Kernel Method 01 Background

Chen Gong

20 November 2019

在 Support Vector Machine 的章节中，我们已经分析了支持向量机前面“两宝”，也就是间隔和对偶，而第三宝，核技巧在这里我们需要抽出来将分析。其实，我最开始学习核的时候，真的是一脸懵逼，这玩意到底是个什么鬼？来龙去脉是什么？这这节有关于 Kernel Method 的背景介绍中，我想分析一下，我们为什么要使用核？以及怎么用核？来给大家一个直观的感受。

本小节主要从 Kernel Method，Kernel Function 和 Kernel Trick，三个方面来进行分析和讨论，我们为什么要用核？我们怎么样用核？

1 Kernel Method

核方法是一种思想，在 Cover Theorem 中提出：高维空间比低维空间更容易线性可分。这句话非常的直观，我们想想就理解了，这里我不做出详细的证明。在这里我们举一个例子，对于经典的线性不可分问题异或问题，图像描述如下所示：

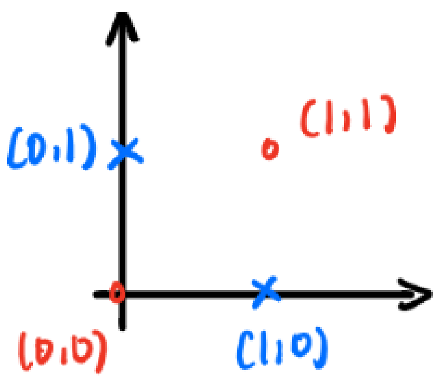

图 1: 异或问题的图像在二维空间中的描述

这二维空间中的点可以被我们记为 $X = (x_1, x_2)$，如果我们使用一个变换函数，将其变换到三维空间中就会发生有意思的事情。我们设定一个变换函数为 $\phi(X)$，将二维空间中的点，变换到一个三维空间 Z 中，并且令 $Z = (x_1, x_2, (x_1 - x_2)^2)$，那么我们再来看看异或问题在三维空间中点的分布，我们惊奇的发现变得线性可分了：

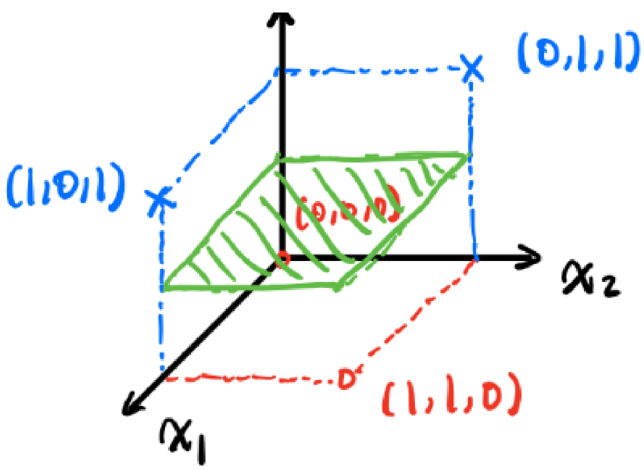

图 2: 异或问题的图像在三维空间中分布

通过这个例子，想想大家都直观的感受到了高维空间带来的好处。实际上对于解决非线性问题，我们有两种思路：

1. 也就是之前提到的，由 Perceptron Layer Analysis (PLA) 引出的多层感知机 (Multilayer Perceptron) 也就我们经常听到的神经网络，以及之后发展得到的 Deep Learning。

2. 而第二种思路就是通过非线性变换 $\phi(x)$，将非线性可分问题转换为线性可分问题。上述的异或问题，可以表述为：

$$\mathcal{X} = (x_1, x_2) \stackrel{\phi(x)}{\longmapsto} \mathcal{Z} = (x_1, x_2, (x_1 - x_2)^2) \tag{1}$$

第二类方法也就是我们讨论的重点，其实在我们机器学习理论的研究中，第二种方法有很大的威力，大部分同学在学习的时候都会忽掉，例子可以看看之前发的再生核希尔伯特空间。

2 Kernel Function

核函数，从模型的角度讲可以带来给非线性带来高维的转换，这个我们上面已经分析过了。从优化的角度讲可以为对偶带来内积，这两个角度可以合在一起看看。

以我们之前学习的 Hard Margin SVM 为例，原问题和对偶问题的表述为：

$$\begin{gathered}\begin{cases} \max_{w,b} \ \frac{1}{2}w^Tw \\ s.t. \quad 1 - y_i(w_i^Tx + b) \leq 0 \end{cases} \\ \begin{cases} \min_\lambda \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \lambda_i\lambda_j y_i y_j (x_i^Tx_j) - \sum_{i=1}^N \lambda_i \\ s.t. \quad \lambda_i \geq 0, \ \sum_{i=1}^N \lambda_i y_i = 0 \end{cases}\end{gathered} \tag{2}$$

在我们的对偶问题中，是不是有一个 $x_i^Tx_j$。在线性可分问题中，我们直接计算就好了，在线性不可分问题中，就需要将 x 通过一个变换 ϕ 转换到高维空间中。那么 $x_i^Tx_j$ 就变成了 $\phi(x_i^T)\phi(x_j)$。那么我们就将两个角度的分析联系起来了。那么核函数我们可以定义为：

对于一个 $K(x, x') = \phi(x)^T \cdot \phi(x) =< \phi(x), \phi(x') >$，

有 $\forall x, x' \in \mathcal{X}, \exists \phi : x \mapsto z \ s.t. \ K(x, x') = \phi(x)^T \cdot \phi(x')$。则称 $K(x, x')$ 是一个核函数。比如：

$$K(x, x') = exp(-\frac{(x - x')^2}{2\sigma^2}) \tag{3}$$

3 Kernel Track

下面我们需要引入核技巧了，也就是想想，核函数有什么用？前面我们讲到将 x 通过一个变换 ϕ 转换到高维空间。但是，有可能 $\phi(x)$ 的维度非常的高，甚至是无限维的，那么这将变得非常的难求。如果还要继续求 $\phi(x_i^T)\phi(x_j)$，这个计算量恐怕会要原地爆炸。

大家通过上面的表达会发现我们实际上关注的不是 $\phi(x_i)$ 本身，而是 $\phi(x_i^T)\phi(x_j)$。那么，我们完全可直接求跳过求 $\phi(x_i)$ 的过程，然后 $\phi(x_i^T)\phi(x_j)$。我们看看核函数的定义，是不是 $K(x_i, x_j)$ 就等于 $\phi(x_i^T)\phi(x_j)$。这就省去了很多麻烦的计算过程，核函数在这实在是太好用了，这就是核技巧的思想。总的来说，就是非线性转换上的一个内积。

我们为什么引入 kernel？就是原来的方法有不足，不能解决非线性可分问题。所以，我们想到利用核函数将 $\mathcal{X} \mapsto \mathcal{Z}$，到更高维的空间来转换成线性可分问题。又因为 $\phi(x_i)$ 的计算很难，我们有想到用核函数来直接求 $\phi(x_i^T)\phi(x_j)$。这里面其实是一环扣一环的，逻辑性非常的强。

对于前面讨论的线性可分问题 Perceptron Layer Analysis 和 Hard Margin SVM。允许出现错误就出现了 Pocket Algorithm 和 Soft Margin SVM。进一步如果是严格的非线性问题，引入了 $\phi(x)$ 就得到了 $\phi(x) + PLA$ 和 $\phi(x) + Hargin$ (Kernel SVM)，就是将输入变量的内积转换为核函数。

那么，我们怎么找一个核函数，核函数具有怎样的性质？我们在下一小节中进行分析。

Kernel Method 02 The Definition of Positive Kernel Function

Chen Gong

21 November 2019

上一节中，我们已经讲了什么是核函数，也讲了什么是核技巧，以及核技巧存在的意义是什么。我们首先想想，上一小节我们提到的核函数的定义。

对于一个映射 K，我们有两个输入空间 $\mathcal{X}\times\mathcal{X},\mathcal{X}\in\mathbb{R}^p$，可以形成一个映射 $\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$。对于，$\forall\ x,z\in\mathcal{X}$，存在一个映射 $\phi:\mathcal{X}\mapsto\mathbb{R}$，使得 $K(x,z)=<\phi(x),\phi(z)>$。那么这个 $K(\cdot)$，就被我们称为核函数。(<> 代表内积运算)

这既是我们上一节中将的核函数的定义，实际上这个核函数的精准定义，应该是正定核函数。在本小节中，我们将会介绍核函数的精准定义，什么是正定核函数？并介绍内积和希尔伯特空间 (Hilbert Space) 的定义。这一小节虽然看着会有些枯燥，实际上非常的重要。

1 核函数的定义

核函数的定义，也就是对于一个映射 K，存在一个映射 $\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$，对于 $x,z\in\mathcal{X}$ 都成立，则称 $K(x,z)$ 为核函数。

对比一下，我们就会发现，这个定义实际上比我们之前学的定义要简单很多。好像是个阉割版，下面我们来介绍两个正定核的定义方法。

2 正定核的定义

正定核函数的定义有两个，我首先分别进行描述一下：

2.1 第一个定义

现在存在一个映射 $K:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$。对于 $\forall x,z\in\mathcal{X}$。如果存在一个 $\phi:\mathcal{X}\mapsto\mathbb{R}^p$，并且 $\phi(x)\in\mathcal{H}$，使得 $K(x,z)=<\phi(x),\phi(z)>$，那么称 $K(x,z)$ 为正定核函数。

2.2 第二个定义

对于一个映射 $K:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$，对于 $\forall x,z\in\mathcal{X}$，都有 $K(x,z)$。如果 $K(x,z)$ 满足，1. 对称性；2. 正定性；那么称 $K(x,z)$ 为一个正定核函数。

我们来分析一个，首先什么是对称性？这个非常的好理解，也就是 $K(x,z)=K(z,x)$。那么什么又是正定性呢？那就是任取 N 个元素，$x_1,x_2,\cdots,x_N\in\mathcal{X}$，对应的 Gram Matrix 是半正定的，其中 Gram Matrix 用 K 来表示为 $K=[K(x_i,x_j)]$。

对于第一个对称性，我们其实非常好理解，不就是内积嘛！有一定数学功底的同学一定知道，内积和距离是挂钩的，距离之间一定是对称的。那么正定性就要好好讨论一下了。我们知道这两个定义之间是等价的，为什么会有正定性呢？我们需要进行证明，这个证明可以被我们描述为：

$$K(x,z) = < \phi(x), \phi(z) > \Longleftrightarrow \text{Gram Matrix 是半正定矩阵}$$

这个等式的证明我们留到下一节再来进行，这里我们首先需要学习一个很重要的概念叫做，希尔伯特空间 ($\mathcal{H}$:Hilbert Space)。小编之前被这个概念搞得头晕，特别还有一个叫再生核希尔伯特空间的玩意，太恶心了。

3 Hilbert Space ($\mathcal{H}$)

Hilbert Space 是一个完备的，可能是无限维的，被赋予内积运算的线性空间。下面我们对这个概念进行逐字逐句的分析。

线性空间：也就是向量空间，这个空间的元素就是向量，向量之间满足加法和乘法的封闭性，实际上也就是线性表示。空间中的任意两个向量都可以由基向量线性表示。

完备的：完备性简单的认为就是对极限的操作是封闭的。我们怎么理解呢？若有一个序列为 $\{K_n\}$，这里强调一下 Hilbert Space 是一个函数空间，空间中的元素就是函数。所以，K_n 就是一个函数。那么就会有：

$$\lim_{n \longrightarrow +\infty} K_n = K \in \mathcal{H} \tag{1}$$

所以，我们理解一下就是会和无限维这个重要的概念挂钩。我理解的主要是 Hilbert Space 在无限维满足线性关系。

内积：内积应该满足三个定义，1. 正定性；2. 对称性；3. 线性。下面我们逐个来进行解释：

1. 对称性：也就是 $f,g \in \mathcal{H}$，那么就会有 $< f,g > = < g,f >$。其中，f,g 是函数，我们可以认为 Hilbert Space 是基于函数的，向量是一个特殊的表达。其实，也就是函数可以看成一个无限维的向量。大家在这里是不是看到了无限维和完备性的引用，他们的定义之间是在相互铺垫的。

2. 正定性：也就是 $< f,f > \leq 0$，等号当且仅当 $f = 0$ 是成立。

3. 线性也就是满足：$< r_1 f_1 + r_2 f_2, g > = r_1 < f_1, g > + r_2 < f_2, g >$。

描述上述三条性质的原因是什么呢？也就是我们要证明一个空间中加入一些运算。如果，判断这个运算是不是内积运算，我们需要知道这个运算满不满足上述三个条件。

现在我们介绍了大致的基本概念了，我们回到这样一个问题，对于正定核我们为什么要有两个定义？这个思想和我们之前学到的 Kernel Trick 非常的类似了，Kernel Trick 跳过了寻找 ϕ 这个过程。因为，直接用定义不好找，

Kernel Method 03 Necessary and Sufficient Conditions

Chen Gong

22 November 2019

在上一小节中，我们描述了正定核的两个定义，并且认为这两个定义之间是相互等价的。下面我们就要证明他们之间的等价性。

1 充分性证明

大家注意到在上一节的描述中，我似乎没有谈到对称性，实际上是因为对称性的证明比较的简单。就没有做过多的解释，那么我重新描述一下我们需要证明的问题。

已知：$K(x,z) =< \phi(x), \phi(z) >$，证：Gram Matrix 是半正定的，且 $K(x,z)$ 是对称矩阵。

对称性：已知：

$$K(x,z) =< \phi(x), \phi(z) > \qquad K(z,x) =< \phi(z), \phi(x) > \tag{1}$$

又因为，内积运算具有对称性，所以可以得到：

$$< \phi(x), \phi(z) >=< \phi(z), \phi(x) > \tag{2}$$

所以，我们很容易得到：$K(x,z) = K(z,x)$，所以对称性得证。

正定性：我们想要证的是 Gram Matrix= $K[K(x_i,x_j)]_{N\times N}$ 是半正定的。那么，对一个矩阵 $A_{N\times N}$，我们如何判断这是一个半正定矩阵？大概有两种方法，1. 这个矩阵的所有特征值大于等于 0；2. 对于 $\forall \alpha \in \mathbb{R}^N$，$\alpha^T A\alpha \geq 0$。这个是充分必要条件。那么，这个问题上我们要使用的方法就是，对于 $\forall \alpha \in \mathbb{R}^N$，$\alpha^T A\alpha \geq 0$。

$$\begin{align}
\alpha^T K\alpha &= \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_N\end{bmatrix} \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & K_{22} & \cdots & K_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ K_{N1} & K_{N2} & \cdots & K_{NN}\end{bmatrix} \begin{bmatrix}\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N\end{bmatrix} \tag{3} \\
&= \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j K_{ij} \tag{4} \\
&= \sum_{i=1}^N \sum_{j=1}^N \alpha_i\alpha_j < \phi(x_i), \phi(x_j) > \tag{5}
\end{align}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \phi(x_i)^T \phi(x_j) &(6)\\
&= \sum_{i=1}^{N} \phi(x_i)^T \sum_{j=1}^{N} \phi(x_j) &(7)\\
&= \left[\sum_{i=1}^{N} \phi(x_i)\right]^T \left[\sum_{j=1}^{N} \phi(x_j)\right] &(8)\\
&= \left\|\sum_{i=1}^{N} \alpha_i \phi(x_i)\right\|^2 \geq 0 &(9)
\end{aligned}
$$

所以，我们可以得到 K 是半正定的，必要性得证。

2 必要性证明

充分性得到证明之后，必要性的证明就会变得很简单了。这个证明可以被我们描述为：

已知：Gram Matrix 是半正定的，且 $K(x,z)$ 是对称矩阵。证：存在一个映射 $\phi : \mathcal{X} \mapsto \mathbb{R}^p$，使得 $K(x,z) =< \phi(x), \phi(z) >$。

对于我们建立的一个映射 $\phi(x) = K(x,\cdot)$,我们可以得到 $K(x,\cdot)K(z,\cdot) = K(x,z)$。所以有 $K(x,z) = K(x,\cdot)K(z,\cdot) = \phi(x)\phi(z)$。我们就得证了，具体的理解可以参考我之前写的关于可再生核希尔伯特空间的理解。

另外一种证明方法：对 K 进行特征值分解，$K = V\Lambda V^T$，那么令 $\phi(x_i) = \sqrt{\lambda_i}V_i$，于是构造了 $K(x_i,x_j) = \sqrt{\lambda_i\lambda_j}V_iV_j$。