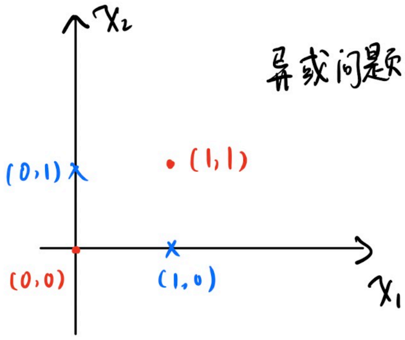
机器学习白板推导系列笔记 (7.1~7.3)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

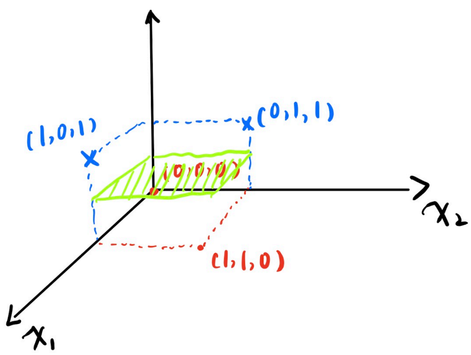
日期：2021 年 11 月 14 日

1 背景介绍

在 SVM 一章中, 说过 SVM 有三宝: 间隔, 对偶, 核技巧. 本章就是对核技巧的讲述. 我们知道, 对于线性可分的数据, 使用硬间隔 SVM 或者感知机就可以完成分类任务. 当存在一点错误时, 通过修改损失函数, 如软间隔 SVM, 就可以解决这个问题. 但是如果数据是严格非线性可分, 那么使用之前的方法不能解决这个问题. 解决这个问题的方法有许多, 核技巧是一种广泛应用的方法.



上图是一个简单的异或问题, 使用感知机 (PLA) 或非核方法 SVM 都无法解决. 二维空间的点可以被记为 $X = (x_1, x_2)$, 如果使用一种变换函数, 将其变换到三维空间, 设变换函数为 $\phi(X)$, 变换后的空间 $Z = (x_1, x_2, (x_1 - x_2)^2)$, 变换后的数据分布如下



可以发现, 通过特定的变换将数据转换成三维后, 原来线性不可分的数据转化为线性可分, 这是高维空间带来的好处, 在Cover Theorem中提出: 高纬空间比低维空间更容易线性可分.

但是高维也会带来一些问题, 如过拟合等等, 这部分在 5. 降维中讨论过. 在实际中, 面对非线性问题的解决思路主要有两种

- 由 PLA 中引入的多层感知机, 也就是神经网络, 到现在发展的深度学习
- 另一种方法是通过非线性变换 $\phi(x)$, 将非线性可分问题转化为线性可分问题, 这是本次讨论的重点

通过变换 $\phi(x)$ 将非线性问题转化为线性问题的思想称为核方法. 其主要有两个作用, 非线性带来高纬转换以及对偶表示带来乘积. 以上内容解释了非线性带来的高纬转换, 下面来解释一下第二个作用.

以 Hard-margin SVM 为例, 通过将原问题转化为对偶问题进行求解, 得到的对偶问题如下

$$\begin{cases} \max\limits_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j + \sum_{i=1}^{N}\lambda_i \\ s.t. \quad \lambda_i \geq 0, \sum_{i=1}^{N}\lambda_i y_i = 0 \end{cases} \tag{1}$$

在线性可分的情况下, 继续进行求解即可, 但在线性不可分的情况下, 需要将输入数据通过 $\phi(x)$ 进行转化后再进行求解. 一种很自然的想法是找到一个转换函数 $\phi(x)$ 然后带入进行计算. 但是, $\phi(x)$ 的寻找是很难的, 因为无法确定其维数, 或许可能就是无穷维的, 因此想要找到 $\phi(x)$ 是比较困难的. 观察式 (1) 可以看到里面存在 $\mathbf{x}_i^T\mathbf{x}_j$ 内积项, 核技巧的想法是, 我们不用去找出转换函数 $\phi(x)$, 而是直接求二者的内积 $\phi(x)^T\phi(z)$, 也称其为核函数.

下面给出核函数正式的定义, 设 $\mathcal{X}$ 是输入空间, 设 $\mathcal{H}$ 为希尔伯特空间, 如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

$$\phi(x) : \mathcal{X} \to \mathcal{H} \tag{2}$$

使得对所有的 $x, z \in \mathcal{X}$, 函数 $K(x, z)$ 满足条件

$$K(x,z) = \phi(x)^T\phi(z) =< \phi(x), \phi(z) > \tag{3}$$

则称 $K(x, z)$ 为核函数. 有了核函数, 式 (1) 在非线性可分的情况下就可以表述为

$$\begin{cases} \max\limits_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j < \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) > + \sum_{i=1}^{N}\lambda_i \\ s.t. \quad \lambda_i \geq 0, \sum_{i=1}^{N}\lambda_i y_i = 0 \end{cases} \tag{4}$$

也就是说, 在核函数给定的情况下, 可以利用求解线性分类问题的方法求解非线性问题, 并且不需要显式的学习特征空间的映射函数, 这样的技巧称为核技巧. 在实际中, 核函数的选取往往是根据领域知识直接选取, 并通过实验进行验证.

2 正定核函数

上节所说的核函数都默认是正定核函数, 正定核的定义有两个, 分别对其进行描述.

(定义 1) 存在一个映射 $\mathcal{X} \times \mathcal{X}$, 对于任意的 $x, z \in \mathcal{X}$, 如果存在 $\phi(x) \in \mathcal{H}$, 使得

$$K(x,z) =< \phi(x), \phi(z) > \tag{5}$$

那么称 $K(x,z)$ 为正定核.

(定义 2) 对于一个映射 K, 对于任意 $x,z\in\mathcal{X}$ 都有 $K(x,z)$, 且 $K(x,z)$ 满足

- 对称性
- 正定性

那么称 $K(x,z)$ 为正定核函数.

对称性指的是 $K(x,z)=K(z,x)$, 很好理解. 正定性指的是任意取 N 个元素 $x_1,x_2,\cdots,x_N$ 对应的 Gram 矩阵是半正定的, 其中 Gram 矩阵使用 $K=[K(x_i,x_j)]_{N\times N}$

为什么要引出两个关于正定核的定义, 观察定义 1 可以发现, 几乎很难根据该定义找到正定核函数, 但是可以利用定义 2 来判断一个函数是否为核函数, 显然二者是等价的.

在证明等价性之前, 先来介绍一下定义 1 中出现的$\mathcal{H}$(希尔伯特空间). 希尔伯特空间是完备的, 可能是无限维的, 被赋予内积运算的线性空间.

- 线性空间这个概念很容易理解, 在线性代数中学过, 线性空间即对加法和数乘封闭的空间.
- 完备完备可以简单的认为是对极限操作是封闭的. 该如何理解呢, 若有一个序列 K_n, 该序列的元素都是属于希尔伯特空间, 有
$$\lim_{n\to\infty}K_n=K\in\mathcal{H} \tag{6}$$
- 内积运算内积运算有需要满足三个要求, 对称性, 正定性, 线性.
  - 对称性 $<f,g>=<g,f>$
  - 正定性 $<f,f>\geq 0$, 当且仅当 $f=0$ 时, 等号成立
  - 线性 $<r_1f_i+r_2f_2,g>=r_1<f_1,g>+r_2<f_2,g>$

3 正定核充要条件证明

根据上节的定义可知

$$\text{对称}+Gram\text{矩阵半正定}\Leftrightarrow<\phi(x),\phi(z)>$$

首先证明必要性, 已知 $K(x,z)=<\phi(x),\phi(z)>$, 对于对称性, 因为是向量内积, 因此有

$$K(x,z)=<\phi(x),\phi(z)>=<\phi(z),\phi(x)> \tag{7}$$

对称性显然成立. 对于正定性, 需要证明 Gram 矩阵 $K=[K(x_i,x_j)]_{N\times N}$ 是半正定的. 在线性代数中学过, 判断矩阵 A 半正定的方法主要有: 1) 矩阵 A 的所有特征值大等于 0; 2) 对于任意向量

$\alpha \in \mathcal{R}^N$, 有 $\alpha^T A \alpha \geq 0$. 在这里使用第二种方法.

$$
\begin{aligned}
\alpha^T K \alpha &= \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \end{pmatrix} \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1N} \\ K_{21} & K_{22} & \cdots & K_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ K_{N1} & K_{N2} & \cdots & K_{NN} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} \\
&= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K_{ij} = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \phi(x_i)^T \phi(x_j) \\
&= \sum_{i=1}^N \alpha_i \phi(x_i)^T \sum_{j=1}^N \alpha_j \phi(x_j) \\
&= (\sum_{i=1}^N \alpha_i \phi(x_i))^T \sum_{j=1}^N \alpha_j \phi(x_j) \\
&= ||\sum_{i=1}^N \alpha_i \phi(x_i)||^2 \geq 0
\end{aligned}
$$

因此 Gram 矩阵是半正定的.

接下来证明充分性, 已知 $K(x,z)$ 满足对称性和正定性, 需要证明 $K(x,z) =< \phi(x), \phi(z) >$. 在证明之前, 需要一些再生希尔伯特空间 (RKHS) 的知识.

3.1 再生希尔伯特空间

假设 K(x,z) 是定义在 $\mathcal{X} \times \mathcal{X}$ 的对称函数, 对于任意的 $x_1, x_2, \cdots, x_m \in \mathcal{X}$, $K(x,z)$ 关于 $x_1, x_2, \cdots, x_m$ 的 Gram 矩阵是半正定的, 那么依据函数 $K(x,z)$, 构成一个希尔伯特空间. 主要包括三个步骤, 定义映射 ϕ 构成向量空间 $\mathcal{S}$; 在 $\mathcal{S}$ 上定义内积, 构成内积空间; 最后将 $\mathcal{S}$ 完备化构成希尔伯特空间.

3.1.1 定义映射 ϕ 构成向量空间 $\mathcal{S}$

首先定义映射 ϕ

$$\phi : x \to K(\cdot, x) \tag{8}$$

根据这个映射, 对任意的 $x_i \in \mathcal{X}, \alpha_i \in \mathbb{R}, i = 1, 2, \cdots, m$, 定义一个线性组合

$$f(\cdot) = \sum_{i=1}^m \alpha_i K(\cdot, x_i) \tag{9}$$

这些元素组成的集合称为 $\mathcal{S}$, 该集合对加法和数乘是封闭的, 因此 $\mathcal{S}$ 构成一个向量 (线性) 空间.

3.1.2 定义内积构成内积空间

在向量空间 $\mathcal{S}$ 上定义一个运算 $*$, 对任意的 $f, g \in \mathcal{S}$,

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

$$g(\cdot) = \sum_{j=1}^{l} \beta_j K(\cdot, z_j)$$

定义运算 $*$

$$f * g = \sum_{i=1}^{m} \sum_{j=1}^{l} \alpha_i \beta_j K(x_i, z_j) \tag{10}$$

为什么上面写成了 $K(x, z)$, 笔者的理解是, 根据式 (8), 定义了 x 到 $K(\cdot, x)$ 的映射 $\phi(x)$, 那么 $K(\cdot, x) = \phi(x)$, 这个如何理解, 比如函数 $y = f(x)$, 这里的 y 相当于 $K(\cdot, x)$, $f(x)$ 相当于 $\phi(x)$. 然后将 $K(\cdot, x)K(\cdot, z)$ 记为 $K(x, z)$, 因此

$$K(x, z) = \phi(x)\phi(z) \tag{11}$$

为了证明运算 $*$ 是空间 $\mathcal{S}$ 的内积, 要证

$$
\begin{aligned}
&1)(cf) * g = c(f * g), \;\; c \in \mathbb{R} \\
&2)(f + g) * h = f * h + g * h, h \in \mathcal{S} \\
&3) f * g = g * f \\
&4) f * f \geq 0 \quad f * f = 0 \Leftrightarrow f = 0
\end{aligned}
$$

对于 1)

$$
\begin{aligned}
(cf) * g &= (\sum_{i=1}^{m} c\alpha_i K(\cdot, x_i)) * \sum_{j=1}^{l} \beta_j K(\cdot, z_j) \\
&= \sum_{i=1}^{m} c\alpha_i K(\cdot, x_i) * \sum_{j=1}^{l} \beta_j K(\cdot, z_j) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{l} c\alpha_i \beta_j K(x_i, z_j) \\
&= c \sum_{i=1}^{m} \sum_{j=1}^{l} \alpha_i \beta_j K(x_i, z_j) \\
&= c(f * g)
\end{aligned}
$$

对于 2)

$$
\begin{aligned}
(f+g)*h &= (\sum_{i=1}^{m}\alpha_i K(\cdot,x_i)+\sum_{j=1}^{l}\beta_j K(\cdot,z_j))*\sum_{k=1}^{p}\eta_k K(\cdot,s_k) \\
&= \sum_{i=1}^{m+l}\alpha_i K(\cdot,x_i)*\sum_{k=1}^{p}\eta_k K(\cdot,s_k) \\
&= \sum_{i=1}^{m+l}\sum_{k=1}^{p}\alpha_i\eta_k K(x_i,s_k) \\
&= \sum_{i=1}^{m}\sum_{k=1}^{p}\alpha_i\eta_k K(x_i,s_k)+\sum_{j=1}^{l}\sum_{k=1}^{p}\beta_j\eta_k K(y_j,s_k) \\
&= f*h+g*h
\end{aligned}
$$

对于 3), 由于 $K(x,z)$ 的对称性, 显然是成立的.
对于 4)

$$f*f=\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j K(x_i,x_j) \tag{12}$$

由 Gram 矩阵的半正定性可知 $\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j K(x_i,x_j)\geq 0$, 下面证明其等号成立的条件.
充分性是显然的, 下面证明必要性, 首先证明下面不等式

$$|f*g|^2\leq (f*f)(g*g) \tag{13}$$

设 $f,g\in\mathcal{S},\lambda\in\mathbb{R}$, 那么 $f+\lambda g\in\mathcal{S}$

$$(f+\lambda g)*(f+\lambda g)\geq 0$$

$$f*f+2\lambda(f*g)+\lambda^2(g*g)\geq 0$$

上式是 λ 得一元二次方程, 其判别式恒小等于 0

$$(f*g)^2-(f*f)(g*g)\leq 0$$

因此式 (13) 成立, 下面开始正式的证明, 对任意 $x\in\mathcal{X}$

$$K(\cdot,x)*f=\sum_{i=1}^{m}\alpha_i K(x,x_i)=f(x)$$

于是

$$|f(x)|^2=|K(\cdot,x),f|^2$$

那么

$$
\begin{aligned}
|K(\cdot,x)*f|^2 &\leq (K(\cdot,x)*K(\cdot,x))(f*f) \\
&= K(x,x)(f*f)
\end{aligned}
$$

得到

$$|f(x)|^2\leq K(x,x)(f*f)$$

因此当 $f*f=0$ 时, 对于任意的 x 都有 $|f(x)|=0$

由此可以得出 $*$ 是向量空间 $\mathcal{S}$ 的内积, 使用 $\cdot$ 表示

$$f \cdot g = \sum_{i=1}^{m}\sum_{j=1}^{l}\alpha_i\beta_j K(x_i, z_j) \tag{14}$$

3.1.3 将 $\mathcal{S}$ 完备化

最后将内积空间完备化, 可以得到范数

$$||f|| = \sqrt{f \cdot f} \tag{15}$$

因此 $\mathcal{S}$ 是一个赋范向量空间, 根据泛函理论, 对于不完备的赋范向量空间一定可以使之完备化, 得到完备的赋范空间 $\mathcal{H}$, 即希尔伯特空间.

这一希尔伯特空间也称为再生希尔伯特空间, 这是由于核 K 具有再生性, 称为再生核.

$$K(\cdot, x) \cdot f = f(x)$$

$$K(\cdot, x) \cdot K(\cdot, z) = K(x, z)$$

下面就可以来证明充分性了, 即已知对称性和正定性, 可以构造从 $\mathcal{X}$ 到某个希尔伯特空间 $\mathcal{H}$ 的映射

$$\phi : x \to K(\cdot, x)$$

$$K(\cdot, x) \cdot K(\cdot, z) = K(x, z)$$

那么

$$K(x, z) = \phi(x) \cdot \phi(z)$$