Kernel Method

- **Introduction**
- Why (positive) kernel?
- Example of KRR with artificial data
- Examples of positive kernels on $\mathbb{R} \times \mathbb{R}$
- Examples of positive kernels on high dimensional spaces

Positive kernel

- A function $K : X \times X \to \mathbb{R}$ is called a ***positive kernel*** (or *positive semi-definite kernel*) if
  - $K(u, v) = K(v, u)$ for any $u, v \in X$ and
  - for any n, for any sequence $x_1, \ldots, x_n \in X$ and any constant sequence $c_1, \ldots, c_n \in \mathbb{R}$, it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

- Write

$$\mathbb{K} = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_n) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_n) \\ \cdots & \cdots & \cdots & \cdots \\ K(x_n, x_1) & K(x_n, x_2) & \cdots & K(x_n, x_n) \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

The matrix $\mathbb{K}$ is called the *Gram matrix*. Note that $\mathbb{K}$ is symmetric. The above inequality then can be rewritten as

$$c' \mathbb{K} c \geq 0,$$

i.e. $\mathbb{K}$ is positive semidefinite if K is a positive kernel.

Examples of positive kernel

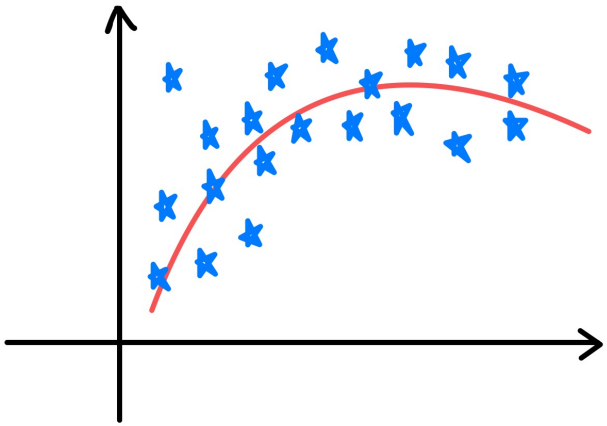
• If for any distinct set $x_1, x_2, \ldots, x_n$ (meaning any two of them are not equal), the matrix $\mathbb{K} = (K(x_i, x_j))_{i,j}$ is strictly positive definite, then we say that K is a *strictly positive* kernel.

• Examples of positive kernels: let $x, y \in \mathbb{R}^p$ for some $p \geq 1$,

- Inner product kernel: $K(x, y) = \langle x, y \rangle$
- Gaussian kernel: $K(x, y) = \exp\left\{-\|x - y\|^2\right\}$
- Polynomial kernel: $K(x, y) = (1 + \langle x, y \rangle)^d$, for some integer $d \geq 1$

Here the Gaussian kernel and the polynomial kernel are strictly positive.

Kernel for nonlinear curve



Nonlinear regression: using data to recover the unknown function $f_{\text{nonlinear}}$.

- Given a kernel K, and given the data $\{(x_i, y_i)\}_{i=1}^n$, we hope to find some coefficients c_i's, such that the estimated output for x_i is close to the observed output y_j.
- Predict the output of the new sample x

$$\hat{f}(x) = \sum_{i=1}^{n} c_i K(x_i, x) \approx f_{\text{nonlinear}}(x),$$

- We do not have to know the exact form of $f_{\text{nonlinear}}$.

Kernel ridge regression

- The kernel ridge regression algorithm is defined by
  1. Training: given dateset $\{(x_i, y^i)\}_{i=1}^n$, find $c^* := (c^{1,*}, \cdots, c^{n,*})'$ s.t.

$$c^* = \arg\min_{c\in\mathbb{R}^n} \left\{ \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^n c_j K(x_j, x_i) - y_i \right)^2 + \lambda c'\mathbb{K}c \right\}.$$

  2. Prediction: $\hat{f}_{\text{krr}}(x) = \sum_{j=1}^n c_j^* K(x_j, x)$; $\quad c^* = (c_1^*, \cdots, c_n^*)'$.
- The training step is equivalent to

$$c^* = \arg\min_{c\in\mathbb{R}^n} \left\{ \frac{1}{n}\|\mathbb{K}c - y\|^2 + \lambda c'\mathbb{K}c \right\} = (\mathbb{K} + \lambda n I)^{-1} y.$$

Remark.

- $\mathbb{K} + n\lambda I$ is positive definite and invertible (Exercise).
- $\frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^n c_j K(x_j, x_i) - y_i \right)^2$ is the "fidelity" term, which is loyal to the data used to fit the "curve".
- $\lambda c'\mathbb{K}c$ is the regularization term, to prevent c from being too large.

Kernel Method

- Introduction
- **Why (positive) kernel?**
- Example of KRR with artificial data
- Examples of positive kernels on $\mathbb{R} \times \mathbb{R}$
- Examples of positive kernels on high dimensional spaces

Why kernel (1/1)

- Nonlinearity.
- Kernel methods implicitly map data to a higher-dimensional space without explicitly calculating the coordinates in that space. This allows for working in high-dimensional spaces without actually storing or computing high-dimensional vectors.

Example. Consider

$$f(x) := (x_1^2, x_1x_2, x_1x_3, x_1x_4, x_2x_1, x_2^2, x_2x_3, x_2x_4, x_3x_1, x_3x_2, x_3^2, x_3x_4, x_4x_1, x_4x_2, x_4x_3, x_4^2)$$

Let $x = (1, 2, 3, 4)$ and $y = (5, 6, 7, 8)$. Calculate $\langle f(x), f(y)\rangle$.

- Calculate $f(x) = \cdots$
- Calculate $f(y) = \cdots$
- Calculate $\langle f(x), f(y)\rangle = \cdots$

Introduce a kernel function $K(x, y) = \langle x, y\rangle^2 = (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4)^2$. Then

$$\langle f(x), f(y)\rangle = K(x, y) = \cdots$$

Why positive kernel (1/5)

Lemma (Eigen decomposition). Let $A \in \mathbb{R}^{n\times n}$ be a real-valued matrix. Then

$$A = Q\Lambda Q',$$

where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, with λ_i being all eigenvalues of A.

Why positive kernel (2/5)

- Recall the KRR algorithm that requires one to find c^*:

$$c^* = \arg\min_{c\in\mathbb{R}^n} \left\{ \|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c \right\}.$$

- **Goal**: choose a appropriate non-positive kernel, such that the corresponding KRR has no solution[5].

Rewrite the regression cost into a polynomial of c:

$$\begin{aligned}\|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c &= c'\mathbb{K}^2 c + n\lambda c'\mathbb{K}c - 2y'\mathbb{K}c + \|y\|^2 \\ &= c'(\mathbb{K}^2 + n\lambda\mathbb{K})c - 2y'\mathbb{K}c + \|y\|^2.\end{aligned}$$

Using eigen decomposition, we have $\mathbb{K} = UDU'$, where U is an orthogonal matrix, and $D = \text{diag}\{d_1, d_2, \ldots, d_n\}$. If K is NOT a positive kernel, then $\mathbb{K}$ is not positive semidefinite. Thus, w.l.o.g., assume that $d_1 < 0$.

[5] No solution means the minimal value of the cost can never be reached.

Why positive kernel (3/5)

- Calculation of $c'(\mathbb{K}^2 + n\lambda\mathbb{K})c$. Recall $\mathbb{K} = U\text{diag}\{d_1, d_2, \cdots, d_n\}U'$.

$$\mathbb{K}^2 + n\lambda\mathbb{K} = U\text{diag}\left\{d_1^2 + n\lambda d_1, d_2^2 + n\lambda d_2, \cdots, d_n^2 + n\lambda d_n\right\}U'.$$

Choose

$$c' = (t, 0, \cdots, 0)U' \Leftrightarrow c = U\begin{bmatrix} t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

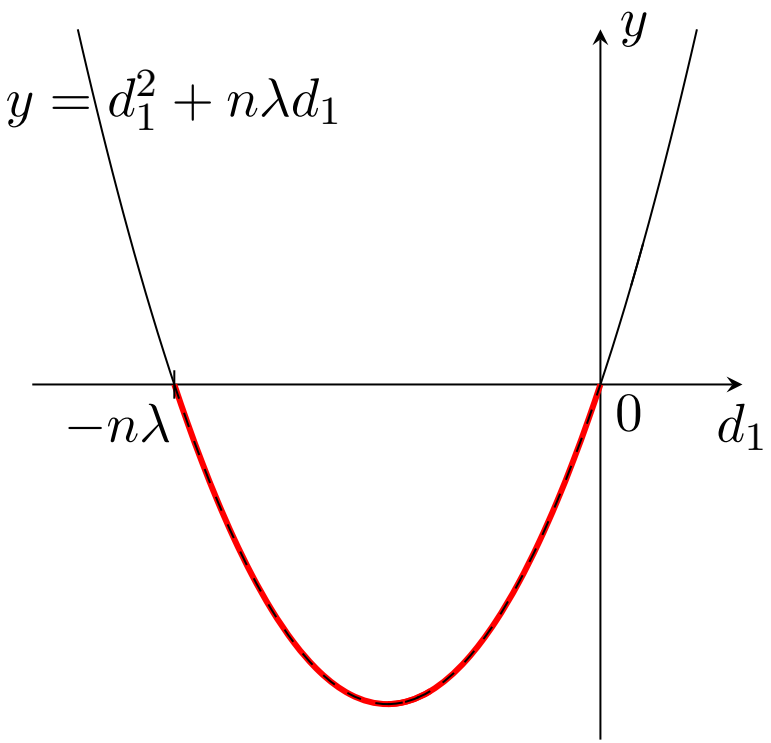
to get

$$c'(\mathbb{K}^2 + \lambda n\mathbb{K})c = (t, 0, \cdots, 0)\begin{bmatrix} d_1^2 + n\lambda d_1 & & 0 \\ & \ddots & \\ 0 & & d_n^2 + n\lambda d_n \end{bmatrix}\begin{pmatrix} t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= t^2(d_1^2 + n\lambda d_1).$$

- Note that for $d_1 < 0$, it is possible that $d_1^2 + n\lambda d_1 < 0$. Refer to the picture:

Why positive kernel (4/5)



Why positive kernel (5/5)

- Calculation of $-2y'\mathbb{K}c + \|y\|^2$. When $c = U[t, 0, \cdots, 0]'$,

$$-2y'\mathbb{K}c + \|y\|^2 = at + b,$$

for some a and b.

- Caculation of $\|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c$:

$$\|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c = (d_1^2 + n\lambda d_1)t^2 + at + b,$$

which means that

$$\begin{aligned} &\min_c \left(\|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c\right) \\ \leq\ &\|\mathbb{K}c - y\|^2 + n\lambda c'\mathbb{K}c = (d_1^2 + n\lambda d_1)t^2 + at + b \to -\infty, \end{aligned}$$

Thus, the kernel ridge regression has no solution!

Kernel Method

- Introduction
- Why (positive) kernel?
- **Example of KRR with artificial data**
- Examples of positive kernels on $\mathbb{R} \times \mathbb{R}$
- Extension: kernels on high dimensional spaces

Kernel Method

- Introduction
- Why (positive) kernel?
- Example of KRR with artificial data
- **Examples of positive kernels on $\mathbb{R} \times \mathbb{R}$**
- Examples of positive kernels on high dimensional spaces

Proof by expansion

$$
\begin{aligned}
e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!} \\
(1+x)^d &= \sum_{m=0}^{d} \binom{d}{m} x^m \\
\frac{1}{1-x} &= \sum_{n=0}^{\infty} x^n, \quad x \in (-1,1) \\
-\log(1-t) &= t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \frac{t^5}{5} + \cdots, \quad t \in (-1,1) \\
&\vdots
\end{aligned}
$$

Show that the function $K(x, y) = \exp(xy)$ defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel.

Exercise. Show that the function $K(x, y) = xy \exp(xy)$ defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel.

Exercise. Show that the function $K(x, y) = e^{x^2} e^{y^2} \exp(xy)$ defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel.

Show that the function $K(x, y) = \exp(-(x - y)^2)$ defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel.

Exercise. Show that the function $K(x,y) = \exp\left\{-\frac{(x-y)^2}{2\sigma^2}\right\}$ for any $\sigma^2 > 0$, defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel. This kernel is the general Gaussian kernel with "variance" σ^2.

Show that the function $K(x, y) = (1 + xy)^d$ with positive integer d, defined on $\mathbb{R} \times \mathbb{R}$ is a positive kernel.

Exercise. Show that the function $K(x, y) = -\log(1 - xy)$, defined on $x, y \in (-1, 1)$ is a positive kernel. Recall that for any $t \in (-1, 1)$,

$$-\log(1-t) = t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \frac{t^5}{5} + \cdots$$

Exercise. Show that the function $K(x,y) = \frac{1}{1-xy}$, defined on $x, y \in (-1,1)$ is a positive kernel.

Kernel Method

- Introduction
- Why (positive) kernel?
- Example of KRR with artificial data
- Examples of positive kernels on $\mathbb{R} \times \mathbb{R}$
- **Examples of positive kernels on high dimensional spaces**

- Inner product kernel: $K(x,y) = \langle x, y \rangle$ for $x, y \in \mathbb{R}^n$.
  Since for $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$, and coefficients $c_1, \ldots, c_n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle x_i, x_j \rangle$$
$$= \left\langle \sum_{i=1}^{n} c_i x_i, \sum_{j=1}^{n} c_j x_j \right\rangle = \left\| \sum_{i=1}^{n} c_i x_i \right\|^2 \geq 0.$$

- Write $x = (x^1, \cdots, x^p)$ for $x \in \mathbb{R}^p$. Let $M(u,v)$ defined on $\mathbb{R} \times \mathbb{R}$ be a positive kernel. Then $K(x,y) := M(x^1, y^1)$ defined on $x, y \in \mathbb{R}^p$ is also a kernel. In fact,

$$\sum\sum c_i c_j K(x_i, x_j) = \sum\sum c_i c_j M(x_i^1, x_j^1) \geq 0.$$

- If M and N are two positive kernels, then $M + N$ is also a positive kernel. In fact,

$$\sum\sum c_i c_j (M(x_i, x_j) + N(x_i, x_j))$$
$$= \left(\sum\sum c_i c_j M(x_i, x_j)\right) + \left(\sum\sum c_i c_j N(x_i, x_j)\right) \geq 0$$

- If M and N are positive kernels then MN is also a positive kernel. This follows the classical Schur product theorem. Before introducing it we need the notation

(Hadamard product)

$$\begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \cdot & \cdot & \cdots & \cdot \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix} \circ \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,n} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,n} \\ \cdot & \cdot & \cdots & \cdot \\ B_{m,1} & B_{m,2} & \cdots & B_{m,n} \end{bmatrix}$$

$$:= \begin{bmatrix} A_{1,1}B_{1,1} & A_{1,2}B_{1,2} & \cdots & A_{1,n}B_{1,n} \\ A_{2,1}B_{2,1} & A_{2,2}B_{2,2} & \cdots & A_{2,n}B_{2,n} \\ \cdot & \cdot & \cdots & \cdot \\ A_{m,1}B_{m,1} & A_{m,2}B_{m,2} & \cdots & A_{m,n}B_{m,n} \end{bmatrix}$$

Example. Let $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, and let $B = \begin{bmatrix} 5 & 6 \\ 7 & 10 \end{bmatrix}$. Find $A \circ B$.

Solution. By definition,

$$A \circ B = \begin{bmatrix} 1 \times 5 & 2 \times 6 \\ 3 \times 7 & 4 \times 10 \end{bmatrix} = \begin{bmatrix} 5 & 12 \\ 21 & 40 \end{bmatrix}.$$

Example. Let $A = \begin{bmatrix} 3 & 8 \\ 2 & 7 \end{bmatrix}$, and let $B = \begin{bmatrix} 1 & 1 \\ -3 & 2 \end{bmatrix}$. Find $A \circ B$.

Schur product theorem. If A and B are positive semi-definite, so is $A \circ B$.

Corollary. If M and N are both positive kernels, then MN is also a positive kernel.

Corollary. If M is a positive kernel, then M^d with d being a positive integer, is also a positive kernel.

The Gaussian kernel $K(x, y) := \exp\left(-\|x - y\|^2\right)$ defined on $\mathbb{R}^p$ is a positive kernel.

Exercise. The polynomial kernel $K(x, y) := (1 + \langle x, y \rangle)^d$ with a positive integer d, is a positive kernel.