1 核方法基础

根据模式识别理论，低维空间线性不可分的模式通过非线性映射到高维特征空间则可能实现线性可分。但是如果直接采用这种技术在高维空间进行分类或回归，则存在确定非线性映射函数的形式和参数、特征空间维数等问题，而最大的障碍则是在高维特征空间运算时存在的“维数灾难”。采用核方法可以有效地解决这样的问题。核函数方法是一种模块化方法，分为核函数设计和算法设计两个部分，它为处理许多问题提供了一个统一的框架。

1.1 映射函数定义

一个特征映射是

$$\phi : x \in \mathcal{X} \mapsto \phi(x) \in \mathcal{H} \tag{1}$$

式中，$\mathcal{X} \in \mathbb{R}^n$ 称为输入空间（Input space），$\mathcal{H} \in \mathbb{R}^N$ 称为特征空间（feature space）。利用映射函数 $\phi(\cdot)$ 将输入空间映射到特征空间，一般取它为 Hilbert 空间。

1.2 核函数定义

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 ϕ: $\mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \,. \tag{2}$$

已知映射函数 ϕ，可以通过 $\phi(x)$ 和 $\phi(x')$ 的内积求得核函数 $k(x, x')$。

由于直接计算 $k(x, x')$ 比较容易，而通过 $\phi(x)$ 和 $\phi(x')$ 计算 $k(x, x')$ 并不容易。核技巧的思想是，在学习与预测中只定义核函数 $k(x, x')$，而不用显示地定义映射函数 ϕ。

Note: All kernel functions are **positive definite**.

1.3 Hilbert 空间

Hilbert 空间从定义角度讲是“完备的内积空间”。具体来讲，“Hilbert 空间 = 线性空间 + 有内积 + 有范数 + 有完备性（极限）”。

A Hilbert space is a space on which an inner product is defined, along with the limits of all Cauchy sequences of functions.

Cauchy squence's definition: A sequence ${f_n}_{n=1}^{\infty}$ of elements in a normed space $\mathcal{H}$ is said to be a *Cauchy sequence* if for every $\epsilon > 0$, there exists $N = N(\varepsilon) \in \mathbb{N}$, such that for all $n, m \geq N$, $\|f_n - f_m\|_{\mathcal{H}} < \epsilon$.

1.4 再生核 Hilbert 空间

The reproducing kernel Hilbert space (RKHS)

1.4.1 Reproducing kernel Hilbert space (first definition)

设 $\mathcal{H}$ 是一个由定义在非空集合 $\mathcal{X}$ 上函数 $f, \mathcal{X} \mapsto \mathbb{R}$ 构成的 Hilbert 函数空间。若函数 k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 满足：

- $\forall x \in \mathcal{X}, \quad k(\cdot, x) \in \mathcal{H}$
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$（重构属性）

则称 k 为 $\mathcal{H}$ 的再生核函数，$\mathcal{H}$ 为再生核 Hilbert 空间。特别地，对于 $\forall x, y \in \mathcal{X}$，有 $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$。

1.4.2 Reproducing kernel Hilbert space (second definition)

$\mathcal{H}$ is an RKHS if for all $x \in \mathcal{X}$, the evaluation operator δ_x is bounded: there exists a corresponding $\lambda_x \geq 0$ such that $\forall f \in \mathcal{H}$,

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}} \tag{3}$$

1.5 积分算子

定义积分算子 T_k: $L_2(\mathcal{X}) \to L_2(\mathcal{X})$ 按下式确定

$$T_k f = (T_k f)(\cdot) = \int_{\mathcal{X}} k(\cdot, x') f(x') dx' \quad \forall f \in L_2(\mathcal{X}) \tag{4}$$

1.6 Mercer 理论

令 $\mathcal{X}$ 是 $\mathbb{R}^n$ 的紧集，k 是 $\mathcal{X} \times \mathcal{X}$ 上的连续对称函数，积分算子 T_k 是半正定的，即

$$\int_{\mathcal{X}} k(x, x') f(x) f(x')\, dx dx' \geq 0, \quad f \in L_2(\mathcal{X}) \tag{5}$$

等价于 k 是可以表示为 $\mathcal{X}\times\mathcal{X}$ 上的一致收敛序列的核函数

$$k(x,x')=\sum_{r=1}^{\infty}\lambda_r\psi_r(x)\psi_r(x')=\langle\phi(x),\phi(x')\rangle \tag{6}$$

其中

$$\phi:x\mapsto\left(\sqrt{\lambda_1}\psi_1(x),\sqrt{\lambda_2}\psi_2(x),\ldots\right)^T \tag{7}$$

$\lambda_r\geq 0$ 是 T_k 的特征值，$\psi_r\in L_2(\mathcal{X})$ 为对应于 λ_r 的特征向量（$\|\psi_r\|_{l2}=1$）。

1.7 The key questions

• point evaluation functional, operator of evaluation 如何理解？与 RKHS 的定义有什么关系？

求值泛函定义：设 $\mathcal{H}$ 是一个由定义在非空集合 $\mathcal{X}$ 上函数空间，对于一个固定的 $x\in\mathcal{X}$，定义映射 $\delta_x:\mathcal{H}\mapsto\mathbb{R}$ 满足 $\delta_x f=f(x)$，则 δ_x 是在 x 点的求值泛函。显然，求值泛函 δ_x 是一个线性泛函，对于 $\forall f,g\in\mathcal{H}$ 和 $\forall\alpha,\beta\in\mathbb{R}$，有

$$\delta_x(\alpha f+\beta g)=(\alpha f+\beta g)(x)=\alpha f(x)+\beta g(x)=\alpha\delta_x(f)+\beta\delta_x(g).$$

RKHS 定义：$\mathcal{H}$ 是再生核 Hilbert 空间，当且仅当对于 $\forall x\in\mathcal{X}$，求值泛函 δ_x 是有界的，即存在一个与 x 有关的常量 $\lambda_x\geq 0$ 满足对于 $\forall f\in\mathcal{H}$，有

$$|f(x)|=|\delta_x f|\leq\lambda_x\,\|f\|_{\mathcal{H}}$$

• 任意给定 RKHS 中的两个元素 f 和 g，他们的内积是怎么定义的？

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

1.$<\alpha_1 f_1+\alpha_2 f_2,g>_{\mathcal{H}}=\alpha_1<f_1,g>_{\mathcal{H}}+\alpha_2<f_2,g>_{\mathcal{H}}$

2.$<f,g>_{\mathcal{H}}=<g,f>_{\mathcal{H}}$

3.$<f,f>_{\mathcal{H}}\geq 0 and <f,f>_{\mathcal{H}}=0 if and only if f=0$

• Representer theorem 阐述了什么内容？有什么意义？

定理：The solution to

$$\min_{f\in\mathcal{H}}[L(f(x_1),\ldots,f(x_n))+\Omega(\|f\|_{\mathcal{H}}^2)]$$

takes the form

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

If Ω is strictly increasing, all solutions have this form.

意义：简化了正则化的经验风险最小化问题；将高维甚至无限维的计算问题简化为标量系数的优化问题；为一般机器学习问题推广到可实现算法提供理论基础。

- Moore-Aronszajn theorem 阐述了什么内容？有什么意义？

定理：每一个正定核 k 都有唯一一个与之相对应的再生核 Hilbert 空间。

意义：Functions in the RKHS can be written as linear combinations of feature maps,

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

as well as the limits of Cauchy sequences (where we can allow $m \to \infty$).

- Mercer's theorem 阐述了什么内容？有什么意义？

定理：If k is a continuous kernel of a positive definite intergral operator on $L_2(\mathcal{X})$ (where $\mathcal{X}$ is some compact space),

$$\int_{\mathcal{X}} k(x, x') f(x) f(x') \, dx \, dx' \geq 0,$$

it can be expanded as

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$$

using eigenfunctions ψ_i and eigenvalues $\lambda_i \geq 0$.

意义：证明核函数可以构成一个 RKHS；核函数可以表示为 $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x') = \langle \phi(x), \phi(x') \rangle$。

2 核对齐基础

核矩阵定义：数据矩阵 X 和 Y 中各取任意向量 x_i 和 y_j 两两之间的核函数值所组成的矩阵。

$$\mathrm{K}(\mathrm{X},\mathrm{Y})=\begin{bmatrix} K(x_1,y_1) & K(x_1,y_2) & \cdots & K(x_1,y_n) \\ K(x_2,y_1) & K(x_2,y_2) & \cdots & K(x_2,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_m,y_1) & K(x_m,y_2) & \cdots & K(x_m,y_n) \end{bmatrix} \tag{8}$$

其中 $K(\cdot,\cdot)$ 为核函数，常见的有线性核以及 RBF 核，矩阵 $X\in\mathbb{R}^{m\times d}$ 和 $Y\in\mathbb{R}^{n\times d}$ 为数据矩阵，每一行代表一个样本，定义为

$$X=[x_1,x_2,\ldots,x_m]^\top$$

$$Y=[y_1,y_2,\ldots,y_m]^\top$$

2.1 核对齐 KTA

Given an (unlabelled) sample $S=\{x_1,\ldots,x_m\}$, we use the following inner product between Gram matrices, $\langle \mathrm{K},\mathrm{K}'\rangle_F=\sum_{i,j=1}^m K(x_i,x_j)K'(x_i,x_j)$.

2.1.1 核对齐（核函数）定义

The (empirical) alignment of a kernel function K with a kernel function K' with respect to the sample S is the quantity

$$A=\frac{E[KK']}{\sqrt{E[K^2]E[K'^2]}} \tag{9}$$

2.1.2 核对齐（核矩阵）定义

The (empirical) alignment of a kernel matrix K with a kernel matrix K' with respect to the sample S is the quantity

$$\widehat{A}=\frac{\langle \mathrm{K},\mathrm{K}'\rangle_F}{\|\mathrm{K}\|_F\,\|\mathrm{K}'\|_F} \tag{10}$$

2.2 中心化核函数

Let D be the distribution according to which training and test points are drawn. Centering a feature mapping $\phi : \mathcal{X} \mapsto \mathcal{H}$ consists of replacing it by $\phi - E_x[\phi]$, where E_x denotes the expected value of ψ when x is drawn according to the distribution D. Centering a positive definite symmetric(PDS) kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ consists of centering any feature mapping ψ associated to K. Thus, the centered kernel K_c associated to K is defined for all $x, x' \in mathcalX$ by

$$\begin{aligned} K_c(x, x') &= (\phi(x) - E_x[\phi(x)])^\top (\phi(x') - E_{x'}[\phi(x')]) \\ &= K(x, x') - E_x[K(x, x')] - E_x[K(x, x')] + E_{x,x'}[K(x, x')] \end{aligned} \tag{11}$$

2.3 中心化核矩阵

Similar definitions can be given for a finite sample $S = (x_1, \ldots, x_m)$ drawn according to D: a feature vector $\phi(x_i)$ with $i \in [1, m]$ is then centered by replacing it with $\phi(x_i) - \overline{\phi}$, with $\overline{\phi} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$, and the kernel matrix K associated to K and the sample S is centered by replacing it with $\mathrm{K_c}$ defined for all $i, j \in [1, m]$ by

$$[\mathrm{K_c}]_{ij} = \mathrm{K}_{ij} - \frac{1}{m} \sum_{i=1}^{m} \mathrm{K}_{ij} - \frac{1}{m} \sum_{j=1}^{m} \mathrm{K}_{ij} + \frac{1}{m^2} \sum_{i,j=1}^{m} \mathrm{K}_{ij}. \tag{12}$$

2.4 核对齐 CKA

2.4.1 核对齐 (核函数) 定义

Let K and K' be two kernel functions defined over $\mathcal{X} \times \mathcal{X}$ such that $0 < E[K_c^2] < +\infty$ and $0 < E[{K'}_c^2] < +\infty$. Then, the alignment between K and K' is defined by

$$\rho(K, K') = \frac{E[K_c {K'}_c]}{\sqrt{E[K_c^2] E[{K'}_c^2]}}. \tag{13}$$

The notion of alignment seeks to capture the correlation between the random variables $K(x, x')$ and $K'(x, x')$ and one could think it natural, as for the standard correlation coefficients, to consider the following definition:

$$\rho(K, K') = \frac{E[(K - E[K])(K' - E[K'])]}{\sqrt{E[(K - E[K])^2]E[(K' - E[K'])^2]}} \tag{14}$$

Note: $0 \leq \rho(K, K') \leq 1$.

2.4.2 核对齐 (核矩阵) 定义

Let $\mathrm{K} \in \mathbb{R}^{m \times m}$ and $\mathrm{K}' \in \mathbb{R}^{m \times m}$ be two kernel matrices such that $\|K_c\|_F \neq 0$ and $\|K'_c\|_F \neq 0$. Then, the alignment between K and K′ is defined by

$$\widehat{\rho}(\mathrm{K}, \mathrm{K}') = \frac{\langle \mathrm{K_c}, \mathrm{K'_c} \rangle_F}{\|\mathrm{K_c}\|_F \, \|\mathrm{K'_c}\|_F}. \tag{15}$$

Note: $\mathrm{K}_c = U_m K U_m = \left[I_m - \frac{\mathbf{1}\mathbf{1}^\top}{m}\right] \mathrm{K} \left[I_m - \frac{\mathbf{1}\mathbf{1}^\top}{m}\right]$, $\mathrm{K'}_c$ the same as K_c. $0 \leq \widehat{\rho}(\mathrm{K}, \mathrm{K}') \leq 1$.

2.5 Single-stage Alignment-based Algorithm

2.6 Two-stage Alignment-based Algorithm

2.6.1 Independent Alignment-based Algorithm (align)

It determines each mixture weight μ_k independently.

The optimization problem with an L_q-norm constraint on μ with $q > 1$:

$$\begin{aligned} &\max_{\mu} \ \widehat{\rho}_u(\mathrm{K}_\mu, \mathrm{K}_Y) = \left\langle \sum_{k=1}^{p} \mu_k \mathrm{K}_{kc}, \mathrm{K}_Y \right\rangle_F \\ &\text{subject to}: \ \sum_{k=1}^{p} \mu_k^q \leq \Lambda. \end{aligned} \tag{16}$$

Let μ^* be the solution of the optimization problem, then

$$\mu_k^* \propto \langle \mathrm{K}_{kc}, \mathrm{K}_Y \rangle_F^{\frac{1}{q-1}} \tag{17}$$

2.6.2 Alignment Maximization Algorithm (alignf)

It determines the mixture weights μ_k jointly by seeking to maximize the alignment between the convex combination kernel $\mathrm{K}_\mu = \sum_{k=1}^{p} \mu_k \mathrm{K}_k$ and the target kernel $\mathrm{K}_Y = \mathrm{y}\mathrm{y}^\top$.

Linear combination with $\mathcal{M} = \{\mu : \|\mu\|_2 = 1\}$.

Convex combination with $\mathcal{M} = \{\mu : \|\mu\|_2 = 1 \wedge \mu \geq 0\}$.

Let

$$a = (< \mathrm{K}_{1c}, \mathrm{yy}^\top >_F, \ldots, < \mathrm{K}_{pc}, \mathrm{yy}^\top >_F)^\top, \tag{18}$$

and let M denote the matrix defined by

$$\mathrm{M}_{kl} =< \mathrm{K}_{kc}, \mathrm{K}_{lc} >_F \tag{19}$$

Note: M and M^{-1} are PSD.

The alignment maximization problem

$$\max_{\mu \in \mathcal{M}} \widehat{\rho}\,(\mathrm{K}_\mu, \mathrm{K}_Y) \tag{20}$$

cam be equivalently written as the following optimization problem

$$\mu^\star = \arg\max_{\mu \in \mathcal{M}} \frac{\mu^\top \mathrm{aa}^\top \mu}{\mu^\top \mathrm{M} \mu} \tag{21}$$

Let $\mu^\star$ be the solution of the optimization problem, then

$$\mu^\star = \frac{\mathrm{M}^{-1}\mathrm{a}}{\|\mathrm{M}^{-1}\mathrm{a}\|} \tag{22}$$