SVM 的前世今生

计试 001 苏悦馨 2204120515

2022 年 11 月 17 日

SVM 是机器学习中非常经典且高效的分类模型，并且其有严格的数学理论支持。对于 SVM 的前世今生这一问题，本文就将从 SVM 的优化建模角度的若干问题来阐释。

首先从分类问题出发，所考虑的即有该分类问题是不是线性可分的，以及该分类问题是否是严格可分。相应的，这样的想法就对应了 SVM 建模中的“硬间隔线性 SVM”，“硬间隔核化 SVM”以及“软间隔核化 SVM”。接下来文章中将逐一介绍这三种 SVM 模型。

在此之前，先简要的讨论“硬间隔软间隔”和“核化”这两个概念

- 硬间隔所做的事情是将正负样本完全分开，即做到训练误差为 0；而软间隔允许有少量样本分类错误的情况出现，有利于降低过拟合的风险。
- 核化的概念涉及到模型是否是线性可分，即如果样本点在样本空间是线性可分的，则不需要使用核函数。而如果样本点在样本空间非线性可分（例如正负样本之间的间隔为曲线），则需要使用核函数，将样本点映射到更高维的空间，期望在高维空间中样本点是线性可分的。

接下来做出本文的符号定义: 二分类训练数据集表示为

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$$

其中 $x_i \in R^d$ 为特征向量，y_i 为样本的标记，正样本 $y_i = 1$，负样本 $y_i = -1$

1 硬间隔线性 SVM

硬间隔线性 SVM 是假定正负样本可在样本空间完全线性可分的，且训练的目标是将正负样本完全分隔开。

对于线性可分的样本点，超平面方程为 $w^T x + b = 0, x \in R^d$。优化的目标在于最大化正样本和负样本点集到超平面的最小距离。由于超平面的参数 b 没有任何限制，显然可以使得正样本点集到超平面最近的点位于方程 $w^T x + b = 1$，负样本 x 点集到超平面最近的点位于方程 $w^T x + b = -1$，即

$$\min\{\|x_k - x\|_2 : w^T x + b = 0, y_k = 1, k \in [1, m]\} = \frac{1}{\|w\|_2^2}$$
$$\min\{\|x_k - x\|_2 : w^T x + b = 0, y_k = -1, k \in [1, m]\} = \frac{1}{\|w\|_2^2}$$

从而正负样本到超平面的最小距离之和为 $\frac{2}{\|w\|_2^2}$，由此，硬间隔线性 SVM 的优化问题为

$$\begin{aligned} \min_{w \in R^d, b \in R} \quad & \tfrac{1}{2}\|w\|_2^2 \\ s.t. \quad & y_k(w^T x_k + b) \geq 1,\ k = 1, 2, \cdots, m \end{aligned}$$

硬间隔线性 SVM 的支持向量

硬间隔线性 SVM 的 KKT 条件为

$$\begin{cases} 1 - y_i(w^T x_i + b) \le 0 \\ \alpha_i \ge 0 \\ \alpha_i(y_i(w^T x_i + b) - 1) = 0 \end{cases}$$

当且仅当 $\alpha_i > 0$ 时，$y_i(w^T x_i + b) = 1$，该样本是离超平面最近的那个样本，此时的 x_i 被称为支持向量。由于 $w^* = \sum_{i=1}^{m} \alpha_i y_i x_i$，则最终建立的超平面只和支持向量有关。

2 硬间隔核化 SVM

所谓硬间隔核化 SVM，仍然认为样本是完全可分的，但是此时样本不一定是线性可分的。所以需要核函数将其投影到高维空间，期望在高维空间中寻找高维超平面来划分样本点集。

SVM 通过映射 $\phi : R^d \to R^{\tilde{d}}$，希望使得在新的空间 $R^{\tilde{d}}$ 中的数据集 $\tilde{D} = \{\phi(x_i), y_i\}, i = 1, 2, \cdots, m$ 是线性可分的。

通过映射将 $x \in R^d$ 映射为 $\phi(x) \in R^{\tilde{d}}$ 之后，w, b 也相应变为 $\tilde{d}$ 维。由此，可以从硬间隔线性 SVM 的问题形式，推出硬间隔核化 SVM 的表达式为

$$d \begin{array}{ll} \min\limits_{w,b} & \frac{1}{2} w^T w \\ s.t. & y_k(w^T \phi(x_k) + b) \ge 1,\ k = 1, 2, \cdots, m \end{array}$$

该问题的对偶问题可以写为如下形式

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) - \sum\limits_{i=1}^{m} \\ s.t. & \alpha_i \ge 0,\ i = 1, 2, \cdots, m, \\ & \sum\limits_{i=1}^{m} \alpha_i y_i = 0 \end{array}$$

可以看到，在其对偶问题中，出现的只有 $R^{\tilde{d}}$ 空间的内积，从而可以想到不需要将 x 映射 $\phi(x)$ 再进行计算，只需要构造出核函数 $\kappa(x_i, x_j) = \phi(x_i)^T \phi(x_j)$，将映射和计算内积这两步过程压缩为一步，使得计算复杂度由 $O(\tilde{d})$ 降为 $O(d)$。

常见的核函数有如下几种

- 线性核，此时硬间隔核化 SVM 退化为硬间隔线性 SVM

$$\kappa(x_i, x_j) = x_i^T x_j$$

- 多项式核

$$\kappa(x_i, x_j) = (\gamma x_i^T x_j + c)^k$$

- 高斯核，对应于向 R^∞ 空间的映射

$$\kappa(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

3 软间隔核化 SVM

在前面介绍的硬间隔线性 SVM 和硬间隔核化 SVM，他们所做的，都是要将正负样本完全分开，做到训练误差为 0。虽然理论上总是能找到一个映射使得数据在高维空间中线性可分 [1]，但是由于数据中噪声的存在，一昧追求将正负样本分开将增加过拟合的风险。而对于“软间隔”，即允许一定的训练样本分类出现错误，来降低过拟合风险。

现在我们希望在硬间隔核化 SVM 的基础上，允许一定错误分类样本的出现，但是又希望这类样本尽可能的少，于是将错误样本个数加入目标函数，以期出错的样本个数最少，改动后的优化问题为

$$\begin{aligned}\min_{w,b}\quad & \tfrac{1}{2}w^Tw + C\sum_{i=1}^{m} l_0(y_i \neq sign(w^T\phi(x_i)+b)) \\ s.t.\quad & y_i(w^T\phi(x_i)+b) \geq 1,\ if\ y_i = sign(w^T\phi(x_i)+b)\end{aligned}$$

其中 l_0 为指示函数，如果条件为真，则指示函数为 1；如果条件为假，则指示函数为 0。

因为 $l_0(x)$ 并不满足连续可导，于是引入松弛变量 ξ_i，用于度量样本 x_i 违背约束的程度。

$$\xi_i = \begin{cases} 0 & if\ y_i(w\phi(x_i)+b) \geq 1 \\ 1 - y_i(w\phi(x_i)+b) & else \end{cases} \tag{1}$$

从而得到软间隔核化 SVM 的形式

$$\begin{aligned}\min_{w,b,\xi}\quad & \tfrac{1}{2}w^Tw + C\sum_{i=1}^{m}\xi_i \\ s.t.\quad & y_i(w^T\phi(x_i)+b) \geq 1-\xi_i,\ i=1,2,\cdots,m \\ & \xi_i \geq 0, i=1,2,\cdots,m\end{aligned}$$

其中 C 为惩罚因子，当 C 比较大时，优化会尽量最小化 $\sum_{i=1}^{m}\xi_i$，使得正负样本之间的间隔比较小。而当 C 比较小时，优化会尽量最小化第一项，即使正负样本之间的间隔较大，而允许一些样本不满足约束。

同时为了使用核函数的技巧，我们写出软间隔核化 SVM 的对偶问题为

$$\begin{aligned}\max_{\alpha,\beta}\quad & -\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(x_i,x_j) + \sum_{i=1}^{m}\alpha_i \\ s.t.\quad & \alpha_i \geq 0,\ i=1,2,\cdots,m \\ & \beta_i \geq 0,\ i=1,2,\cdots,m \\ & \alpha_i + \beta_i = C,\ i=1,2,\cdots,m \\ & \sum_{i=1}^{m}\alpha_i y_i = 0\end{aligned}$$

参考文献

[1] [Vapnik, 2000] Vladimir Vapnik. The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science, Springer, 2000.

[2] 优化方法基础课件

[3]《机器学习》（周志华著）ISBN:978-7-302-42328-7