核方法（1）-介绍

Y.Q. Wang

2016 年 1 月 25 日

1 什么是核方法？

数据无时无刻不在生活中产生，数据的价值在于找到数据中的模式（例如，判别、分类、排序、关联等），并将其应用到实际问题中去，指导甚至发现各类问题。但就数据本身，随着互联网化的进一步加速，数据的广度和深度都较以往发生了革命性的变化，这也使得数据所能覆盖的应用领域在这个时代中被极速拓展。这种数据形式的变化也使得我们已经很难再像过去几十年甚至百年前那样，对一个问题进行假设、建模、归纳、总结，最终找到问题所对应的模型。在这个时代，我们被迫于这个社会的发展，为了更合理的利用数据资源，使得我们常不求甚解。

感慨归感慨，如何找到一个通用模型，使其能够"标准化"地发现数据中的模式，这就是核方法发展至今的初衷。核方法的基本认识是：原始数据中的模式可以在某一维度的空间中被简单表达（见图1）。这儿所谓的"简单表达"指的是线性关系。核方法所解决的是如何将数据映射到特定的维度空间。一般核方法的处理步骤为：1) 把各种类型的数据处理为对应的 kernel matrix；2) 利用 kernel matrix 获得数据中的模式。（对应的处理框架见图2。）从另一角度看，核方法还可以理解为数学化的构建了一个更具有解释性的单层神经网络（见图3）。

过去大量的研究工作表明，核方法的思路是可行的。目前已在文本、语音、图像、检索、生物、金融、社会科学等领域有了广泛应用。

2 核方法是如何起作用的？

我们以线性回归为例具体解释核方法的作用原理。

令训练集为 $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_1)\}$，其中数据点 $\mathbf{x_i} \in R^n$，标签 $y_i \in R$。一个简单的线性回归问题是：在实数空间中找到（训练获得）一个线性函数（模型）

$$g(\mathbf{x}) = \omega^T \mathbf{x}, \tag{1}$$

能够使得 $f((\mathbf{x}, y)) = |y - g(\mathbf{x})| \to 0$。定义 $\xi = f((\mathbf{x}, y))$ 为模型对应一个数据点的训练误差。对所有的 l 个数据点和对应标签，希望能够找到这样一个模型，使得在整体训练样本上做到训练误差最

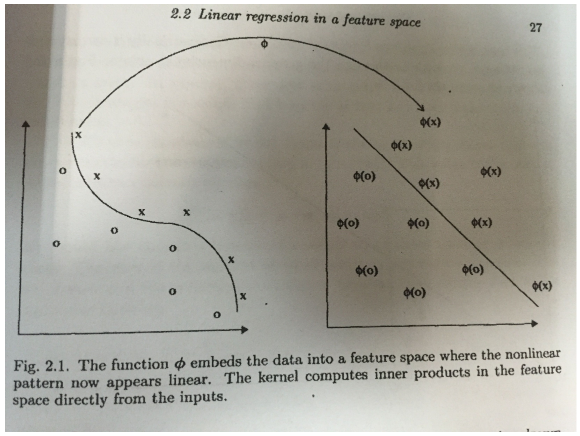

Figure 1: 非线性映射到特定维度空间后的线性关系表达（来自《Kernel methods for pattern analysis》）

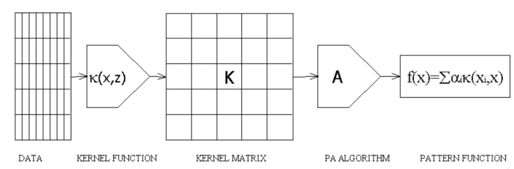

Figure 2: 核方法的一般处理框架

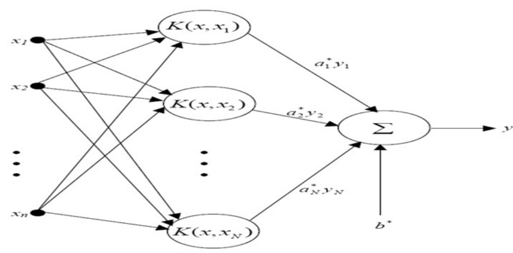

Figure 3: 核方法的网络结构

小，即

$$\mathcal{L}(g, S) = \mathcal{L}(\omega, S) = \sum_{i=1}^{l} \xi^2 = \sum_{i=1}^{l} (y_i - g(\mathbf{x}_i)^2). \tag{2}$$

为了计算方便，可以用矩阵来表示模型的在训练集上的整体误差：$\xi = \mathbf{y} - \mathbf{X}\omega$，将其带入公式 (2) 中可得：

$$\mathcal{L}(\omega, S) = \|\xi\|_2^2 = (\mathbf{y} - \mathbf{X}\omega)^T(\mathbf{y} - \mathbf{X}\omega). \tag{3}$$

为了提高模型的泛化能力，加上正则化项使其成为岭回归形式：

$$\min \mathcal{L}_\lambda(\omega, S) = \min \lambda \|\omega\|^2 + \mathcal{L}(\omega, S). \tag{4}$$

最小化公式 (4) 的解即是梯度为 0 的解，此时的最优解参数 ω 为

$$
\begin{aligned}
&\frac{\partial \mathcal{L}(\omega, S)}{\partial \omega} = 2\lambda\omega - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\omega = 0 \\
&\Rightarrow \omega = \mathbf{X}^T\alpha = \sum_{i=1}^{l} \alpha_i \mathbf{x}_i,
\end{aligned}
$$

其中 $\alpha = \lambda^{-1}(\mathbf{y} - \mathbf{X}\omega)$。可以将 α 改写为

$$
\begin{aligned}
&\alpha = \lambda^{-1}(\mathbf{y} - \mathbf{X}\omega) \\
&\Rightarrow \lambda\alpha = \mathbf{y} - \mathbf{X}\mathbf{X}^T\alpha \\
&\Rightarrow (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_l)\alpha = \mathbf{y} \\
&\Rightarrow \alpha = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_l)^{-1}\mathbf{y} \\
&\Rightarrow \alpha = (\mathbf{G} + \lambda\mathbf{I}_l)^{-1}\mathbf{y},
\end{aligned}
$$

其中 $\mathbf{I}_l$ 是大小为 $l \times l$ 的单元矩阵，$\mathbf{G} = \mathbf{X}\mathbf{X}^T$。将新得的 α 代入公式 (1) 得，

$$
g(\mathbf{x}) = \omega^T\mathbf{x} = \sum_{i=1}^{l} \alpha_i \mathbf{x}_i \mathbf{x} = \sum_{i=1}^{l} \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle = \mathbf{y}^T(\mathbf{G} + \lambda\mathbf{I}_l)^{-1}\mathbf{k}, \tag{5}
$$

其中向量 $\mathbf{k}$ 中的元素 $k_i = \langle \mathbf{x}_i, \mathbf{x} \rangle$。

可以发现，对参数 α 的求解就是对原问题（公式 (4)）求解 ω 的一种对偶求解形式。可以称变量 α 为对偶变量，矩阵 $\mathbf{G}$ 为 Gram 矩阵（后面会详细讲）。对对偶问题的求解，能够使得模型的训练复杂度从 $O(n^3)$ 变为 $O(l^3)$。当模型的复杂度远大于数据量（$n \gg l$）时，转变为对偶问题的直接好处就是可以有效降低训练的计算复杂度。

原问题的对偶形式仍然没有解决数据模式中的非线性问题。核方法考虑了一种特征映射方式将特征映射到其他对应空间（在线性回归的问题中，数据在映射后的空间里呈线性），从而解决这一问题。考虑一种映射关系

$$
\phi : \mathbf{x} \in R^n \to \phi(\mathbf{x}) \in F \subseteq R^N. \tag{6}
$$

特征映射后的原线性回归问题可以改写为

$$
\mathcal{L}(\omega, S) = |\mathbf{y} - \omega^T\phi(\mathbf{x})|. \tag{7}
$$

对应的 Gram 矩阵和向量 $\mathbf{k}$ 可以改写为

$$
\begin{aligned}
&\mathbf{G}_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle \\
&k_i = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle
\end{aligned} \tag{8}
$$

定义：核函数（kernel function）$\mathcal{K}$，对所有的 $\mathbf{x}, \mathbf{z} \in \mathbf{X}$ 满足 $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$，其中 ϕ 是从 $\mathbf{X}$ 到 $\mathbf{F}$ 的映射

$$\phi : \mathbf{x} \to \phi(\mathbf{x}) \in F.$$

核函数的选择决定了特征空间的映射关系，直接影响了原问题的求解效率。如何选择核函数，将会在接下来的内容中逐步介绍。这里举一个简单的示例演示核函数的作用。

例 1：考虑特征映射的形式为：

$$\phi : \mathbf{x} = (x_1, x_2) \to \phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in F = R^3.$$

则特征映射后的特征空间的内积形式为：

$$\begin{aligned}
\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle &= \left\langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2) \right\rangle \\
&= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (x_1z_1 + x_2z_2)^2 = \langle \mathbf{x}, \mathbf{z} \rangle^2
\end{aligned}$$

那么这种核函数的的形式即为：

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^2.$$

通过这种核函数的特征映射使得可以发现“圆”的线性关系。

当然，分类、排序等问题都可以通过类似的方法找到原问题的对偶形式，用核方法求解。具体可以参照工作 SVM 以及 SVM$^{\text{MAP}}$ 自行阅读。