Tutorial on Kernels (IML Tutorial V)

Max B. Paulus

March 23, 2019

Table of contents

1. Maths & Intuition
   - What is a kernel?
   - How can we construct a kernel?
   - Positive Definiteness

2. Applications
   - Why care?
   - Highlight: Kernel Ridge Regression
   - Digression: Feature Selection

What is a kernel?

Definition (Inner Product)

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

What is a kernel?

Definition (Inner Product)

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

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$

What is a kernel?

Definition (Inner Product)

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

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$

What is a kernel?

Definition (Inner Product)

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

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0 \quad \forall f \in \mathcal{H}$ and $\langle f, f \rangle_{\mathcal{H}} = 0 \iff f = 0$

Definition (Kernel)

What is a kernel?

Definition (Inner Product)

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

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle_{\mathcal{H}} \geq 0 \quad \forall f \in \mathcal{H}$ and $\langle f, f \rangle_{\mathcal{H}} = 0 \iff f = 0$

Definition (Kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$, such that $\forall x, x' \in \mathcal{X}$,

What is a kernel?

Definition (Inner Product)

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

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g\rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g\rangle_{\mathcal{H}}$
2. $\langle f, g\rangle_{\mathcal{H}} = \langle g, f\rangle_{\mathcal{H}}$
3. $\langle f, f\rangle_{\mathcal{H}} \geq 0 \quad \forall f \in \mathcal{H}$ and $\langle f, f\rangle_{\mathcal{H}} = 0 \iff f = 0$

Definition (Kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$, such that $\forall x, x' \in \mathcal{X}$,

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

Some Examples of Kernels

- Linear Kernel: $k(x, x') = x^\top x'$ ($\mathcal{X} = \mathbb{R}^n$, $\mathcal{H} = \mathbb{R}^n$, $\phi(x) = x$)
- Polynomial kernel $k_d(x, x') = (x^\top x' + 1)^d$ ($\mathcal{X} = \mathbb{R}^n$, $\mathcal{H} = \mathbb{R}^{\binom{n+d}{d}}$)
- Gaussian kernel $k_h(x, x') = \exp(-\frac{\|x-x'\|^2}{2h^2})$ ($\mathcal{X} = \mathbb{R}^n$, $\mathcal{H} = \mathbb{R}^\infty$)
- String kernels, let $x \in \mathcal{A}^{,}$ and $x' \in \mathcal{A}^n$, now define $\phi_s(x) := \#\{s \text{ appears in } x\}$, $k(x, x') = \sum_{s \in \mathcal{A}^*} w_s \phi_s(x) \phi_s(x')$

Exercise

Find the feature map ϕ associated with the Gaussian kernel on $\mathbb{R}$ with $h = 1$, i.e. $k(x, y) = e^{-(x-y)^2}$ for $x, y \in \mathbb{R}$.

Exercise

Find the feature map ϕ associated with the Gaussian kernel on $\mathbb{R}$ with $h = 1$, i.e. $k(x, y) = e^{-(x-y)^2}$ for $x, y \in \mathbb{R}$.

$$\begin{aligned}
k(x, y) &= e^{-(x-y)^2} \\
&= e^{-x^2+2xy-y^2} \\
&= e^{-x^2}e^{-y^2}[e^{2xy}] \\
&= e^{-x^2}e^{-y^2}[1 + 2xy + \frac{(2xy)^2}{2!} + \frac{(2xy)^3}{3!} + \dots] \\
&= e^{-x^2}e^{-y^2}[1 + \sqrt{2}x\sqrt{2}y + \sqrt{\frac{2^2}{2!}}x^2\sqrt{\frac{2^2}{2!}}y^2 + \dots] \\
&= \phi(x)^\top\phi(y)
\end{aligned}$$

with $\phi(x) = e^{-x^2}[1, \sqrt{2}x, \sqrt{\frac{2^2}{2!}}x^2, \dots]^\top$

How can we construct a kernel?

Lemma (Positive Scaling Rule)

Given $\alpha > 0$ and k, a kernel on $\mathcal{X}$, then αk is a kernel on $\mathcal{X}$.

Lemma (Sum Rule)

Given k_1 and k_2, kernels on $\mathcal{X}$, then $k_1 + k_2$ is a kernel on $\mathcal{X}$.

Lemma (Product Rule)

Given k_1 and k_2, kernels on $\mathcal{X}$, then $k_1 k_2$ is a kernel on $\mathcal{X}$. If k_1 on $\mathcal{X}_1$, and k_2 on $\mathcal{X}_2$, then $k_1 k_2$ on $\mathcal{X}_1 \times \mathcal{X}_2$.

Lemma (Mapping Rule)

Given sets $\mathcal{X}$ and $\tilde{\mathcal{X}}$ and a map $A : \mathcal{X} \to \tilde{\mathcal{X}}$. Let k be a kernel on $\tilde{\mathcal{X}}$, then $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$.

Exercise

Let $\mathcal{H}_1$ corresponding to k_1 be R^m and $\mathcal{H}_2$ corresponding to k_2 be R^n. Let $k_1(x_1, y_1) = x_1^\top y_1$ and $k_1(x_2, y_2) = x_2^\top y_2$. Show that $k_1 k_2$ is a kernel on $\mathbb{R}^m \times \mathbb{R}^n$ using the inner product between two matrices A, B of same dimensions is $\langle A, B \rangle = \text{trace}(A^\top B)$.

Exercise

Let $\mathcal{H}_1$ corresponding to k_1 be R^m and $\mathcal{H}_2$ corresponding to k_2 be R^n. Let $k_1(x_1, y_1) = x_1^\top y_1$ and $k_1(x_2, y_2) = x_2^\top y_2$. Show that $k_1 k_2$ is a kernel on $\mathbb{R}^m \times \mathbb{R}^n$ using the inner product between two matrices A, B of same dimensions is $\langle A, B \rangle = \text{trace}(A^\top B)$.

$$
\begin{aligned}
k_1(x_1, y_1)k_2(x_2, y_2) &= (x_1^\top y_1)(x_2^\top y_2) \\
&= (x_1^\top y_1)(y_2^\top x_2) \\
&= (x_1^\top y_1)\text{trace}(y_2^\top x_2) \\
&= (x_1^\top y_1)\text{trace}(x_2 y_2^\top) \\
&= \text{trace}(x_2(x_1^\top y_1)y_2^\top) \\
&= \text{trace}((x_2 x_1^\top)(y_1 y_2^\top)) \\
&= \text{trace}((x_2 x_1^\top)(y_1 y_2^\top)) \\
&= \langle x_1 x_2^\top, y_1 y_2^\top \rangle
\end{aligned}
$$

Postive Definiteness & Kernels

Definition (Positive Definiteness)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positve definite if $\forall n \geq 1, \forall (a_1, \dots, a_n) \in \mathbb{R}^n, \forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Postive Definiteness & Kernels

Definition (Positive Definiteness)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positve definite if $\forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Lemma (Every kernel is positive definite)

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, x') := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function.

Postive Definiteness & Kernels

Definition (Positive Definiteness)

A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positve definite if $\forall n \geq 1, \forall (a_1, \dots, a_n) \in \mathbb{R}^n, \forall (x_1, \dots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Lemma (Every kernel is positive definite)

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, x') := \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ is a positive definite function.

Lemma (Every symmetric positive definite function is a kernel.)

Let $\mathcal{X}$ be a non-empty set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric, positive definite function. Then k is a kernel. [See also Mercer's Theorem for a characterisation of k.]

Examples & Exercises

Proof the Sum Rule.

Examples & Exercises

Proof the Sum Rule.

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j[k_1(x_i,x_j)+k_1(x_i,x_j)] \\
&=\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_1(x_i,x_j)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_2(x_i,x_j) \\
&\geq 0
\end{aligned}
$$

Some reasons why we care about kernels?

- **Kernel Machines:** Define feature vectors in terms of kernels, e.g. .
- **Kernelize linear algorithms**, i.e. a computationally efficient way to handle data that may be linearly separable in a higher-dimensional space
- Deal with **structured data**, e.g. natural language, amino acid sequencing, etc.

Highlight: Kernel Ridge Regression

Primal Formulation

- feature vector $x \in \mathbb{R}^D$, design matrix X is $N \times D$
- $L(w) = (y - Xw)^\top (y - Xw) + \lambda \|w\|^2$
- $w^* = (X^\top X + \lambda I)^{-1} X^\top y$

Dual Formulation

- $w^* = X^\top (XX^\top + \lambda I)^{-1} y$
- Define $\alpha := (XX^\top + \lambda I)^{-1} y$
- $w^* = X^\top \alpha = \sum_{i=1}^N \alpha_i x_i$
- $\hat{f}(x_{\text{test}}) = w^{*\top} x = \sum_{i=1}^N \alpha_i x_i^\top x_{\text{test}}$

In the dual formulation, $\hat{f}(x_{\text{test}})$ only depends on **inner products**: To kernelise ridge regression, for $XX^\top$ substitute K, a matrix of inner products between data points, and for $x_i^\top x_{\text{test}}$ substitute $k(x_i, x_{\text{test}})$.

Digression: Feature Selection

See blackboard or Bishop (144-146)