x)|$
  - Exercise: What is f^* for the above losses?

Bayes Risk and Bayes Function

- Idea: Choose f that has the smallest risk.

$$f^* := \arg \inf_{f:X\to\mathbb{R}} \mathcal{R}_{L,\mathbf{P}}(f),$$

where the infimum is taken over the set of all measurable functions.

- f^* is called the Bayes function and $\mathcal{R}_{L,\mathbf{P}}(f^*)$ is called the Bayes risk.
- If **P** is known, finding f^* is often a relatively easy task and there is nothing to learn.
  - Example: $L(y, f(x)) = (y - f(x))^2$ and $L(y, f(x)) = |y - f(x)|$
  - Exercise: What is f^* for the above losses?

Universal Consistency

- But $\mathbf{P}$ is unknown.
- However "partially known" from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*).$$

- Universally consistent learning algorithm: for all $\mathbf{P}$ on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$

in probability.

Universal Consistency

- But $\mathbf{P}$ is unknown.
- However "partially known" from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that
$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*).$$
- Universally consistent learning algorithm: for all $\mathbf{P}$ on $X \times Y$, we have
$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$
in probability.

Universal Consistency

- But **P** is unknown.
- However "partially known" from the training set, $D := \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Given D, the goal is to construct $f_D : X \to \mathbb{R}$ such that

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \approx \mathcal{R}_{L,\mathbf{P}}(f^*).$$

- Universally consistent learning algorithm: for all **P** on $X \times Y$, we have

$$\mathcal{R}_{L,\mathbf{P}}(f_D) \to \mathcal{R}_{L,\mathbf{P}}(f^*), \quad n \to \infty$$

in probability.

Empirical Risk Minimization

- Since $\mathbf{P}$ is unknown but is known through D, it is tempting to replace $\mathcal{R}_{L,\mathbf{P}}(f)$ by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

called the empirical risk and find f_D by

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- Is it a good idea?
- No! Choose f_D such that $f_D(x) = y_i,\ x = x_i,\ \forall\, i$ and $f_D(x) = 0,$ *otherwise*.
- $\mathcal{R}_{L,D}(f_D) = 0$ but can be very far from $\mathcal{R}_{L,\mathbf{P}}(f^*)$.

Overfitting!!

Empirical Risk Minimization

- Since $\mathbf{P}$ is unknown but is known through D, it is tempting to replace $\mathcal{R}_{L,\mathbf{P}}(f)$ by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

called the empirical risk and find f_D by

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- Is it a good idea?

- No! Choose f_D such that $f_D(x) = y_i,\ x = x_i,\ \forall\, i$ and $f_D(x) = 0,\ \textit{otherwise}$.

- $\mathcal{R}_{L,D}(f_D) = 0$ but can be very far from $\mathcal{R}_{L,\mathbf{P}}(f^*)$.

Overfitting!!

Method of Sieves (Structural Risk Minimization)

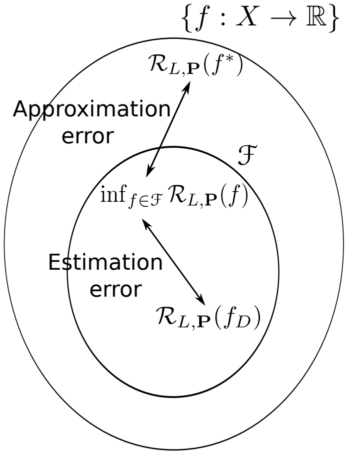
- How to avoid overfitting: Perform ERM on a small set $\mathcal{F}$ of functions $f : X \to Y$ (class of smooth functions) where the size of $\mathcal{F}$ grows appropriately with n.
- Do minimization over $\mathcal{F}$:

$$f_D := \arg \inf_{f \in \mathcal{F}} \mathcal{R}_{L,D}(f)$$

- Total error: Define $\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} := \inf_{f \in \mathcal{F}} \mathcal{R}_{L,\mathbf{P}}(f)$

$$\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P}} = \overbrace{\mathcal{R}_{L,\mathbf{P}}(f_D) - \mathcal{R}^*_{L,\mathbf{P},\mathcal{F}}}^{\text{Estimation error}} + \overbrace{\mathcal{R}^*_{L,\mathbf{P},\mathcal{F}} - \mathcal{R}^*_{L,\mathbf{P}}}^{\text{Approximation error}}$$

Approximation and Estimation Errors



How to choose $\mathcal{F}$?

$$f_D = \arg\inf_{f\in\mathcal{F}} \mathcal{R}_{L,D}(f) = \arg\inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, \underbrace{f(x_i)}_{\delta_{x_i}(f)})$$

- An <u>evaluation functional</u> is a linear functional δ_x that evaluates each function in the space at the point x, i.e.,

$$\delta_x(f) = f(x),\ \forall f \in \mathcal{F}.$$

- Bounded evaluation functional: An evaluation functional is bounded if there exists a M such that

$$|\delta_x(f)| = |f(x)| \leq M_x \|f\|_{\mathcal{F}},\ \forall x, \in \mathcal{X},\ f \in \mathcal{F}$$

  where $\mathcal{F}$ is a normed vector space (continuity of δ_x).

- Evaluation functionals are not always bounded.
- Example: $L^2[a,b]$
  - $\|f\|_2$ remains the same if f is changed at a countable set of points.

Choice of $\mathcal{F}$

- Various choices for $\mathcal{F}$ (with evaluation functional bounded):
  - Lipschitz functions
  - Bounded Lipschitz functions
  - Bounded continuous functions
- If $\mathcal{F}$ is a Hilbert space of functions with bounded evaluation functionals for all $x \in \mathcal{X}$, computationally efficient estimators can be obtained.

Reproducing Kernel Hilbert Space

Choice of $\mathcal{F}$

- Various choices for $\mathcal{F}$ (with evaluation functional bounded):
  - Lipschitz functions
  - Bounded Lipschitz functions
  - Bounded continuous functions
- If $\mathcal{F}$ is a Hilbert space of functions with bounded evaluation functionals for all $x \in \mathcal{X}$, computationally efficient estimators can be obtained.

Reproducing Kernel Hilbert Space

Summary

Points of view:

- Feature map, Φ: trick to achieve non-linear methods from linear ones
- Function space, $\mathcal{F}$: statistical generalization and computational efficiency

History

- Mathematics (Functional analysis): Introduced in 1907 by Stanisław Zaremba for studying boundary value problems; developed by Mercer, Szegö, Bergman, Bochner, Moore, Aronszajn; reached maturity by late 1950's.
- Statistics: Started by Emmanuel Parzen (early 1960's) and pursued by Wahba (between 1970 and 1990).
- Pattern recognition/Machine learning: Started by Aizerman, Braverman and Rozonoer (1964) but fury of activity following the work of Boser, Guyon and Vapnik (1992).

Other areas: Signal processing, control, probability theory, stochastic processes, numerical analysis

Kernels

(Feature space view point)

Hilbert Space

Inner product: Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A map $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

- Linear in the first argument: for any $f_1, f_2, g \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$

$$\langle \alpha f_1 + \beta f_2, g \rangle_{\mathcal{H}} = \alpha \langle f_1, g \rangle_{\mathcal{H}} + \beta \langle f_2, g \rangle_{\mathcal{H}};$$

- Symmetric: for any $f, g \in \mathcal{H}$,

$$\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}};$$

- Positive definiteness: for any $f \in \mathcal{H}$,

$$\langle f, f \rangle_{\mathcal{H}} \geq 0 \quad \text{and} \quad \langle f, f \rangle_{\mathcal{H}} = 0 \Leftrightarrow f = 0.$$

Define $\| \cdot \|_{\mathcal{H}} := \langle \cdot, \cdot \rangle_{\mathcal{H}}$ as the norm on $\mathcal{H}$ induced by the inner product.

A complete (by adding the limits of all Cauchy sequences w.r.t. $\| \cdot \|_{\mathcal{H}}$) inner product space is defined as a Hilbert space.

Measure of similarity

Hilbert Space

Inner product: Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A map $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an inner product on $\mathcal{H}$ if

- Linear in the first argument: for any $f_1, f_2, g \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$

$$\langle \alpha f_1 + \beta f_2, g \rangle_{\mathcal{H}} = \alpha \langle f_1, g \rangle_{\mathcal{H}} + \beta \langle f_2, g \rangle_{\mathcal{H}};$$

- Symmetric: for any $f, g \in \mathcal{H}$,

$$\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}};$$

- Positive definiteness: for any $f \in \mathcal{H}$,

$$\langle f, f \rangle_{\mathcal{H}} \geq 0 \quad \text{and} \quad \langle f, f \rangle_{\mathcal{H}} = 0 \Leftrightarrow f = 0.$$

Define $\| \cdot \|_{\mathcal{H}} := \langle \cdot, \cdot \rangle_{\mathcal{H}}$ as the norm on $\mathcal{H}$ induced by the inner product.

A complete (by adding the limits of all Cauchy sequences w.r.t. $\| \cdot \|_{\mathcal{H}}$) inner product space is defined as a Hilbert space.

Measure of similarity

Kernel

(Steinwart and Christmann, 2008)

Throughout, we assume that $\mathcal{X}$ is a non-empty set (input space)

Kernel: A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall x, x' \in \mathcal{H}.$$

Φ: Feature map and $\mathcal{H}$: Feature space

Non-uniqueness of Φ and $\mathcal{H}$: Suppose $k(x, x') = xx'$, $x, x' \in \mathbb{R}$. Then

$$\Phi_1(x) = x \quad \text{and} \quad \Phi_2(x) = \frac{1}{2}(x, x)$$

are feature maps with corresponding feature spaces being $\mathbb{R}$ and $\mathbb{R}^2$.

Kernel

(Steinwart and Christmann, 2008)

Throughout, we assume that $\mathcal{X}$ is a non-empty set (input space)

Kernel: A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$k(x, x') := \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}, \quad \forall x, x' \in \mathcal{H}.$$

Φ: Feature map and $\mathcal{H}$: Feature space

Non-uniqueness of Φ and $\mathcal{H}$: Suppose $k(x, x') = xx'$, $x, x' \in \mathbb{R}$. Then

$$\Phi_1(x) = x \quad \text{and} \quad \Phi_2(x) = \frac{1}{2}(x, x)$$

are feature maps with corresponding feature spaces being $\mathbb{R}$ and $\mathbb{R}^2$.

Properties

- For any $\alpha > 0$, αk is a kernel.

$$\alpha k(x, x') = \alpha \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \langle \sqrt{\alpha}\Phi(x), \sqrt{\alpha}\Phi(x') \rangle_{\mathcal{H}}.$$

- Conic sum of kernels is a kernel: If $(k_i)_{i=1}^m$ is a collection of kernels, then for any $(\alpha_i)_{i=1}^m \subset \mathbb{R}^+$, $\sum_{i=1}^m \alpha_i k_i$ is a kernel.

$$\begin{aligned}\sum_{i=1}^m \alpha_i k_i(x, x') = \sum_{i=1}^m \alpha_i \langle \Phi_i(x), \Phi_i(x') \rangle_{\mathcal{H}_i} &= \sum_{i=1}^m \langle \sqrt{\alpha_i}\Phi_i(x), \sqrt{\alpha_i}\Phi_i(x') \rangle_{\mathcal{H}_i} \\ &= \langle \tilde{\Phi}(x), \tilde{\Phi}(x') \rangle_{\tilde{\mathcal{H}}}\end{aligned}$$

for all $x, x' \in \mathcal{X}$ where

$$\tilde{\Phi}(x) = (\sqrt{\alpha_1}\Phi_1(x), \ldots, \sqrt{\alpha_m}\Phi_m(x)) \quad \text{and} \quad \tilde{\mathcal{H}} = \underbrace{\mathcal{H}_1 \oplus \ldots \oplus \mathcal{H}_m}_{\text{direct sum}}.$$

$(\mathbb{R} \oplus \mathbb{R} = \mathbb{R}^2)$

Properties

- For any $\alpha > 0$, αk is a kernel.

$$\alpha k(x,x') = \alpha\langle\Phi(x),\Phi(x')\rangle_{\mathcal{H}} = \langle\sqrt{\alpha}\Phi(x),\sqrt{\alpha}\Phi(x')\rangle_{\mathcal{H}}.$$

- Conic sum of kernels is a kernel: If $(k_i)_{i=1}^m$ is a collection of kernels, then for any $(\alpha_i)_{i=1}^m \subset \mathbb{R}^+$, $\sum_{i=1}^m \alpha_i k_i$ is a kernel.

$$\begin{aligned}\sum_{i=1}^m \alpha_i k_i(x,x') = \sum_{i=1}^m \alpha_i\langle\Phi_i(x),\Phi_i(x')\rangle_{\mathcal{H}_i} &= \sum_{i=1}^m\langle\sqrt{\alpha_i}\Phi_i(x),\sqrt{\alpha_i}\Phi_i(x')\rangle_{\mathcal{H}_i}\\ &= \langle\tilde{\Phi}(x),\tilde{\Phi}(x')\rangle_{\tilde{\mathcal{H}}}\end{aligned}$$

for all $x,x' \in \mathcal{X}$ where

$$\tilde{\Phi}(x) = (\sqrt{\alpha_1}\Phi_1(x),\ldots,\sqrt{\alpha_m}\Phi_m(x)) \quad \text{and} \quad \tilde{\mathcal{H}} = \underbrace{\mathcal{H}_1\oplus\ldots\oplus\mathcal{H}_m}_{\text{direct sum}}.$$

$(\mathbb{R}\oplus\mathbb{R} = \mathbb{R}^2)$

Properties

- For any $\alpha > 0$, αk is a kernel.

$$\alpha k(x,x') = \alpha\langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} = \langle \sqrt{\alpha}\Phi(x), \sqrt{\alpha}\Phi(x')\rangle_{\mathcal{H}}.$$

- Conic sum of kernels is a kernel: If $(k_i)_{i=1}^m$ is a collection of kernels, then for any $(\alpha_i)_{i=1}^m \subset \mathbb{R}^+$, $\sum_{i=1}^m \alpha_i k_i$ is a kernel.

$$\begin{aligned}\sum_{i=1}^m \alpha_i k_i(x,x') &= \sum_{i=1}^m \alpha_i \langle \Phi_i(x), \Phi_i(x')\rangle_{\mathcal{H}_i} = \sum_{i=1}^m \langle \sqrt{\alpha_i}\Phi_i(x), \sqrt{\alpha_i}\Phi_i(x')\rangle_{\mathcal{H}_i} \\ &= \langle \tilde{\Phi}(x), \tilde{\Phi}(x')\rangle_{\tilde{\mathcal{H}}}\end{aligned}$$

for all $x, x' \in \mathcal{X}$ where

$$\tilde{\Phi}(x) = (\sqrt{\alpha_1}\Phi_1(x), \ldots, \sqrt{\alpha_m}\Phi_m(x)) \quad \text{and} \quad \tilde{\mathcal{H}} = \underbrace{\mathcal{H}_1 \oplus \ldots \oplus \mathcal{H}_m}_{\text{direct sum}}.$$

$(\mathbb{R} \oplus \mathbb{R} = \mathbb{R}^2)$

Properties

- Difference of kernels is NOT a kernel:
  - Suppose $\exists\, x \in \mathcal{X}$ such that $k_1(x,x) - k_2(x,x) < 0$.
  - If $k_1 - k_2$ is a kernel, then $\exists\, \Phi$ and $\mathcal{H}$ such that for all $x, x' \in \mathcal{H}$,

$$k_1(x,x') - k_2(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$

  - Choose $x = x'$.

- Product of kernels is a kernel: If k_1 and k_2 are kernels, then $k_1 \cdot k_2$ is a kernel.

$$\begin{aligned} k((x_1,x_2),(x_1',x_2')) &= k_1(x_1,x_1') \cdot k_2(x_2,x_2') \\ &= \langle \Phi_1(x_1), \Phi_1(x_1') \rangle_{\mathcal{H}_1} \cdot \langle \Phi_2(x_2), \Phi_2(x_2') \rangle_{\mathcal{H}_2} \\ &= \langle \Phi_1(x_1) \otimes \Phi_2(x_2), \Phi_1(x_1') \otimes \Phi_2(x_2') \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \end{aligned}$$

where $\otimes$ denotes the tensor product.

Properties

- Difference of kernels is NOT a kernel:
  - Suppose $\exists\, x \in \mathcal{X}$ such that $k_1(x,x) - k_2(x,x) < 0$.
  - If $k_1 - k_2$ is a kernel, then $\exists\, \Phi$ and $\mathcal{H}$ such that for all $x, x' \in \mathcal{H}$,

$$k_1(x,x') - k_2(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$

  - Choose $x = x'$.

- Product of kernels is a kernel: If k_1 and k_2 are kernels, then $k_1 \cdot k_2$ is a kernel.

$$\begin{aligned} k((x_1,x_2),(x_1',x_2')) &= k_1(x_1,x_1') \cdot k_2(x_2,x_2') \\ &= \langle \Phi_1(x_1), \Phi_1(x_1') \rangle_{\mathcal{H}_1} \cdot \langle \Phi_2(x_2), \Phi_2(x_2') \rangle_{\mathcal{H}_2} \\ &= \langle \Phi_1(x_1) \otimes \Phi_2(x_2), \Phi_1(x_1') \otimes \Phi_2(x_2') \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} \end{aligned}$$

where $\otimes$ denotes the tensor product.

Properties

- Suppose k_1 is defined on $\{0, 1\}$ and k_2 is defined on $\{A, B, C\}$. Then clearly $k_1 \cdot k_2$ is defined on $\{0, 1\} \times \{A, B, C\}$.
- Suppose for simplicity, we assume $\mathcal{H}_1 = \mathbb{R}^2$ and $\mathcal{H}_2 = \mathbb{R}^5$. Then

$$\begin{aligned} k_1(x_1, x_1') \cdot k_2(x_2, x_2') &= \langle \Phi_1(x_1), \Phi_1(x_1') \rangle_{\mathbb{R}^2} \cdot \langle \Phi_2(x_2), \Phi_2(x_2') \rangle_{\mathbb{R}^5} \\ &= \Phi_1^\top(x_1')\Phi_1(x_1)\Phi_2^\top(x_2)\Phi_2(x_2') \\ &= \mathrm{Tr}\left(\underbrace{\Phi_2(x_2')\Phi_1^\top(x_1')}_{\mathbb{R}^2 \to \mathbb{R}^5} \underbrace{\Phi_1(x_1)\Phi_2^\top(x_2)}_{\mathbb{R}^5 \to \mathbb{R}^2} \right) \\ &= \left\langle \Phi_1(x_1)\Phi_2^\top(x_2), \Phi_1(x_1')\Phi_2^\top(x_2') \right\rangle_{\mathbb{R}^2 \otimes \mathbb{R}^5} \\ &=: \langle \Phi_1(x_1) \otimes \Phi_2(x_2), \Phi_1(x_1') \otimes \Phi_2(x_2') \rangle_{\mathbb{R}^2 \otimes \mathbb{R}^5} \end{aligned}$$

where $\mathbb{R}^2 \otimes \mathbb{R}^5$ is the space of 2×5 matrices.

Properties

- Suppose k_1 is defined on $\{0,1\}$ and k_2 is defined on $\{A,B,C\}$. Then clearly $k_1 \cdot k_2$ is defined on $\{0,1\} \times \{A,B,C\}$.

- Suppose for simplicity, we assume $\mathcal{H}_1 = \mathbb{R}^2$ and $\mathcal{H}_2 = \mathbb{R}^5$. Then

$$
\begin{aligned}
k_1(x_1, x_1') \cdot k_2(x_2, x_2') &= \langle \Phi_1(x_1), \Phi_1(x_1') \rangle_{\mathbb{R}^2} \cdot \langle \Phi_2(x_2), \Phi_2(x_2') \rangle_{\mathbb{R}^5} \\
&= \Phi_1^\top(x_1')\Phi_1(x_1)\Phi_2^\top(x_2)\Phi_2(x_2') \\
&= \mathrm{Tr}\left(\underbrace{\Phi_2(x_2')\Phi_1^\top(x_1')}_{\mathbb{R}^2 \to \mathbb{R}^5} \underbrace{\Phi_1(x_1)\Phi_2^\top(x_2)}_{\mathbb{R}^5 \to \mathbb{R}^2} \right) \\
&= \left\langle \Phi_1(x_1)\Phi_2^\top(x_2), \Phi_1(x_1')\Phi_2^\top(x_2') \right\rangle_{\mathbb{R}^2 \otimes \mathbb{R}^5} \\
&=: \langle \Phi_1(x_1) \otimes \Phi_2(x_2), \Phi_1(x_1') \otimes \Phi_2(x_2') \rangle_{\mathbb{R}^2 \otimes \mathbb{R}^5}
\end{aligned}
$$

where $\mathbb{R}^2 \otimes \mathbb{R}^5$ is the space of 2×5 matrices.

Properties

- For any arbitrary function $f : \mathcal{X} \to \mathbb{R}$,

$$\tilde{k}(x, x') = f(x)k(x, x')f(x') \tag{1}$$

is a kernel.

$$\begin{aligned}\tilde{k}(x, x') = f(x)k(x, x')f(x') &= f(x)\langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} f(x') \\ &= \langle \underbrace{f(x)\Phi(x)}_{\Phi_f(x)}, \underbrace{f(x')\Phi(x')}_{\Phi_f(x')} \rangle_{\mathcal{H}}.\end{aligned}$$

- $k(x, x) \geq 0$: $k(x, x) = \langle \Phi(x), \Phi(x)\rangle_{\mathcal{H}} = \|\Phi(x)\|^2_{\mathcal{H}} \geq 0.$
- Cauchy-Schwartz: $|k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(x', x')}$

$$|k(x, x')| = |\langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}}| \leq \|\Phi(x)\|_{\mathcal{H}} \|\Phi(x')\|_{\mathcal{H}}.$$

Properties

- For any arbitrary function $f : \mathcal{X} \to \mathbb{R}$,

$$\tilde{k}(x, x') = f(x)k(x, x')f(x') \qquad (1)$$

is a kernel.

$$\begin{aligned}\tilde{k}(x, x') = f(x)k(x, x')f(x') &= f(x)\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} f(x') \\ &= \langle \underbrace{f(x)\Phi(x)}_{\Phi_f(x)}, \underbrace{f(x')\Phi(x')}_{\Phi_f(x')} \rangle_{\mathcal{H}}.\end{aligned}$$

- $k(x, x) \geq 0$: $k(x, x) = \langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}} = \|\Phi(x)\|^2_{\mathcal{H}} \geq 0$.
- Cauchy-Schwartz: $|k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(x', x')}$

$$|k(x, x')| = |\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}| \leq \|\Phi(x)\|_{\mathcal{H}} \|\Phi(x')\|_{\mathcal{H}}.$$

Properties

- For any arbitrary function $f : \mathcal{X} \to \mathbb{R}$,

$$\tilde{k}(x, x') = f(x)k(x, x')f(x') \tag{1}$$

is a kernel.

$$\begin{aligned}\tilde{k}(x, x') = f(x)k(x, x')f(x') &= f(x)\langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} f(x') \\ &= \langle \underbrace{f(x)\Phi(x)}_{\Phi_f(x)}, \underbrace{f(x')\Phi(x')}_{\Phi_f(x')}\rangle_{\mathcal{H}}.\end{aligned}$$

- $k(x, x) \geq 0$: $k(x, x) = \langle \Phi(x), \Phi(x)\rangle_{\mathcal{H}} = \|\Phi(x)\|^2_{\mathcal{H}} \geq 0.$
- Cauchy-Schwartz: $|k(x, y)| \leq \sqrt{k(x, x)}\sqrt{k(x', x')}$

$$|k(x, x')| = |\langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}}| \leq \|\Phi(x)\|_{\mathcal{H}}\|\Phi(x')\|_{\mathcal{H}}.$$

Properties

- Infinite dimensional feature map:

$$k(x, x') = \sum_{i \in I} \phi_i(x)\phi_i(x') \quad \text{is a kernel}$$

  if $\|(\phi_i(x))_i\|^2_{\ell_2(I)} := \sum_{i \in I} \phi_i^2(x) < \infty$ for all $x \in \mathcal{X}$.

- Proof:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

  where $\Phi(x) = (\phi_i(x))_{i \in I}$ and $\mathcal{H} = \ell_2(I)$, which is the space of square summable sequences on I.

If I is countable, then $\Phi(x)$ is infinite dimensional.

Properties

- Infinite dimensional feature map:

$$k(x, x') = \sum_{i \in I} \phi_i(x)\phi_i(x') \quad \text{is a kernel}$$

if $\|(\phi_i(x))_i\|^2_{\ell_2(I)} := \sum_{i \in I} \phi_i^2(x) < \infty$ for all $x \in \mathcal{X}$.

- Proof:

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$$

where $\Phi(x) = (\phi_i(x))_{i \in I}$ and $\mathcal{H} = \ell_2(I)$, which is the space of square summable sequences on I.

If I is countable, then $\Phi(x)$ is infinite dimensional.

Examples

- Polynomial kernel: $k(x,x') = (c + \langle x, x' \rangle_2)^m$, $x, x' \in \mathbb{R}^d$ for $c \geq 0$ and $m \in \mathbb{N}$. Use binomial theorem to expand, apply sum and product rules.
- Linear kernel: $c = 0$ and $m = 1$.
- Exponential kernel: $k(x,x') = \exp(\langle x, x' \rangle_2)$, $x, x' \in \mathbb{R}^d$.

  Use Taylor series expansion,

$$k(x,x') = \exp(\langle x, x' \rangle_2) = \sum_{i=0}^{\infty} \frac{\langle x, x' \rangle_2^i}{i!}.$$

- Gaussian kernel: $k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{\gamma^2}\right)$, $x, x' \in \mathbb{R}^d$. Note that

$$k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{\gamma^2}\right) = \frac{\exp\left(-2\frac{\langle x,x' \rangle_2}{\gamma^2}\right)}{\exp\left(-\frac{\|x\|_2^2}{\gamma^2}\right)\exp\left(-\frac{\|x'\|_2^2}{\gamma^2}\right)}$$

  and apply (1).

Examples

- Polynomial kernel: $k(x,x') = (c + \langle x, x' \rangle_2)^m$, $x, x' \in \mathbb{R}^d$ for $c \geq 0$ and $m \in \mathbb{N}$. Use binomial theorem to expand, apply sum and product rules.
- Linear kernel: $c = 0$ and $m = 1$.
- Exponential kernel: $k(x,x') = \exp(\langle x, x' \rangle_2)$, $x, x' \in \mathbb{R}^d$.

  Use Taylor series expansion,

$$k(x,x') = \exp(\langle x, x' \rangle_2) = \sum_{i=0}^{\infty} \frac{\langle x, x' \rangle_2^i}{i!}.$$

- Gaussian kernel: $k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{\gamma^2}\right)$, $x, x' \in \mathbb{R}^d$. Note that

$$k(x,x') = \exp\left(-\frac{\|x - x'\|_2^2}{\gamma^2}\right) = \frac{\exp\left(-2\frac{\langle x, x' \rangle_2}{\gamma^2}\right)}{\exp\left(-\frac{\|x\|_2^2}{\gamma^2}\right)\exp\left(-\frac{\|x'\|_2^2}{\gamma^2}\right)}$$

  and apply (1).

Examples

- Polynomial kernel: $k(x, x') = (c + \langle x, x' \rangle_2)^m$, $x, x' \in \mathbb{R}^d$ for $c \geq 0$ and $m \in \mathbb{N}$. Use binomial theorem to expand, apply sum and product rules.
- Linear kernel: $c = 0$ and $m = 1$.
- Exponential kernel: $k(x, x') = \exp(\langle x, x' \rangle_2)$, $x, x' \in \mathbb{R}^d$.

  Use Taylor series expansion,

$$k(x, x') = \exp(\langle x, x' \rangle_2) = \sum_{i=0}^{\infty} \frac{\langle x, x' \rangle_2^i}{i!}.$$

- Gaussian kernel: $k(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{\gamma^2}\right)$, $x, x' \in \mathbb{R}^d$. Note that

$$k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{\gamma^2}\right) = \frac{\exp\left(-2\frac{\langle x, x' \rangle_2}{\gamma^2}\right)}{\exp\left(-\frac{\|x\|_2^2}{\gamma^2}\right)\exp\left(-\frac{\|x'\|_2^2}{\gamma^2}\right)}$$

  and apply (1).

Positive Definiteness

- Kernels are symmetric and positive definite: EASY
  - Symmetry: $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}} = k(x', x)$
  - Positive definiteness:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.$$

- Symmetric and positive definite functions are kernels: NOT OBVIOUS

  The proof is based on the construction of a reproducing kernel Hilbert space.

In general, checking for positive definiteness is also NOT easy.

Positive Definiteness

- Kernels are symmetric and positive definite: EASY
  - Symmetry: $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}} = k(x', x)$
  - Positive definiteness:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0.$$

- Symmetric and positive definite functions are kernels: NOT OBVIOUS

  The proof is based on the construction of a reproducing kernel Hilbert space.

In general, checking for positive definiteness is also NOT easy.

Positive Definiteness

- Kernels are symmetric and positive definite: EASY
  - Symmetry: $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}} = k(x', x)$
  - Positive definiteness:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|_{\mathcal{H}}^{2} \geq 0.$$

- Symmetric and positive definite functions are kernels: NOT OBVIOUS

  The proof is based on the construction of a reproducing kernel Hilbert space.

In general, checking for positive definiteness is also NOT easy.

Positive Definiteness: Translation Invariant Kernels

Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y), \; x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x,\omega\rangle_2} \, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Given a continuous integrable function ψ, i.e., $\int_{\mathbb{R}^d} |\psi(x)| \, dx < \infty$, compute

$$\hat{\psi}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle \omega,x\rangle_2} \psi(x) \, dx.$$

If $\hat{\psi}(\omega)$ is non-negative for all $\omega \in \mathbb{R}^d$, then ψ is positive definite and $k(x, x') = \psi(x - x')$ is a kernel.

Positive Definiteness: Translation Invariant Kernels

Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y),\ x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x,\omega\rangle_2}\, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Given a continuous integrable function ψ, i.e., $\int_{\mathbb{R}^d} |\psi(x)|\, dx < \infty$, compute

$$\hat{\psi}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle \omega, x\rangle_2}\, \psi(x)\, dx.$$

If $\hat{\psi}(\omega)$ is non-negative for all $\omega \in \mathbb{R}^d$, then ψ is positive definite and $k(x, x') = \psi(x - x')$ is a kernel.

Positive Definiteness: Translation Invariant Kernels

Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y), \; x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x, \omega \rangle_2} \, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Given a continuous integrable function ψ, i.e., $\int_{\mathbb{R}^d} |\psi(x)| \, dx < \infty$, compute

$$\hat{\psi}(\omega) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-\sqrt{-1}\langle \omega, x \rangle_2} \, \psi(x) \, dx.$$

If $\hat{\psi}(\omega)$ is non-negative for all $\omega \in \mathbb{R}^d$, then ψ is positive definite and $k(x, x') = \psi(x - x')$ is a kernel.

Exercise

- Show that
$$\psi(x) = (1 - |x|)1_{[-1,1]}(x),\ x \in \mathbb{R}$$
is positive definite.

- Show that
$$\psi(x) = \frac{1}{2}(2 - |x|)^2 1_{\{(2-|x|)\in[0,1]\}} + \left(1 - \frac{x^2}{2}\right) 1_{[-1,1]}(x),\ x \in \mathbb{R}$$
is NOT positive definite.

So far...

Kernels $\Leftrightarrow$ Symmetric and positive definite functions

Reproducing Kernel Hilbert Space

(Function space view point)

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of <u>real-valued functions</u> on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \left\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \right\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.
- The evaluation functional is bounded:

$$\begin{aligned}|\delta_x(f)| = |f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| &\le \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \sqrt{k(x, x)}\|f\|_{\mathcal{H}}, \ \forall x \in \mathcal{X},\ f \in \mathcal{H}.\end{aligned}$$

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \left\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \right\rangle_{\mathcal{H}}, \quad x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.
- The evaluation functional is bounded:

$$\begin{aligned} |\delta_x(f)| = |f(x)| = |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| &\leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \sqrt{k(x, x)} \|f\|_{\mathcal{H}},\ \forall x \in \mathcal{X},\ f \in \mathcal{H}. \end{aligned}$$

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall x \in \mathcal{X},\ \ k(\cdot, x) \in \mathcal{H};$
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ \ \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x).$
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \Big\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \Big\rangle_{\mathcal{H}}, \ x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.
- The evaluation functional is bounded:

$$|\delta_x(f)| = |f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} = \sqrt{k(x, x)} \|f\|_{\mathcal{H}}, \ \forall x \in \mathcal{X}, \ f \in \mathcal{H}.$$

Reproducing Kernel Hilbert Space

- A Hilbert space $\mathcal{H}$ of <u>real-valued functions</u> on $\mathcal{X}$ is said to be a reproducing kernel Hilbert space (RKHS) with $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as the reproducing kernel, if
  - $\forall x \in \mathcal{X}, \; k(\cdot, x) \in \mathcal{H}$;
  - $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \; \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.
- The reproducing kernel (r.k.) k of $\mathcal{H}$ is a kernel:

$$k(x, x') = \Big\langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, x')}_{\Phi(x')} \Big\rangle_{\mathcal{H}}, \; x, x' \in \mathcal{X}.$$

  We refer to $\Phi(x) = k(\cdot, x)$ as the canonical feature map.
- Every r.k. is a symmetric and positive definite function.
- The evaluation functional is bounded:

$$\begin{aligned} |\delta_x(f)| = |f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| &\leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} \\ &= \sqrt{k(x, x)} \|f\|_{\mathcal{H}}, \; \forall x \in \mathcal{X}, \; f \in \mathcal{H}. \end{aligned}$$

Reproducing Kernel Hilbert Space

- Every Hilbert function space with a reproducing kernel is an RKHS.
- The converse is true: Every RKHS has a unique reproducing kernel.
- (Moore-Aronszajn Theorem)

  If k is a positive definite kernel, then there exists a unique RKHS with k as the reproducing kernel.

  (Proof: Define $H = \{f : f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ \alpha_i \in \mathbb{R},\ x_i \in \mathcal{X}\}$ endowed with the bilinear form

$$\langle f, g \rangle_H = \sum_{i,j=1}^{n} \alpha_i \beta_j k(x_i, x_j).$$

  Verify that $\langle \cdot, \cdot \rangle_H$ is an inner product and $\langle f, k(\cdot, x) \rangle_H = f(x)$ for any $f \in H$. Complete H to obtain an RKHS.)

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Reproducing Kernel Hilbert Space

- Every Hilbert function space with a reproducing kernel is an RKHS.
- The converse is true: Every RKHS has a unique reproducing kernel.
- (Moore-Aronszajn Theorem)

  If k is a positive definite kernel, then there exists a unique RKHS with k as the reproducing kernel.

  (Proof: Define $H = \{f : f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ \alpha_i \in \mathbb{R},\ x_i \in \mathcal{X}\}$ endowed with the bilinear form

  $$\langle f, g \rangle_H = \sum_{i,j=1}^{n} \alpha_i \beta_j k(x_i, x_j).$$

  Verify that $\langle \cdot, \cdot \rangle_H$ is an inner product and $\langle f, k(\cdot, x) \rangle_H = f(x)$ for any $f \in H$. Complete H to obtain an RKHS.)

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Reproducing Kernel Hilbert Space

- Every Hilbert function space with a reproducing kernel is an RKHS.
- The converse is true: Every RKHS has a unique reproducing kernel.
- (Moore-Aronszajn Theorem)

  If k is a positive definite kernel, then there exists a unique RKHS with k as the reproducing kernel.

  (Proof: Define $H = \{f : f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ \alpha_i \in \mathbb{R},\ x_i \in \mathcal{X}\}$ endowed with the bilinear form

  $$\langle f, g\rangle_H = \sum_{i,j=1}^{n} \alpha_i \beta_j k(x_i, x_j).$$

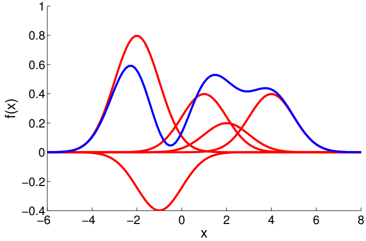
  Verify that $\langle \cdot, \cdot \rangle_H$ is an inner product and $\langle f, k(\cdot, x)\rangle_H = f(x)$ for any $f \in H$. Complete H to obtain an RKHS.)

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

Functions in the RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) \;:\; x \in \mathcal{X}\}}$ (linear span of kernel functions)
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for arbitrary $m \in \mathbb{N}$, $\{\alpha_i\} \subset \mathbb{R}$, $x \in \mathcal{X}$ and $\{x_i\} \subset \mathcal{X}$.

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$



Picture credit: A. Gretton

Properties of RKHS

- k is bounded if and only every $f \in \mathcal{H}$ is bounded.
- If $\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mu(x) < \infty$, then for every $f \in \mathcal{H}$, $\int_{\mathcal{X}} f(x)\, d\mu(x) < \infty$.
- Every $f \in \mathcal{H}$ is continuous if and only if $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$.
- Every $f \in \mathcal{H}$ is m-times continuously differentiable if k is m-times continuously differentiable.

k controls the properties of $\mathcal{H}$

Explicit Realization of RKHS

- $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$ where ψ is a positive definite function.
- Assume ψ satisfies $\int_{\mathbb{R}^d} |\psi(x)|\, dx < \infty$. Denote $\hat{\psi}$ to be the Fourier transform of ψ.
- Define $L^2(\mathbb{R}^d) := \{f : \int_{\mathbb{R}^d} |f(x)|^2\, dx < \infty\}$. Then

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^d) \,\middle|\, \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\psi}(\omega)}\, d\omega < \infty \right\}$$

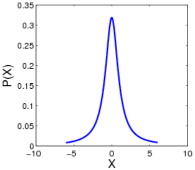
endowed with

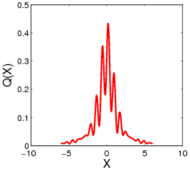
$$\langle f, g \rangle_{\mathcal{H}} = (2\pi)^{-d/2} \int \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\psi}(\omega)}\, d\omega$$

is an <u>RKHS with k as the r.k.</u>

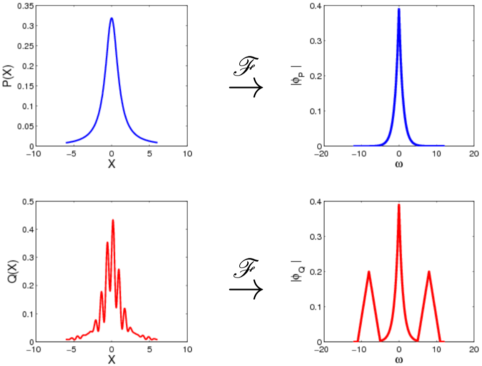
(Wendland, 2005)

Fourier Transform

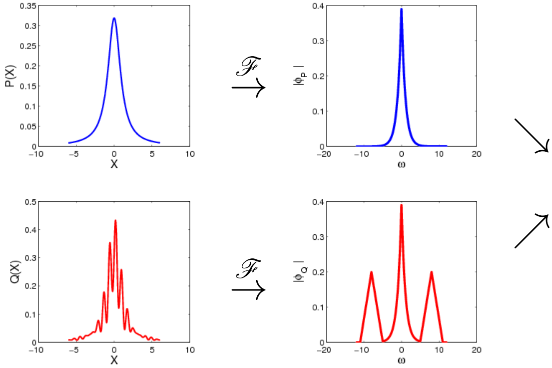




Fourier Transform



Fourier Transform



Smooth function

$\Updownarrow$

Fast rate of decay of
Fourier transform

Gaussian RKHS

- Gaussian kernel:

$$k(x,y) = \psi(x-y) = e^{-\|x-y\|_2^2/\gamma^2},\ x,y \in \mathbb{R}^d$$

- Fourier transform:

$$\hat{\psi}(\omega) = \left(\frac{\gamma^2}{2}\right)^{d/2} e^{-\frac{\gamma^2\|\omega\|_2^2}{4}},\ \omega \in \mathbb{R}^d$$

-

$$\mathcal{H}_\gamma(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) \ :\ \underbrace{\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{\frac{\gamma^2\|\omega\|_2^2}{4}}\, d\omega}_{\|f\|^2_{\mathcal{H}_\gamma}} < \infty \right\}$$

Fast decay of $\hat{\psi} \Rightarrow$ Smooth $\mathcal{H}$

Gaussian RKHS

- Gaussian kernel:

$$k(x, y) = \psi(x - y) = e^{-\|x-y\|_2^2/\gamma^2},\ x, y \in \mathbb{R}^d$$

- Fourier transform:

$$\hat{\psi}(\omega) = \left(\frac{\gamma^2}{2}\right)^{d/2} e^{-\frac{\gamma^2\|\omega\|_2^2}{4}},\ \omega \in \mathbb{R}^d$$

- $$\mathcal{H}_\gamma(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) \ :\ \underbrace{\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{\frac{\gamma^2\|\omega\|_2^2}{4}}\, d\omega}_{\|f\|^2_{\mathcal{H}_\gamma}} < \infty \right\}$$

- $\{f : \|f\|_{\mathcal{H}_\gamma} \leq \alpha\} \subset \{f : \|f\|_{\mathcal{H}_\gamma} \leq \beta\} \subset \mathcal{H}_\gamma$ for any $\alpha < \beta$.

More smoothness

Sobolev RKHS

- Laplacian kernel:

$$k(x,y) = \psi(x-y) = \sqrt{\frac{\pi}{2}}e^{-|x-y|},\ x,y \in \mathbb{R}$$

- Fourier transform:

$$\hat{\psi}(\omega) = \frac{1}{1+|\omega|^2},\ \omega \in \mathbb{R}$$

-

$$\mathcal{H}_1^2(\mathbb{R}) := \left\{ f \in L^2(\mathbb{R}) \ : \ \underbrace{\int_{\mathbb{R}} |\hat{f}(\omega)|^2(1+|\omega|^2)\,d\omega}_{\|f\|^2_{\mathcal{H}_1^2}} < \infty \right\}$$

- $\{f : \|f\|_{\mathcal{H}_1^2} \le \alpha\} \subset \{f : \|f\|_{\mathcal{H}_1^2} \le \beta\} \subset \mathcal{H}_1^2$ for any $\alpha < \beta$.

Extension to $\mathbb{R}^d$: Matérn Kernel

Summing Up

- Kernels: Feature map Φ and feature space $\mathcal{H}$
- Positive definiteness and Bochner's theorem
- RKHS: Canonical feature map $\Phi(x) = k(\cdot, x)$
- Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS
- Properties of k control the properties of the RKHS.
- Smoothness

Application: Ridge Regression

(Kernel Trick: Feature map point of view)

Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{primal}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{primal}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n} \left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right)^{-1} \mathbf{X}\mathbf{y}}_{primal}$$

- Easy:

$$\left(\frac{1}{n} \mathbf{X}\mathbf{X}^\top + \lambda I_d \right) \mathbf{X} = \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)$$

$$w = \underbrace{\frac{1}{n} \mathbf{X} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Ridge regression

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- Task: Find a linear regressor $f = \langle w, \cdot \rangle_2$ s.t. $f(x_i) \approx y_i$,

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle_2 - y_i)^2 + \lambda \|w\|_2^2 \quad (\lambda > 0)$$

- Solution: For $\mathbf{X} := (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$w = \underbrace{\frac{1}{n}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top + \lambda I_d\right)^{-1} \mathbf{X}\mathbf{y}}_{primal}$$

- Easy:

$$\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top + \lambda I_d\right)\mathbf{X} = \mathbf{X}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda I_n\right)$$

$$w = \underbrace{\frac{1}{n}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \lambda I_n\right)^{-1} \mathbf{y}}_{dual}$$

Ridge regression

- Prediction: Given $t \in \mathbb{R}^d$

$$f(t) = \langle w, t \rangle_2 = \mathbf{y}^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + n\lambda I_d \right)^{-1} t$$
$$= \mathbf{y}^\top \left(\mathbf{X}^\top \mathbf{X} + n\lambda I_n \right)^{-1} \mathbf{X}^\top t$$

- How does $\mathbf{X}^\top\mathbf{X}$ look like?

$$\mathbf{X}^\top\mathbf{X} = \underbrace{\begin{bmatrix} \langle x_1, x_1 \rangle_2 & \langle x_1, x_2 \rangle_2 & \cdots & \langle x_1, x_n \rangle_2 \\ \langle x_2, x_1 \rangle_1 & \langle x_2, x_2 \rangle_2 & \cdots & \langle x_2, x_n \rangle_2 \\ \vdots & \langle x_i, x_j \rangle_2 & \ddots & \vdots \\ \langle x_n, x_1 \rangle_1 & \langle x_n, x_2 \rangle_2 & \cdots & \langle x_n, x_n \rangle_2 \end{bmatrix}}_{\text{Matrix of inner products: Gram Matrix}}$$

Ridge regression

- Prediction: Given $t \in \mathbb{R}^d$

$$
\begin{aligned}
f(t) = \langle w, t \rangle_2 &= \mathbf{y}^\top \mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + n\lambda I_d \right)^{-1} t \\
&= \mathbf{y}^\top \left(\mathbf{X}^\top \mathbf{X} + n\lambda I_n \right)^{-1} \mathbf{X}^\top t
\end{aligned}
$$

- How does $\mathbf{X}^\top \mathbf{X}$ look like?

$$
\mathbf{X}^\top \mathbf{X} = \underbrace{\begin{bmatrix} \langle x_1, x_1 \rangle_2 & \langle x_1, x_2 \rangle_2 & \cdots & \langle x_1, x_n \rangle_2 \\ \langle x_2, x_1 \rangle_1 & \langle x_2, x_2 \rangle_2 & \cdots & \langle x_2, x_n \rangle_2 \\ \vdots & \langle x_i, x_j \rangle_2 & \ddots & \vdots \\ \langle x_n, x_1 \rangle_1 & \langle x_n, x_2 \rangle_2 & \cdots & \langle x_n, x_n \rangle_2 \end{bmatrix}}_{\text{Matrix of inner products: Gram Matrix}}
$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n}\left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})}\right)^{-1}\Phi(\mathbf{X})\mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n}\Phi(\mathbf{X})\left(\frac{1}{n}\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) + \lambda I_n\right)^{-1}\mathbf{y}}_{dual}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{primal}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{dual}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Given: $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$, $y_i \in \mathbb{R}$
- Task: Find a regressor $f \in \mathcal{H}$ (some feature space) s.t. $f(x_i) \approx y_i$.
- Idea: Map x_i to $\Phi(x_i)$ and do linear regression,

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (\langle f, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0)$$

- Solution: For $\Phi(\mathbf{X}) := (\Phi(x_1), \ldots, \Phi(x_n)) \in \mathbb{R}^{\dim(\mathcal{H}) \times n}$ and $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$,

$$f = \underbrace{\frac{1}{n} \left(\frac{1}{n} \Phi(\mathbf{X}) \Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(\mathbf{X}) \mathbf{y}}_{\textit{primal}}$$

$$= \underbrace{\frac{1}{n} \Phi(\mathbf{X}) \left(\frac{1}{n} \Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \mathbf{y}}_{\textit{dual}}$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n}\mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$
$$= \frac{1}{n}\mathbf{y}^\top \left(\frac{1}{n}\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top \Phi(t) = [\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}}]^\top$$

Kernel Ridge regression: Feature Map and Kernel Trick

- Prediction: Given $t \in \mathcal{X}$

$$f(t) = \langle f, \Phi(t) \rangle_{\mathcal{H}} = \frac{1}{n}\mathbf{y}^\top \Phi(\mathbf{X})^\top \left(\frac{1}{n}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \lambda I_{\dim(\mathcal{H})} \right)^{-1} \Phi(t)$$
$$= \frac{1}{n}\mathbf{y}^\top \left(\frac{1}{n}\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) + \lambda I_n \right)^{-1} \Phi(\mathbf{X})^\top \Phi(t)$$

As before

$$\Phi(\mathbf{X})^\top\Phi(\mathbf{X}) = \underbrace{\begin{bmatrix} \langle \Phi(x_1), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_1), \Phi(x_n) \rangle_{\mathcal{H}} \\ \langle \Phi(x_2), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_2), \Phi(x_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \Phi(x_n), \Phi(x_1) \rangle_{\mathcal{H}} & \cdots & \langle \Phi(x_n), \Phi(x_n) \rangle_{\mathcal{H}} \end{bmatrix}}_{k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}$$

and

$$\Phi(\mathbf{X})^\top \Phi(t) = [\langle \Phi(x_1), \Phi(t) \rangle_{\mathcal{H}}, \ldots, \langle \Phi(x_n), \Phi(t) \rangle_{\mathcal{H}}]^\top$$

Feature Map and Kernel Trick: Remarks

- The primal formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.
- Suppose we have access to a kernel function, k (Recall: not easy to verify that k is a kernel). Then the dual formulation is entirely determined by k (Gram matrix or kernel matrix).
- Linear regression in the dual uses a linear kernel.

Kernel trick or heuristic

Replace $\langle x_i, x_j \rangle_2$ in your linear method by $k(x_i, x_j)$ where k is your favorite kernel

Application: Principal Component Analysis

(Kernel Trick: Feature map point of view)

Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$
\begin{aligned}
&\max_{\|w\|_2=1} \mathrm{Var}\left(\langle w, x_1\rangle_2, \langle w, x_2\rangle_2, \ldots, \langle w, x_n\rangle_2\right)\\
&\equiv \max_{\|w\|_1=1} \frac{1}{n}\sum_{i=1}^{n}\langle w, x_i\rangle_2^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\langle w, x_i\rangle_2\right)^2\\
&\equiv \max_{\|w\|_1=1} w^\top \hat{\boldsymbol{\Sigma}} w
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}} &:= \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^\top\\
&= \frac{1}{n}\mathbf{X}\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{X}^\top =: \mathbf{X}\mathbf{H}\mathbf{X}^\top.
\end{aligned}
$$

Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$\max_{\|w\|_2=1} \mathrm{Var}\left(\langle w, x_1\rangle_2, \langle w, x_2\rangle_2, \ldots, \langle w, x_n\rangle_2\right)$$

$$\equiv \max_{\|w\|_1=1} \frac{1}{n}\sum_{i=1}^n \langle w, x_i\rangle_2^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle w, x_i\rangle_2\right)^2$$

$$\equiv \max_{\|w\|_1=1} w^\top \hat{\boldsymbol{\Sigma}} w$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^\top$$

$$= \frac{1}{n}\mathbf{X}\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{X}^\top =: \mathbf{X}\mathbf{H}\mathbf{X}^\top.$$

Principal Component Analysis

- Dimensionality reduction
- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$
- Task: Find a low-dimensional representation for (x_i).

$$
\begin{aligned}
&\max_{\|w\|_2=1} \operatorname{Var}\left(\langle w, x_1\rangle_2, \langle w, x_2\rangle_2, \ldots, \langle w, x_n\rangle_2\right) \\
&\equiv \max_{\|w\|_1=1} \frac{1}{n}\sum_{i=1}^n \langle w, x_i\rangle_2^2 - \left(\frac{1}{n}\sum_{i=1}^n \langle w, x_i\rangle_2\right)^2 \\
&\equiv \max_{\|w\|_1=1} w^\top \hat{\mathbf{\Sigma}} w
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{\mathbf{\Sigma}} &:= \frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i\right)^\top \\
&= \frac{1}{n}\mathbf{X}\left(I_d - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{X}^\top =: \mathbf{X}\mathbf{H}\mathbf{X}^\top.
\end{aligned}
$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} w &= \lambda_1 w \\ \mathbf{X}\mathbf{H}\mathbf{X}^\top w &= \lambda_1 w \\ \underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} &= \lambda_1 \mathbf{X}^\top w \\ \mathbf{K}\mathbf{H}\alpha &= \lambda_1 \alpha \end{aligned}$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\boldsymbol{\Sigma}}$
- $$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$

  The eigenvector of $\hat{\boldsymbol{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

  $$w = \frac{1}{\lambda_1}\mathbf{X}\mathbf{H}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1}\boldsymbol{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n}(\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\mathbf{\Sigma}} w = \lambda_1 w$$

$$\mathbf{XHX}^\top w = \lambda_1 w$$

$$\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$

$$\mathbf{KH}\alpha = \lambda_1 \alpha$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\mathbf{\Sigma}}$
-

$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{XH}\alpha = \mathbf{XHX}^\top w = \hat{\mathbf{\Sigma}} w = \lambda_1 w$$

The eigenvector of $\hat{\mathbf{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

$$w = \frac{1}{\lambda_1} \mathbf{XH}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1} \mathbf{\Phi}(\mathbf{X}) \mathbf{H}\alpha = \sum_{i=1}^{n} (\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\mathbf{\Sigma}} w = \lambda_1 w$$

$$\mathbf{X}\mathbf{H}\mathbf{X}^\top w = \lambda_1 w$$

$$\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$

$$\mathbf{K}\mathbf{H}\alpha = \lambda_1 \alpha$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\mathbf{\Sigma}}$
-
$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\mathbf{\Sigma}} w = \lambda_1 w$$

  The eigenvector of $\hat{\mathbf{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

$$w = \frac{1}{\lambda_1}\mathbf{X}\mathbf{H}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1}\mathbf{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n}(\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$

$$\mathbf{X}\mathbf{H}\mathbf{X}^\top w = \lambda_1 w$$

$$\underbrace{\mathbf{X}^\top\mathbf{X}}_{\mathbf{K}}\mathbf{H}\ \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$

$$\mathbf{K}\mathbf{H}\alpha = \lambda_1 \alpha$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\boldsymbol{\Sigma}}$
-
$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$

  The eigenvector of $\hat{\boldsymbol{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

$$w = \frac{1}{\lambda_1}\mathbf{X}\mathbf{H}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1}\boldsymbol{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n}(\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\mathbf{\Sigma}} w = \lambda_1 w$$
$$\mathbf{X}\mathbf{H}\mathbf{X}^\top w = \lambda_1 w$$
$$\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$
$$\mathbf{K}\mathbf{H}\alpha = \lambda_1 \alpha$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\mathbf{\Sigma}}$
-
$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\mathbf{\Sigma}} w = \lambda_1 w$$

  The eigenvector of $\hat{\mathbf{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

$$w = \frac{1}{\lambda_1} \mathbf{X}\mathbf{H}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1} \mathbf{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n} (\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\mathbf{\Sigma}} w = \lambda_1 w$$
$$\mathbf{XHX}^\top w = \lambda_1 w$$
$$\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$
$$\mathbf{KH}\alpha = \lambda_1 \alpha$$

- The eigenvalues of **KH** are the same as that of $\hat{\mathbf{\Sigma}}$
-
$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{XH}\alpha = \mathbf{XHX}^\top w = \hat{\mathbf{\Sigma}} w = \lambda_1 w$$
The eigenvector of $\hat{\Sigma}$ can be computed from the eigenvector of **KH** as
$$w = \frac{1}{\lambda_1} \mathbf{XH}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1} \mathbf{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n} (\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\begin{aligned} \hat{\boldsymbol{\Sigma}} w &= \lambda_1 w \\ \mathbf{X}\mathbf{H}\mathbf{X}^\top w &= \lambda_1 w \\ \underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} &= \lambda_1 \mathbf{X}^\top w \\ \mathbf{K}\mathbf{H}\alpha &= \lambda_1 \alpha \end{aligned}$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\boldsymbol{\Sigma}}$
-

$$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$

The eigenvector of $\hat{\boldsymbol{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as

$$w = \frac{1}{\lambda_1}\mathbf{X}\mathbf{H}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1}\boldsymbol{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n}(\mathbf{H}\alpha)_i \Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\begin{aligned}\hat{\mathbf{\Sigma}}w &= \lambda_1 w\\ \mathbf{X}\mathbf{H}\mathbf{X}^\top w &= \lambda_1 w\\ \underbrace{\mathbf{X}^\top\mathbf{X}}_{\mathbf{K}}\mathbf{H}\ \underbrace{\mathbf{X}^\top w}_{\alpha} &= \lambda_1 \mathbf{X}^\top w\\ \mathbf{K}\mathbf{H}\alpha &= \lambda_1\alpha\end{aligned}$$

- The eigenvalues of $\mathbf{KH}$ are the same as that of $\hat{\mathbf{\Sigma}}$
- $$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{X}\mathbf{H}\alpha = \mathbf{X}\mathbf{H}\mathbf{X}^\top w = \hat{\mathbf{\Sigma}}w = \lambda_1 w$$
  The eigenvector of $\hat{\mathbf{\Sigma}}$ can be computed from the eigenvector of $\mathbf{KH}$ as
  $$w = \frac{1}{\lambda_1}\mathbf{X}\mathbf{H}\alpha$$
- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1}\mathbf{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n}(\mathbf{H}\alpha)_i\Phi(x_i)$$

Principal Component Analysis

- Solution: Find the eigenvector corresponding to the maximum eigenvalue of $\hat{\Sigma}$,

$$\hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$
$$\mathbf{XHX}^\top w = \lambda_1 w$$
$$\underbrace{\mathbf{X}^\top \mathbf{X}}_{\mathbf{K}} \mathbf{H} \underbrace{\mathbf{X}^\top w}_{\alpha} = \lambda_1 \mathbf{X}^\top w$$
$$\mathbf{KH}\alpha = \lambda_1 \alpha$$

- The eigenvalues of **KH** are the same as that of $\hat{\boldsymbol{\Sigma}}$
- $$\alpha = \mathbf{X}^\top w \Longrightarrow \mathbf{XH}\alpha = \mathbf{XHX}^\top w = \hat{\boldsymbol{\Sigma}} w = \lambda_1 w$$

  The eigenvector of $\hat{\boldsymbol{\Sigma}}$ can be computed from the eigenvector of **KH** as

$$w = \frac{1}{\lambda_1} \mathbf{XH}\alpha$$

- Kernel PCA (Schölkopf et al., 1998):

$$w = \frac{1}{\lambda_1} \boldsymbol{\Phi}(\mathbf{X})\mathbf{H}\alpha = \sum_{i=1}^{n} (\mathbf{H}\alpha)_i \Phi(x_i)$$

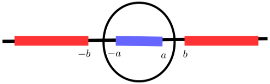
Feature Map and Kernel Trick

Same idea yields: (Schölkopf and Smola, 2002)

- Linear SVM $\rightarrow$ Kernel SVM
- Fisher discriminant analysis (FDA) $\rightarrow$ Kernel FDA
- Canonical correlation analysis (CCA) $\rightarrow$ Kernel CCA

many more ...

Revisiting Nonlinear Classification: 1

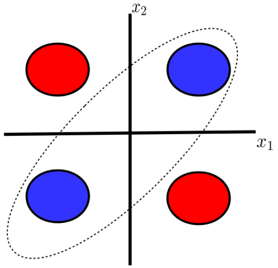


- The following function perfectly separates red and blue regions

$$f(x) = x^2 - r = \left\langle \underbrace{(1, -r)}_{w}, \underbrace{(x^2, 1)}_{\Phi(x)} \right\rangle_2 , \; a < r < b.$$

- Apply kernel trick with $k(x, y) = x^2 y^2 + 1$.

Revisiting Nonlinear Classification: 2



- A conic section, however, perfectly separates them

$$f(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2 + dx_1 + ex_2 + g$$
$$= \left\langle \underbrace{(a, b, c, d, e, g)}_{w}, \underbrace{(x_1^2, x_1x_2, x_2^2, x_1, x_2, 1)}_{\Phi(x)} \right\rangle_2 .$$

- Apply kernel trick with $k(x, y)$. Exercise: Find the kernel $k(x, y)$.

Application: Ridge Regression

(Representer Theorem: Function space point of view)

Learning Theory: Revisit

- Empirical risk: $\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$

$$f_D := \arg\min_{f:X\to\mathbb{R}} \mathcal{R}_{L,D}(f)$$

- To avoid overfitting: Perform ERM on a small set $\mathfrak{F}$ of functions (class of smooth functions)

$$f_D := \arg\inf_{f\in\mathfrak{F}} \mathcal{R}_{L,D}(f)$$

- Choice of $\mathcal{F}$: Evaluation functionals are bounded.

$$|\delta_x(f)| = |f(x)| \le M_x \|f\|_{\mathfrak{F}},\ \forall\, x \in \mathcal{X},\ f \in \mathfrak{F}$$

Pick $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le \alpha\}$; $\mathcal{H}$ is an RKHS

Classification with Lipschitz functions (von Luxburg and Bousquet, JMLR 2004)

Penalized Estimation

- We have

$$f_D = \arg \inf_{\|f\|_{\mathcal{H}} \leq \alpha} R_{L,D}(f)$$

$$= \arg \inf_{\|f\|_{\mathcal{H}} \leq \alpha} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

- In the Lagrangian formulation, we have

$$f_D = \arg \inf_{f \in \mathcal{H}} R_{L,D}(f) + \lambda \|f\|_{\mathcal{H}}^2$$

$$= \arg \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

where $\lambda > 0$.

Optimization over (possibly infinite dimensional) function space

Representer Theorem

Consider the penalized estimation problem,

$$\inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \theta(\|f\|_{\mathcal{H}})$$

where $\theta : [0, \infty) \to \mathbb{R}$ is a non-decreasing function.

- (Kimeldorf and Wahba, 1971; Schölkopf et al., ALT 2001) The solution to the above minimization problem is achieved by a function of the form

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

where $(\alpha_i)_{i=1}^{n} \subset \mathbb{R}$.

The infinite dimensional optimization problem reduces to a finite dimensional optimization problem in $\mathbb{R}^n$.

Proof

- Decomposition:
$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^{\perp},$$
where $\mathcal{H}_0 = \mathrm{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$, $\mathcal{H}_0^{\perp}$: orthogonal complement. Decompose
$$f = f_0 + f^{\perp}$$
accordingly.

- The loss function L does not change by replacing f with f_0 because
$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_0, k(\cdot, x_i)\rangle_{\mathcal{H}} + \underbrace{\langle f^{\perp}, k(\cdot, x_i)\rangle_{\mathcal{H}}}_{=0}.$$

- Penalty term:
$$\|f_0\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \qquad \Rightarrow \qquad \theta(\|f_0\|_{\mathcal{H}}) \leq \theta(\|f\|_{\mathcal{H}}).$$

- Thus the optimum lies in $\mathcal{H}_0$.

Kernel Ridge Regression

- $f : \mathcal{X} \to \mathbb{R}$ and $L(y, f(x)) = (y - f(x))^2$ (Squared loss)

$$\inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}}\right)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- By representer theorem, the solution is of the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ which on substitution yields

$$\inf_{\alpha} \frac{1}{n} \|\mathbf{Y} - \mathbf{K}\alpha\|^2 + \lambda \alpha^\top \mathbf{K} \alpha$$

where $\mathbf{K}$ is the Gram matrix with $\mathbf{K}_{ij} = k(x_i, x_j)$.

- Solution: $\hat{\alpha} = (\mathbf{K} + n\lambda I_n)^{-1}\mathbf{Y}$ (assuming $\mathbf{K}$ is invertible). For any $t \in \mathcal{X}$,

$$\hat{f}(t) = \sum_{i=1}^{n} \hat{\alpha}_i k(t, x_i) = \mathbf{Y}^\top (\mathbf{K} + n\lambda I_n)^{-1} \mathbf{k}_t,$$

where $(\mathbf{k}_t)_i := k(t, x_i)$. (Same solution as the feature map view point)

Kernel Ridge Regression

- $f : \mathcal{X} \to \mathbb{R}$ and $L(y, f(x)) = (y - f(x))^2$ (Squared loss)

$$\inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- By representer theorem, the solution is of the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ which on substitution yields

$$\inf_{\alpha} \frac{1}{n} \|\mathbf{Y} - \mathbf{K}\alpha\|^2 + \lambda \alpha^\top \mathbf{K} \alpha$$

where $\mathbf{K}$ is the Gram matrix with $\mathbf{K}_{ij} = k(x_i, x_j)$.

- Solution: $\hat{\alpha} = (\mathbf{K} + n\lambda I_n)^{-1} \mathbf{Y}$ (assuming $\mathbf{K}$ is invertible). For any $t \in \mathcal{X}$,

$$\hat{f}(t) = \sum_{i=1}^{n} \hat{\alpha}_i k(t, x_i) = \mathbf{Y}^\top (\mathbf{K} + n\lambda I_n)^{-1} \mathbf{k}_t,$$

where $(\mathbf{k}_t)_i := k(t, x_i)$. (Same solution as the feature map view point)

Kernel Ridge Regression

- $f : \mathcal{X} \to \mathbb{R}$ and $L(y, f(x)) = (y - f(x))^2$ (Squared loss)

$$\inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- By representer theorem, the solution is of the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$ which on substitution yields

$$\inf_{\alpha} \frac{1}{n} \|\mathbf{Y} - \mathbf{K}\alpha\|^2 + \lambda \alpha^\top \mathbf{K} \alpha$$

  where $\mathbf{K}$ is the Gram matrix with $\mathbf{K}_{ij} = k(x_i, x_j)$.

- Solution: $\hat{\alpha} = (\mathbf{K} + n\lambda I_n)^{-1} \mathbf{Y}$ (assuming $\mathbf{K}$ is invertible). For any $t \in \mathcal{X}$,

$$\hat{f}(t) = \sum_{i=1}^{n} \hat{\alpha}_i k(t, x_i) = \mathbf{Y}^\top (\mathbf{K} + n\lambda I_n)^{-1} \mathbf{k}_t,$$

  where $(\mathbf{k}_t)_i := k(t, x_i)$. (Same solution as the feature map view point)

Application: Principal Component Analysis

(Representer Theorem: Function space point of view)

Principal Component Analysis

- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$.
- $$\begin{aligned}
&\min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(f(x_1),\ldots,f(x_n)\right) \\
&= \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(\langle f, k(\cdot,x_1)\rangle_{\mathcal{H}},\ldots,\langle f, k(\cdot,x_n)\rangle_{\mathcal{H}}\right) \\
&= \min_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, k(\cdot,x_i)\rangle_{\mathcal{H}}^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right\rangle_{\mathcal{H}}^2 \\
&= \min_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f\rangle_{\mathcal{H}}
\end{aligned}$$

  where

  $$\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\otimes k(\cdot,x_i) - \left(\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right)$$

  is the covariance operator.
- The solution to the above problem is the eigenfunction of $\hat{\Sigma}$ corresponding to the maximum eigenvalue. (CAUTION!!)

Principal Component Analysis

- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$.
-

$$
\begin{aligned}
&\min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(f(x_1), \ldots, f(x_n)\right) \\
&= \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(\langle f, k(\cdot, x_1)\rangle_{\mathcal{H}}, \ldots, \langle f, k(\cdot, x_n)\rangle_{\mathcal{H}}\right) \\
&= \min_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}}^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right\rangle_{\mathcal{H}}^2 \\
&= \min_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}}
\end{aligned}
$$

where

$$
\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) \otimes k(\cdot, x_i) - \left(\frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right)
$$

is the covariance operator.

- The solution to the above problem is the eigenfunction of $\hat{\boldsymbol{\Sigma}}$ corresponding to the maximum eigenvalue. (CAUTION!!)

Principal Component Analysis

- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$.
- $$\begin{aligned} & \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}(f(x_1), \ldots, f(x_n)) \\ = & \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}(\langle f, k(\cdot, x_1)\rangle_{\mathcal{H}}, \ldots, \langle f, k(\cdot, x_n)\rangle_{\mathcal{H}}) \\ = & \min_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}}^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right\rangle_{\mathcal{H}}^2 \\ = & \min_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}} \end{aligned}$$

  where
  $$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) \otimes k(\cdot, x_i) - \left(\frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n k(\cdot, x_i)\right)$$
  is the covariance operator.
- The solution to the above problem is the eigenfunction of $\hat{\boldsymbol{\Sigma}}$ corresponding to the maximum eigenvalue. (CAUTION!!)

Principal Component Analysis

- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$.
-

$$\begin{aligned}
& \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(f(x_1), \ldots, f(x_n)\right) \\
= & \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(\langle f, k(\cdot, x_1)\rangle_{\mathcal{H}}, \ldots, \langle f, k(\cdot, x_n)\rangle_{\mathcal{H}}\right) \\
= & \min_{\|f\|_{\mathcal{H}}=1} \frac{1}{n} \sum_{i=1}^n \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}}^2 - \left\langle f, \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) \right\rangle_{\mathcal{H}}^2 \\
= & \min_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}}
\end{aligned}$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) \otimes k(\cdot, x_i) - \left(\frac{1}{n} \sum_{i=1}^n k(\cdot, x_i)\right) \otimes \left(\frac{1}{n} \sum_{i=1}^n k(\cdot, x_i)\right)$$

is the covariance operator.

- The solution to the above problem is the eigenfunction of $\hat{\boldsymbol{\Sigma}}$ corresponding to the maximum eigenvalue. (CAUTION!!)

Principal Component Analysis

- Given: $\{(x_i)\}_{i=1}^n$ where $x_i \in \mathcal{X}$.
-

$$\begin{aligned}
&\min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(f(x_1),\ldots,f(x_n)\right)\\
&= \min_{\|f\|_{\mathcal{H}}=1} \operatorname{Var}\left(\langle f, k(\cdot,x_1)\rangle_{\mathcal{H}},\ldots,\langle f, k(\cdot,x_n)\rangle_{\mathcal{H}}\right)\\
&= \min_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^n \langle f, k(\cdot,x_i)\rangle_{\mathcal{H}}^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right\rangle_{\mathcal{H}}^2\\
&= \min_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\boldsymbol{\Sigma}} f\rangle_{\mathcal{H}}
\end{aligned}$$

where

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\otimes k(\cdot,x_i) - \left(\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right)\otimes\left(\frac{1}{n}\sum_{i=1}^n k(\cdot,x_i)\right)$$

is the covariance operator.

- The solution to the above problem is the eigenfunction of $\hat{\boldsymbol{\Sigma}}$ corresponding to the maximum eigenvalue. (CAUTION!!)

Principal Component Analysis (Schölkopf et al., 1998)

- By the representer theorem, the solution to the optimization problem is of the form

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

- This yields

$$\min_{\alpha^\top \mathbf{K}\alpha = 1} \alpha^\top \mathbf{KHK}\alpha$$

and therefore α satisfies

$$\mathbf{KHK}\alpha = \lambda_1 \mathbf{K}\alpha.$$

- If $\mathbf{K}$ is invertible, then α is an eigenvector to $\mathbf{KH}$ (same solution as the feature map view point)

Principal Component Analysis (Schölkopf et al., 1998)

- By the representer theorem, the solution to the optimization problem is of the form

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

- This yields

$$\min_{\alpha^\top \mathbf{K} \alpha = 1} \alpha^\top \mathbf{KHK} \alpha$$

and therefore α satisfies

$$\mathbf{KHK}\alpha = \lambda_1 \mathbf{K} \alpha.$$

- If $\mathbf{K}$ is invertible, then α is an eigenvector to $\mathbf{KH}$ (same solution as the feature map view point)

Principal Component Analysis (Schölkopf et al., 1998)

- By the representer theorem, the solution to the optimization problem is of the form

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

- This yields

$$\min_{\alpha^\top \mathbf{K}\alpha = 1} \alpha^\top \mathbf{KHK}\alpha$$

and therefore α satisfies

$$\mathbf{KHK}\alpha = \lambda_1 \mathbf{K}\alpha.$$

- If $\mathbf{K}$ is invertible, then α is an eigenvector to $\mathbf{KH}$ (same solution as the feature map view point)

How to choose $\mathcal{H}$?

Large RKHS: Universal Kernel/RKHS

- Universal kernel (Steinwart, JMLR 2001): A kernel k on a compact metric space, $\mathcal{X}$ is said to be universal if the RKHS, $\mathcal{H}$ is dense (w.r.t. uniform norm) in the space of continuous functions on $\mathcal{X}$.

  Any continous function on $\mathcal{X}$ can be approximated arbitrarily by a function in $\mathcal{H}$.

- (Steinwart and Christmann, 2008) For certain conditions on L, if k is universal, then

$$\inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathbf{P}}(f) = \mathcal{R}_{L,\mathbf{P}}(f^*),$$

  i.e., approximation error is zero.

  - Squared loss, Hinge loss,...

Large RKHS: Universal Kernel/RKHS

- Universal kernel (Steinwart, JMLR 2001): A kernel k on a compact metric space, $\mathcal{X}$ is said to be universal if the RKHS, $\mathcal{H}$ is dense (w.r.t. uniform norm) in the space of continuous functions on $\mathcal{X}$.

  Any continous function on $\mathcal{X}$ can be approximated arbitrarily by a function in $\mathcal{H}$.

- (Steinwart and Christmann, 2008) For certain conditions on L, if k is universal, then

$$\inf_{f \in \mathcal{H}} \mathcal{R}_{L,\mathbf{P}}(f) = \mathcal{R}_{L,\mathbf{P}}(f^*),$$

  i.e., approximation error is zero.

  - Squared loss, Hinge loss,...

When is k Universal?

k is universal if and only if

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, d\mu(x)\, d\mu(y) > 0$$

for all non-zero finite signed measures, μ on $\mathcal{X}$.

(Carmeli et al., 2010; S et al., 2011)

Generalization of strictly positive definite kernels

- In Lecture 2, we will explore more by relating it to the Hilbert space embedding of measures.
- Examples: Gaussian, Laplacian, etc. (No finite dimensional RKHS is universal!!)

Advanced Topics

- Consistency and convergence rates of KRR (Caponnetto and De Vito, 2007; Steinwart et al., 2009)
- Consistency and convergence rates of KPCA (Blanchard et al., 2007; Rudi et al., 2013)
- Spectral regularization
- Stochastic gradient descent methods for KRR

Questions

References I

Aronszajn, N. (1950).
Theory of reproducing kernels.
Trans. Amer. Math. Soc., 68:337–404.

Blanchard, G., Bousquet, O., and Zwald, L. (2007).
Statistical properties of kernel principal component analysis.
Machine Learning, 66(2):259–294.

Caponnetto, A. and Vito, E. D. (2007).
Optimal rates for regularized least-squares algorithm.
Foundations of Computational Mathematics, 7:331–368.

Carmeli, C., Vito, E. D., Toigo, A., and Umanità, V. (2010).
Vector valued reproducing kernel Hilbert spaces and universality.
Analysis and Applications, 8:19–61.

Kimeldorf, G. S. and Wahba, G. (1971).
Some results on Tchebycheffian spline functions.
Journal of Mathematical Analysis and Applications, 33:82–95.

Rudi, A., Canas, G. D., and Rosasco, L. (2013).
On the sample complexity of subspace learning.
In *Advances in Neural Information Processing Systems 26*, pages 2067–2075.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).
A generalized representer theorem.
In *Proc. of the 14th Annual Conference on Learning Theory*, pages 416–426.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998).
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319.

Schölkopf, B. and Smola, A. J. (2002).
Learning with Kernels.
MIT Press, Cambridge, MA.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011).
Universality, characteristic kernels and RKHS embedding of measures.
Journal of Machine Learning Research, 12:2389–2410.

References II

Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 2:67–93.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.

Steinwart, I., Hush, D. R., and Scovel, C. (2009).
Optimal rates for regularized least squares regression.
In *Conference on Learning Theory*.

von Luxburg, U. and Bousquet, O. (2004).
Distance-based classification with Lipschitz functions.
Journal for Machine Learning Research, 5:669–695.

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press, Cambridge, UK.

Suggested Readings

Machine Learning

- Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels.* MIT Press, Cambridge, MA.
- Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, UK.
- Shawe-Taylor, J. and Cristianini, N. (2000). *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, UK.

Learning Theory

- Cucker, F. and Zhou, D-X. (2007). *Learning Theory: An Approximation Theory Viewpoint.* Cambridge University Press, Cambridge, UK.
- Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* Springer, NY.

Non-parametric Statistics

- Berlinet, A. and Thomas-Agnan, C. (2004.) *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, MA.
- Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Mathematics

- Paulsen, V. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.* Cambridge University Press, Cambridge, UK.