核方法（2）-核的性质

Y.Q. Wang

2016 年 1 月 25 日

1 摘要

上一回且说到核方法到底是什么，为了具体说明，以线性回归为例，通过找原问题的对偶形式，做特征映射，改写原问题中的 Gram 矩阵，从而达到了核方法的目的：将特征空间做非线性映射，用线性方法在映射后的空间中找到数据的模式。通过示例可以发现，核方法的物理意义其实是将数据点映射到一个特定的空间，在特定的空间里，构建数据点对间的关系矩阵，依赖关系矩阵，发现数据的模式。由于映射关系可以是非线性的，甚至可以映射到一个无穷维的空间中去，所以核方法在理论上可以解决任何问题。在核方法中，原始信息不再被使用，Gram 矩阵会包含所有的有效信息，因此核方法的关键是选取一个合适的 kernel，构建 Gram 矩阵。一般化的，Gram 矩阵的构建过程中也可以不对特征空间做显式的特征映射，这样将极大的增加 kernel 的可选择性，进而优化对问题的求解。那么，在基本了解核方法的目的、过程和物理意义后，接下来需要进一步的开始明确核方法中的关键概念 kernel。到底什么是 kernel？又该怎样构造一个 kernel 呢？

2 名词辨析：kernel

在机器学习中，kernel 一词常会在不同的场景下出现，因而在正式介绍核方法中的 kernel 概念前，简要明确 kernel 一词在各个场景下的不同涵义。

- Kernel density estimation/Kernel smoother： 这两个场景下的 “kernel” 的概念基本相似，指的是数据点对通过核函数（kernel function）产生映射的这种映射形式。目标是非参化的估计概率密度函数（kernel density estimation）或实值函数 $f(x)$。方法是将数据点做 n 段切分，在每段切分的区间中，计算所有的数据点和分段中心点在核函数作用下做变换。例如，最常见的方法是采用 Gaussian Kernel，$\mathcal{K}(\mathbf{x}, \mathbf{x}_0) = \exp(-(\mathbf{x} - \mathbf{x}_0)^2/2\sigma^2)$。这类估计方法可以有效的将不规则的数据点用一条（分段）平滑的曲线进行拟合。（见图1）
- Kernel 在 CNN 中：在卷积神经网络中，“kernel” 指 $R^d \times R^d \to R$ 的映射方式。目标是利用 kernel 构建图像某一区域的像素矩阵的抽象。这种映射的计算方式和将要阐述的 kernel 形似但不具有相关性质，为了避免产生不必要的误会，不再做展开。（见图2）

- Kernel 在核方法中：这里的 “kernel” 指的是核函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 的映射形式（核函数是具体的函数形式，kernel 指的是一类方式）。目标将原特征空间映射到高维空间中去，使得在高维空间中可以用线性函数发现数据的模式。其性质将在本章内容中详细阐述。

- Reproducting kernel 在 Reproducing Kernel Hilbert Space(RKHS)：（话外音：好吧，这个名词洒家已经木有办法翻译成中文了……）这里的 “reproducing kernel” 有些特殊，在于这种 kernel 满足 reproducing property，即 $\forall f \in \mathcal{H}, f(\mathbf{x}) = \langle f, \mathcal{K}_{\mathbf{x}} \rangle_{\mathcal{H}}$，其中 $\mathcal{K}_{\mathbf{x}} \in \mathcal{H}$。（这里用到了泛函的知识，具体可以先不做了解。）若令 $\phi(\mathbf{x}) = \mathcal{K}_{\mathbf{x}}$，则 reproducing kernel 可以用于核方法。

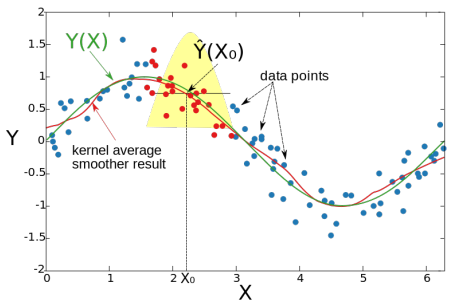

Figure 1: Kernel density estimation/Kernel smoother 示例

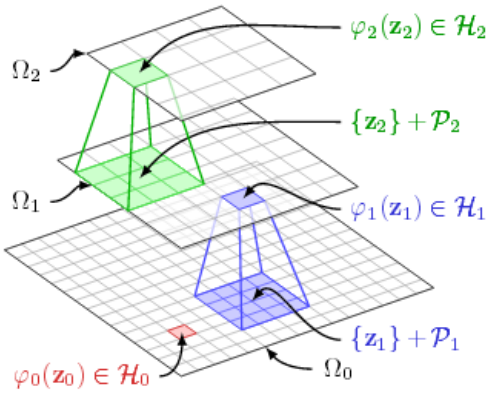

Figure 2: Kernel 在 CNN 中的作用方式

3 希尔伯特空间（Hilbert spaces）

希尔伯特空间 $\mathcal{H}$ 定义了（向量）内积空间（inner product space），它是构造 kernel 的关键基础。从数学定义上着手理解希尔伯特空间的确会有些困扰，用一些简单的例子逐步引出希尔伯特空间的定义。

例 1： 欧式空间 R^2。令 $\mathbf{a} = (a_1, a_2)^T, \mathbf{b} = (b_1, b_2)^T$ 是欧式空间 R^2 的两个向量，这两个向量的内积为 $\langle \mathbf{a}, \mathbf{b} \rangle = a_1 b_1 + a_2 b_2$。

例 2： 收敛的积分。积分可以看作将函数的定义域分为等间隔的区间（Δx），计算各区间的面积，累加结果。收敛的积分的极限方式定义为：$\lim_{n \to \infty} \sum_{i=1}^{\infty} f(x_{i-1}) \Delta x = c$。

这里所举例的欧式空间和收敛的积分都是典型的希尔伯特空间。希尔伯特空间必须是可分的（separable）和完备的（complete）。“可分” 指的是内积（$\langle \mathbf{x}, \mathbf{y} \rangle$）的元素 $\mathbf{x}, \mathbf{y}$ 必须是由一组正交基

定义的线性空间的向量。“完备” 指的是内积的结果必须小于 ∞。所有符合性质的内积空间就称为希尔伯特空间。

以上是非常山寨的个人理解。接下来放出原版解释以供大家参考。

这是直接从希尔伯特空间出发的定义：

A Hilbert Space $\mathcal{H}$ is a strict inner product space with the additional properties that is *separable* and *complete*. Completeness refers to the property that every Cauchy sequence $\{h_n\}_{n\geq 1}$ of elements of $\mathcal{H}$ converges to a element $h \in \mathcal{H}$, where a Cauchy sequence is one satisfying the property that

$$\sup_{m>n} \|h_n - h_m\| \to 0, \text{as } n \to \infty.$$

A space $\mathcal{H}$ is seperable if there is a countable set of elements $h_1, \ldots, h_i$ of $\mathcal{H}$ such that for all $h \in \mathcal{H}$ and $\epsilon > 0$ there exists i such that

$$\|h_i - h\| < \epsilon.$$

还有更为详细的wikipedia 版。

还可以这么理解：线性空间引入了基的概念，从而可以在基坐标中描述一个点（向量）；赋范线性空间引入了长度的概念，从而量化了线性空间中一个向量的长度；由于线性空间和赋范线性空间不能描述极限点的情况（空间极限的概念），为了完备化这一特殊情况，引入了完备的内积空间概念，即希尔伯特空间。（摘自知乎）

希尔伯特空间在物理上可以描述两个向量的差异（covariance），并衍生出了向量的距离，向量的夹角，向量的范数等概念。

4 Gram 矩阵

Gram 矩阵 $\mathbf{G}$ 也被称作核方法中的 kernel matrix。定义 Gram 矩阵中的元素 $\mathbf{G}_{ij}$ 为

$$\mathbf{G}_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{1}$$

根据定义可知 Gram 矩阵是对称的，这就导致 Gram 矩阵丢失了原始数据中的方向信息，同时将原始输入变换为 Gram 矩阵还会丢失一些其他信息。利用核方法求解问题，意味着所有的原始输入必须要变换为 Gram 矩阵（参见介绍章节中的内容），故而 Gram 矩阵成为了利用核方法的瓶颈。接下来就 Gram 矩阵的性质，如何构建 Gram 矩阵做详细的阐述。

4.1 特征值和特征向量

（线性代数的入门知识，不多做介绍）提一下 Rayleigh 商（Rayleigh quotient）（会在未来的章节中用到），Rayleigh 商可以用来求解矩阵的特征值，

$$\frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \lambda\frac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \lambda, \tag{2}$$

其中 $\mathbf{A}, \mathbf{v}, \lambda$ 分别指矩阵，特征向量和特征值。特征值非 0 的任意特征向量对都是正交的。矩阵 $\mathbf{A}$ 的特征值集合 $\lambda(\mathbf{A})$ 称为谱（spectrum）。另外，矩阵 $\mathbf{A}^k$ 的谱就是 $\{\lambda^k : \lambda \in \lambda(\mathbf{A})\}$。

基于 Rayleigh 商可以定义求解矩阵 $\mathbf{A}$ 的最大特征值问题

$$\max_{\mathbf{v}} \frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \tag{3}$$

注意到在 Rayleigh 商中，特征向量的大小是可以缩放的，故而在接下来的讨论中归一化所有的特征向量，使得 $\|v\| = 1$。

借鉴 Rayleigh 商的形式还可以定义矩阵的谱模（spectral norm）

$$\max_{\mathbf{v}} \frac{\|\mathbf{A}\mathbf{v}\|}{\|\mathbf{v}\|} = \sqrt{\max_{\mathbf{v}} \frac{\mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v}}{\mathbf{v}^T\mathbf{v}}} \tag{4}$$

4.2 半正定矩阵

根据半正定矩阵的定义，半正定矩阵 $\mathbf{A}$ 的特征值必须是非负的，由 Rayleigh 商可得，

$$\mathbf{v}^T\mathbf{A}\mathbf{v} \geq 0 \tag{5}$$

对所有的特征向量 $\mathbf{v}$ 成立。

推论 1：当且仅当存在一个实值矩阵 $\mathbf{B}$ 使得 $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ 时，矩阵 $\mathbf{A}$ 为半正定矩阵。

证明：令 $\mathbf{A} = \mathbf{B}^T\mathbf{B}$，对任意的特征向量 $\mathbf{v}$ 都有

$$\mathbf{v}^T\mathbf{A}\mathbf{v} = \mathbf{v}^T\mathbf{B}^T\mathbf{B}\mathbf{v} = \|\mathbf{B}\mathbf{v}\|^2 \geq 0.$$

根据半正定矩阵的定义，证明成立。

注意矩阵 $\mathbf{B}$ 的形式并非是唯一的，例如，利用 Cholesky 分解可以将矩阵 $\mathbf{A}$ 分解为两个上三角矩阵的乘积。

4.3 Gram 矩阵的特点

在一大堆预备知识后，我们正式引出主角 “Gram 矩阵” 登场，并对它的特点逐一描述。

首先，是的，**Gram 矩阵是半正定矩阵。**

证明：根据 Gram 矩阵定义，

$$\mathbf{G}_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \text{, for } i, j = 1, \ldots, l.$$

对任意特征向量 $\mathbf{v}$ 都有

$$\begin{aligned}
\mathbf{v}^T \mathbf{G} \mathbf{v} &= \sum_{i,j=1}^{l} v_i v_j \mathbf{G}_{ij} == \sum_{i,j=1}^{l} v_i v_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\
&= \left\langle \sum_{i=1}^{l} v_i \phi(\mathbf{x}_i), \sum_{j=1}^{l} v_j \phi(\mathbf{x}_j) \right\rangle \\
&= \left\| \sum_{i=1}^{l} v_i \phi(\mathbf{x}_i) \right\|^2 \geq 0.
\end{aligned}$$

证明成立。

由 Gram 矩阵的特质，决定了在一些情况下定义核函数并非一定要显式地对特征空间做映射。关键是确保核函数 $\mathcal{K} : \mathbf{X} \times \mathbf{X} \to R$ 具有 “半正定函数” 的性质。利用Mercer 定理定义半正定 kernel 如下，

定义 1：对称连续函数 $\mathcal{K} : \mathbf{X} \times \mathbf{X} \to R$ 是半正定 kernel，当且仅当该函数满足

$$\sum_{i,j=1}^{n} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) c_i c_j \geq 0$$

对有限序列中的所有元素 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 以及所有选择的实数 $c_1, \ldots, c_n$ 成立。

定理 1：定义在连续或者可数域的函数 $\mathcal{K} : \mathbf{X} \times \mathbf{X} \to R$，当且仅当其满足半正定性质时，其可被分解为

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

在希尔波特空间 $\mathcal{H}$ 中的特征映射 ϕ 的内积。

定理 1 的提出为 kernel 的构造提供了理论保证。

5 Kernel 构造

只要构造的 kernel 满足半正定性质就是一个有效的 kernel。这就使得在 kernel 构造的过程中可以通过组合的方式去获得更为复杂的 kernel。这里将 kernel 的运算规则进行整理，归纳为 6 点，称为 kernel 的闭包性质

闭包性质：令 $\mathcal{K}_1$，$\mathcal{K}_2$ 是定义在 $\mathbf{X} \times \mathbf{X}, \mathbf{X} \subseteq R^n$ 的 kernel，$a \in R^+$，$f(\cdot)$ 是定义在 $\mathbf{X}$ 上的实值函

数，$\phi : \mathbf{X} \to R^N$，$\mathcal{K}_3$ 是定义在 $R^N \times R^N$ 的 kernel，$\mathbf{B}$ 是大小为 $n \times n$ 对称半正定矩阵。那么，下面的函数都是 kernel：

1) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \mathcal{K}_1(\mathbf{x}, \mathbf{z}) + \mathcal{K}_2(\mathbf{x}, \mathbf{z})$

2) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = a\mathcal{K}_1(\mathbf{x}, \mathbf{z})$

3) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \mathcal{K}_1(\mathbf{x}, \mathbf{z})\mathcal{K}_2(\mathbf{x}, \mathbf{z})$

4) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$

5) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \mathcal{K}_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$

6) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{B}\mathbf{z}$.

另外，以下三种情况也可以构造 kernel：
令 $\mathcal{K}_1(\mathbf{x}, \mathbf{z})$ 是定义在 $\mathbf{X} \times \mathbf{X}$ 的 kernel，$p(x)$ 是系数为正的多项式。那么，下面的函数都是 kernel：

1) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = p(\mathcal{K}_1(\mathbf{x}, \mathbf{z}))$

2) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \exp(\mathcal{K}_1(\mathbf{x}, \mathbf{z}))$

3) $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \exp(-\left\|\mathbf{x} - \mathbf{z}\right\|^2 / (2\sigma^2))$.

提示：可以用一个非常简单的 trick 证明以上的很多等式。将特征映射做链式拆解 $\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x})]$，那么

$$\begin{aligned}\mathcal{K}(\mathbf{x}, \mathbf{z}) &= \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \langle [\phi_1(\mathbf{x}), \phi_2(\mathbf{x})], [\phi_1(\mathbf{z}), \phi_2(\mathbf{z})] \rangle \\ &= \langle [\phi_1(\mathbf{x}), \phi_1(\mathbf{z})] \rangle + \langle [\phi_2(\mathbf{x}), \phi_2(\mathbf{z})] \rangle \\ &= \mathcal{K}_1(\mathbf{x}, \mathbf{z}) + \mathcal{K}_2(\mathbf{x}, \mathbf{z}).\end{aligned}$$

其他证明从略。