

Kernel Methods in Machine Learning - Course Notes

Hugo Cisneros

Contents

1	Kernels and RKHS	1
1.1	Positive Definite Kernels	1
1.2	Reproducing Kernel Hilbert Spaces (RKHS)	2
1.3	Examples	3
2	Kernel tricks	4
2.1	Kernel trick	4
2.2	Representer theorem	4
3	Kernel Methods: Supervised Learning	5
3.1	Kernel Ridge regression	5
3.2	Kernel logistic regression	5
3.3	Support vector machines (SVM)	6
4	Kernel Methods: Unsupervised Learning	6
4.1	Kernel K-means and spectral clustering	6
4.2	Kernel PCA	7
4.3	A quick note on kernel CCA	7
5	The Kernel Jungle	7
5.1	Green, Mercer, Herglotz, Bochner and friends	7
6	Open Problems and Research Topics	8
A	Proofs	8
A.1	Kernels and RKHS	8
A.2	Kernels Tricks	9
A.3	Kernel Methods: Unsupervised Learning	10
A.4	The Kernel Jungle	10

1 Kernels and RKHS

1.1 Positive Definite Kernels

Definition 1

A kernel K is a comparison function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.
With n data point $\{x_1, x_2, \dots, x_n\}$ a $n \times n$ matrix \mathbf{K} can be defined by $\mathbf{K}_{ij} = K(x_i, x_j)$.
A kernel K is **positive definite** (p.d.) if it is **symmetric** ($K(x, x') = K(x', x)$) and for

all sets of a and x

$$\sum_i \sum_j a_i a_j K(x_i, x_j) \geq 0$$

This is equivalent to the kernel matrix being **positive semi-definite**.

Examples:

- Kernel on $\mathbb{R} \times \mathbb{R}$ defined by $K(x, x') = xx'$ is p.d. ($xx' = x'x$ and $\sum_i \sum_j a_i a_j K(x_i, x_j) = (\sum_i a_i x_i)^2 \geq 0$).
- Linear kernel ($K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}$) is p.d
- More generally for any set \mathcal{X} , and function $\Phi : \mathcal{X} \rightarrow \mathbb{R}^d$, the kernel defined by $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^d}$ is p.d.

Theorem 1 – Aronszajn, 1950

K is a p.d. kernel on the set \mathcal{X} if and only if there exists a **Hilbert space** \mathcal{H} and a **mapping** $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ such that, for any x, x' in \mathcal{X} :

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

Proof.

(A Hilbert space is a vector space with an inner product and complete for the corresponding norm).

1.2 Reproducing Kernel Hilbert Spaces (RKHS)

Let \mathcal{X} be a set and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ a class of functions forming a Hilbert space.

Definition 2 – Reproducing kernel

A kernel K is called a **reproducing kernel** (r.k.) of \mathcal{H} if

- \mathcal{H} contains all functions of the form

$$\forall x \in \mathcal{X}, K_x : t \rightarrow K(x, t)$$

- For every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$

If there exists a r.k., \mathcal{H} is called a RKHS.

Theorem 2 – Equivalent Definition of RKHS

\mathcal{H} is a RKHS if and only if for any $x \in \mathcal{X}$, the mapping

$$\begin{aligned} F : \mathcal{H} &\rightarrow \mathbb{R} \\ f &\mapsto f(x) \end{aligned}$$

is **continuous**.

Proof.

As a corollary, convergence in a RKHS implies point-wise convergence.

Theorem 3 – Uniqueness of RKHS

If \mathcal{H} is a RKHS, it has a **unique r.k.**, and a function K can be **the r.k of at most one RKHS**.

Proof.

Theorem 4

A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is **p.d. if and only if it is a r.k.**

Proof.

1.3 Examples

1.3.1 Steps for finding the RKHS of a Kernel

1. Look for an **inner product** ($K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$)
2. Propose a **candidate RKHS** \mathcal{H}
3. Check that the candidate \mathcal{H} is a **Hilbert space** (inner product and complete)
4. Check that \mathcal{H} is **the RKHS**
 - \mathcal{H} contains all the functions $K_x : t \mapsto K(x, t)$
 - For all $f \in \mathcal{H}$ and $x \in \mathcal{X}$, $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$.

1.3.2 Linear Kernel

Definition 3 – Linear Kernel

In \mathbb{R}^d , the linear kernel is defined by $K(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$

Theorem 5 – RKHS of a linear Kernel

The RKHS of the linear kernel is the set of linear functions of the form $f_w(x) = \langle w, x \rangle_{\mathbb{R}^d}$ for $w \in \mathbb{R}^d$, endowed with the inner product $\langle f_w, f_v \rangle_{\mathcal{H}} = \langle w, v \rangle_{\mathbb{R}^d}$

1.3.3 Polynomial Kernel

Definition 4 – Polynomial Kernel

In \mathbb{R}^d , the polynomial kernel is defined by $K(x, y) = \langle x, y \rangle_{\mathbb{R}^d}^2$

Theorem 6 – RKHS of a polynomial Kernel

The RKHS \mathcal{H} of the polynomial kernel is the set of quadratic functions of the form $f_S(x) = x^T S x$ for $S \in \mathcal{S}^{d \times d}$

1.3.4 Properties of kernels

If K_1, K_2 are p.d. kernels,

- $K_1 + K_2$ is a p.d. kernel
- $K_1 \cdot K_2$ is a p.d. kernel
- cK_1 for $c \geq 0$ is a p.d. kernel
- The point-wise limits of a sequence of p.d. kernels is a p.d. kernel.
- $\exp(K_1)$ is a p.d. kernel

Small norms in the RKHS space means slow variations in the original space \mathcal{X} with respect to the geometry defined by the kernel.

2 Kernel tricks

2.1 Kernel trick

Statement: All expression of vectors that can be written in terms of pairwise inner products can be transposed to a infinite dimensional space by replacing inner products with kernel evaluations.

2.2 Representer theorem

Theorem 7 – Representer theorem

Let \mathcal{X} a set with a p.d. kernel K and corresponding RKHS \mathcal{H} , $S = \{x_1, \dots, x_n\} \subset \mathcal{X}$ a set of points of \mathcal{X} .

Let $\Phi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ a function strictly increasing w.r.t. the last variable.

Any solution to the optimization problem

$$\min_{f \in \mathcal{H}} \Phi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}})$$

admits a representation in the form

$$\forall x \in \mathcal{X}, f(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$$

Proof

One of the main consequences of the theorem is that problems of the form

$$\min_{f \in \mathcal{H}} \Phi(f(x_1), \dots, f(x_n), \|f\|_{\mathcal{H}})$$

can be re-written as

$$\min_{\alpha \in \mathbb{R}^n} \Phi([\mathbf{K}\alpha]_1, \dots, [\mathbf{K}\alpha]_n, \alpha^T \mathbf{K}\alpha)$$

which is a n-dimensional optimization problem (instead of a possibly infinite dimensional one).

3 Kernel Methods: Supervised Learning

3.1 Kernel Ridge regression

The problem can be described as minimizing a RKHS norm regularized MSE criterion

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

Effects of regularization:

- **Penalize non smooth functions** (avoid overfitting)
- **Simplify the solution** (representer theorem)

The problem can be re-written

$$\hat{\alpha} = \arg \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} (\mathbf{K}\alpha - y)^T (\mathbf{K}\alpha - y) + \lambda \alpha^T \mathbf{K}\alpha$$

One solution is to take

$$\alpha = (\mathbf{K} + \lambda n \mathbf{I})^{-1} y$$

(Uniqueness: If \mathbf{K} is singular, all $\alpha + \varepsilon$ with $\varepsilon \in \text{Ker}(\mathbf{K})$ are solutions leading to the same function f .)

3.2 Kernel logistic regression

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i f(x_i))) + \lambda \|f\|_{\mathcal{H}}^2$$

This problem can also be reformulated in terms of the Gram matrix of the kernel and a parameter α

$$\min_{\alpha \in \mathbb{R}^n} J(\alpha) \triangleq \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i [\mathbf{K}\alpha]_i)) + \frac{\lambda}{2} \alpha^T \mathbf{K}\alpha$$

By writing and computing the terms of the Taylor expansion of J near a point α_0 , we can explicitly solve the problem with Newton's method.

$$J_q(\alpha) = J(\alpha_0) + (\alpha - \alpha_0)^T \nabla J(\alpha_0) + \frac{1}{2} (\alpha - \alpha_0)^T \nabla^2 J(\alpha_0) (\alpha - \alpha_0)$$

$$\nabla J(\alpha) = \frac{1}{n} \mathbf{K} \mathbf{P}(\alpha) y + \lambda \mathbf{K} \alpha$$

$$\nabla^2 J(\alpha) = \frac{1}{n} \mathbf{K} \mathbf{W}(\alpha) \mathbf{K} + \lambda \mathbf{K}$$

where $\mathbf{P}(\alpha) = \text{diag}(\ell'_{\text{logistic}}(y_i [\mathbf{K}\alpha]_i))$ and $\mathbf{W}(\alpha) = \text{diag}(\ell''_{\text{logistic}}(y_i [\mathbf{K}\alpha]_i))$. By developing the approximation, we obtain the following equality

$$2J_q(\alpha) = \frac{1}{n} (\mathbf{K}\alpha - z)^T \mathbf{w} (\mathbf{K}\alpha - z) + \lambda \alpha^T \mathbf{K}\alpha + C$$

with $z = (\mathbf{K}\alpha_0 - \mathbf{W}^{-1} \mathbf{P} y)$. This is exactly the formulation of a weighted kernel ridge regression problem. This problem can be iteratively solved by updating W^t and z^t until convergence (kernel IRLS).

3.3 Support vector machines (SVM)

Definition 5 – Hinge loss

The Hinge loss is a function $\mathbb{R} \rightarrow \mathbb{R}_+$ defined by

$$\varphi_{\text{hinge}}(u) = \max(0, 1 - u) = \begin{cases} 0 & \text{if } u \geq 1 \\ 1 - u & \text{otherwise} \end{cases}$$

Definition 6 – SVM problem

SVM is the large margin classifier that solves

$$\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi_{\text{hinge}}(y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

It can be reformulated by using the Representer theorem as

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi_{\text{hinge}}(y_i [\mathbf{K}\alpha]_i) + \lambda \alpha^T \mathbf{K} \alpha \right\}$$

Then, by introducing slack variables and using the definition of the Hinge loss, the following formulation is obtained

$$\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x)$$

where $\hat{\alpha}$ solves

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \quad & \frac{1}{n} \sum_{i=1}^n \xi_i + \lambda \alpha^T \mathbf{K} \alpha \\ \text{s.t.} \quad & y_i [\mathbf{K}\alpha]_i + \xi_i - 1 \geq 0, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \end{aligned}$$

4 Kernel Methods: Unsupervised Learning

4.1 Kernel K-means and spectral clustering

The objective is similar to K-means, but transposed in the RKHS. Given data points x_1, \dots, x_n and a p.d. kernel K and RKHS \mathcal{H} the objective reads

$$\min_{\substack{\mu_j \in \mathcal{H} \\ s_i \in \{1, \dots, k\}}} \sum_{i=1}^n \|\varphi(x_i) - \mu_{s_i}\|_{\mathcal{H}}^2$$

Proposition 1

The center of mass $\varphi_n = \frac{1}{n} \sum_{i=1}^n \varphi(x_i)$ solves the optimization problem

$$\min_{\mu \in \mathcal{H}} \sum_{i=1}^n \|\varphi(x_i) - \mu\|_{\mathcal{H}}^2$$

Proof.

Greedy (K-means) approach:

Centroid update Given a centroid assignment, update the centroids

$$\forall j, \quad \mu_j = \frac{1}{|C_j|} \sum_{i \in C_j} \varphi(x_i)$$

Cluster assignment

4.2 Kernel PCA

4.3 A quick note on kernel CCA

5 The Kernel Jungle

5.1 Green, Mercer, Herglotz, Bochner and friends

5.1.1 Green Kernel

Theorem 8 – Green Kernel in dimension 1

The set defined by

$$\mathcal{H} = \{f : [0, 1] \rightarrow \mathbb{R}, \text{absolutely continuous}, f' \in L^2([0, 1]), f(0) = 0\}$$

endowed with the inner product $\forall (f, g) \in \mathcal{F}^2 \langle f, g \rangle = \int_0^1 f'(u)g'(u)du$, is a RKHS with with r.k.

$$\forall (x, y) \in [0, 1]^2, K(x, y) = \min(x, y)$$

Theorem 9 – General Green Kernel

If D is a differential operator on a class of functions of \mathcal{H} such that the inner product $\langle f, g \rangle_{\mathcal{H}} = \langle Df, Dg \rangle_{L^2(\mathcal{X})}$

Then \mathcal{H} is a RKHS and admits for r.k. the Green function of the operator D^*D

5.1.2 Mercer Kernels

Definition 7 – Mercer Kernels

A kernel K on a set \mathcal{X} is called a Mercer kernel if:

- \mathcal{X} is a compact metric space (typically, a closed bounded subset of \mathbb{R}^d)
- $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a continuous p.d kernel (w.r.t the Borel topology)

6 Open Problems and Research Topics

A Proofs

A.1 Kernels and RKHS

Proof of Theorem 2

(\Rightarrow) If a r.k. exists in \mathcal{H} then for any $(x, f) \in \mathcal{X} \times \mathcal{H}$:

$$\begin{aligned} |f(x)| &= |\langle f, K_x \rangle_{\mathcal{H}}| \\ &\leq \|f\|_{\mathcal{H}} \cdot \|K_x\|_{\mathcal{H}} \quad (\text{Cauchy-Schwarz}) \\ &\leq \|f\|_{\mathcal{H}} \cdot K(x, x)^{\frac{1}{2}} \end{aligned}$$

Therefore, $f \in \mathcal{H} \rightarrow f(x) \in \mathbb{R}$ is a linear continuous mapping because F is linear and $\lim_{f \rightarrow 0} F(f) = 0$

(\Leftarrow) F is continuous, by the Riesz representation theorem: there exists a unique $g_x \in \mathcal{H}$ such that $f(x) = \langle f, g_x \rangle_{\mathcal{H}}$.

The function $K : (x, y) \mapsto g_x(y)$ is then a r.k. for \mathcal{H}

Proof of Theorem 3

(Uniqueness) If K and K' are two r.k. of a RKHS, then for any x

$$\|K_x - K'_x\|^2 = K_x(x) - K'_x(x) - K_x(x) + K'_x(x) = 0$$

So $K_x = K'_x$

Proof of Theorem 4

(\Leftarrow) A r.k. is symmetric, and $\sum_{i,j} a_i a_j K(x_i, x_j) = \|\sum_i a_i K_{x_i}\|_{\mathcal{H}}^2 \geq 0$

(\Rightarrow) Let \mathcal{H}_0 be the subspace spanned by the functions $(K_x)_{x \in \mathcal{X}}$. If $f = \sum_i a_i K_{x_i}$ and $g = \sum_j b_j K_{y_j}$. Let (not an inner product yet)

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}_0} &= \sum_{i,j} a_i b_j K(x_i, y_j) \\ &= \sum_i a_i g(x_i) \\ &= \sum_j b_j f(y_j) \end{aligned}$$

$\langle f, g \rangle_{\mathcal{H}_0}$ does not depend on the expansion of f or g) For any $x \in \mathcal{X}$ and $f \in \mathcal{H}_0$, $\langle f, K_x \rangle_{\mathcal{H}_0} = f(x)$.

$$\|f\|_{\mathcal{H}_0}^2 = \sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

And since Cauchy-Schwarz is valid,

$$|f(x)| = |\langle f, K_x \rangle_{\mathcal{H}_0}| \leq \|f\|_{\mathcal{H}_0} \cdot K(x, x)^{\frac{1}{2}}$$

Therefore $\|f\|_{\mathcal{H}_0} = 0 \implies f = 0$. $\langle \cdot, \cdot \rangle$ is an inner product on \mathcal{H}_0 .

For a Cauchy sequence $(f_n)_{n \geq 0}$,

$$|f_m(x) - f_n(x)| \leq \|f_m - f_n\|_{\mathcal{H}_0} \cdot K(x, x)^{\frac{1}{2}}$$

For any x the sequence $(f_n(x))$ is Cauchy in \mathbb{R} and therefore converges.

If the functions defined as the point-wise limits of Cauchy sequences are added \mathcal{H}_0 , it becomes a Hilbert space with K as r.k..

Proof of Aronszajn's theorem

If K is p.d. over a set \mathcal{X} , it is the r.k. of a Hilbert space \mathcal{H} . The mapping Φ is defined by $\forall x \in \mathcal{X}, \quad \Phi(x) = K_x$.

By the reproducing property

$$\forall (x, y) \in \mathcal{X}^2, \quad \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \langle K_x, K_y \rangle_{\mathcal{H}} = K(x, y)$$

A.2 Kernels Tricks

Proof of the Representer theorem

Let $\xi(f)$ the functional that is minimized in the optimization problem of the theorem, and $\mathcal{H}_{\mathcal{S}}$ the linear span of all the K_{x_i} functions.

Since $\mathcal{H}_{\mathcal{S}}$ is a finite dimensional space, every function $f \in \mathcal{H}$ can be decomposed as $f = f_{\mathcal{S}} + f_{\perp}$, with $f_{\mathcal{S}}$ the orthogonal projection of f on $\mathcal{H}_{\mathcal{S}}$.

Because \mathcal{H} is a RKHS,

$$\forall i \leq n, \quad f_{\perp}(x_i) = \langle f_{\perp}, K_{x_i} \rangle_{\mathcal{H}} = 0$$

Therefore

$$\forall i \leq n, \quad f(x_i) = f_{\mathcal{S}}(x_i)$$

From Pythagora's theorem in \mathcal{H} , $\|f\|_{\mathcal{H}}^2 = \|f_{\mathcal{S}}\|_{\mathcal{H}}^2 + \|f_{\perp}\|_{\mathcal{H}}^2$.

We therefore have $\xi(f) \geq \xi(f_{\mathcal{S}})$ with equality if and only if $\|f_{\perp}\|_{\mathcal{H}}^2 = 0$, the minimum belongs to $\mathcal{H}_{\mathcal{S}}$.

A.3 Kernel Methods: Unsupervised Learning

Proof of Proposition 1

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\varphi(x_i) - \mu\|_{\mathcal{H}}^2 &= \frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\|_{\mathcal{H}}^2 - \left\langle \frac{2}{n} \sum_{i=1}^n \varphi(x_i), \mu \right\rangle_{\mathcal{H}} + \|\mu\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\|_{\mathcal{H}}^2 - 2\langle \varphi_n, \mu \rangle_{\mathcal{H}} + \|\mu\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\|_{\mathcal{H}}^2 - \|\varphi_n\|_{\mathcal{H}}^2 + \|\varphi_n - \mu\|_{\mathcal{H}}^2\end{aligned}$$

which is minimum for $\varphi_n = \mu$.

A.4 The Kernel Jungle

Proof