机器学习常见知识汇总之支持向量机

田晓彬
xiaobin9652@163.com

支持向量机（SVM）

传说天使和魔鬼玩了一个游戏，魔鬼在桌子上放了两种颜色的球，如下图所示。魔鬼让天使用一根木棍将它们分开，天使不加思索的一摆，将球分成了两部分。随后魔鬼又加入了更多的球，有的球不能再被木棍正确的分开了。

 图一

天使重新调整了木棍的摆放位置，使得两边的球都离分割他们的木棍足够远。并且在添加更多的球时，依然可以使加入的球摆放正确。天使调整后的木棒位置，魔鬼按照刚才的方式加入新球，木棍依旧可以很好的将两类球正确的分开。

接下来，魔鬼给了天使一个新的挑战，按照这种摆法，似乎没有一根木棍可以将他们完美的分开。但天使有法力，他一拍桌子，让这些球都飞到了空中，然后抓起一张纸片，插入了两类球中间，这些球被完美的分开了。

 图二

 图三

上面的例子形象地描述了SVM的分类过程，球就是**数据**，木棍和纸片就是**分类超平面**，找到最优的摆放位置就叫做**优化**，拍飞球叫做**核映射**，将数据从输入空间映射到高维特征空间。

1.支持向量机的优化目标

如上面例子中所描述的，SVM的基本思想就是找到一个划分超平面使得不同类别的样本被正确分开。但是能将训练样本划分开的超平面可能有很多，SVM尽可能的找到对未出现样本的泛化能力最强的超平面。

在样本空间中，超平面可以被表示为 $\boldsymbol{w}^T\boldsymbol{x}+b=0$,那么样本空间中任意点 $\boldsymbol{x}$ 到超平面 $(\boldsymbol{w},b)$ 的距离可以表示为

$$r=\frac{|\boldsymbol{w}^T\boldsymbol{x}+b|}{||\boldsymbol{w}||}.$$

假设超平面 $(\boldsymbol{w}, b)$ 可以将训练样本正确分类，即：

$$\begin{cases} \boldsymbol{w}^T \boldsymbol{x}_i + b \geq +1, y_i = +1; \\ \boldsymbol{w}^T \boldsymbol{x}_i + b \leq -1, y_i = -1. \end{cases} \tag{1}$$

定义支持向量为距离超平面最近数据，那么支持向量$(\boldsymbol{x}_i, y_i)$必定满足 $\boldsymbol{w}^T \boldsymbol{x}_i + b = \pm 1$。两个异类支持向量到超平面的距离之和为

$$\gamma = \frac{2}{||\boldsymbol{w}||}. \tag{2}$$

γ 被定义为间隔，如果希望找到具有最大间隔的超平面，也就是找到最大的 γ ，即如下所示：

$$\begin{aligned} &\max_{\boldsymbol{w}, b} \frac{2}{||\boldsymbol{w}||} \\ &\text{s.t.} \quad y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, i = 1, 2, ..., m. \end{aligned} \tag{3}$$

最大化式子(3)可以转化成如下的最小化问题

$$\begin{aligned} &\min_{\boldsymbol{w}, b} \frac{1}{2} ||\boldsymbol{w}||^2 \\ &\text{s.t.} y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, i = 1, 2, ..., m. \end{aligned} \tag{4}$$

这个式子就是基本的支持向量机的待优化目标。

2.优化目标的对偶问题

SVM的优化目标为通过求解公式(4)得到最优超平面的参数 $\boldsymbol{w}$ 和 b。为了更高效的求解公式(4)，我们使用拉格朗日乘子法得到其拉格朗日函数：

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} ||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{w}^T x_i + b)), \tag{5}$$

其中 $\boldsymbol{\alpha} = \{\alpha_1; \alpha_2; ...; \alpha_m\}$， α_i 为拉格朗日乘子。分别对 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 求 $\boldsymbol{w}$ 和 b 的偏导为零可得

$$\begin{aligned} \boldsymbol{w} &= \sum_{i=1}^{m} \alpha_i y_i x_i \\ 0 &= \sum_{i=1}^{m} \alpha_i y_i \end{aligned} \tag{6}$$

将式子(6)代入式子(5)中，可以将 $\boldsymbol{w}$ 和 b 消去，得到带约束的对偶问题

$$L(\boldsymbol{\alpha})=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

$$\text{s.t.}\sum_{i=1}^{m}\alpha_i y_i=0, \tag{7}$$

$$\alpha_i\geq 0, i=1,2,...,m$$

由凸优化可以知道，对偶问题是原最小化问题的下界，在求解原问题的过程中，我们肯定希望得到原问题的最优下界。最优下界其实就是对偶问题的最大值，即最大化 $L(\boldsymbol{\alpha})$。求解对偶问题得到 $\boldsymbol{\alpha}$ 后，求解 $\boldsymbol{w}$和b 即可得到最优划分超平面。

3.由优化问题的KKT条件导出支持向量

在第一节中，我们求解的优化问题为支持向量距离超平面的距离，那么为什么要定义支持向量？为什么只考虑优化支持向量与超平面的距离？

重新回到公式(4)，该优化问题的约束为不等式问题，因此我们可以得到该优化问题的KKT条件

$$\begin{cases}\alpha_i\geq 0;\\ y_i f(\boldsymbol{x}_i)-1\geq 0;\\ \alpha_i(y_i f(\boldsymbol{x}_i)-1)=0.\end{cases}$$

由公式很容易可以看出，对于任意 $(\boldsymbol{x}_i,y_i)$ 总有 $\alpha_i=0$ 或者 $y_i f(\boldsymbol{x}_i)=1$。若 $\alpha_i=0$ ，则该样本不会在优化时进行任何贡献，即样本与划分超平面无关；若 $y_i f(\boldsymbol{x}_i)=1$， 则 $\alpha_i\neq 0$，即该样本对优化过程产生了贡献，影响了 $\boldsymbol{w}$ 和 b 的值。可以看出，该样本必定在划分超平面的最大间隔边界上，即为我们在之前定义的支持向量。

从上面的推论可以看出，在SVM的求解过程中，仅仅有一部分的样本对求结果过程产生了影响。而这一部分样本都在划分超平面的最大间隔边界上，即距离划分超平面最近的样本。所以我们将这一部分样本成为支持向量，最终构建得到模型仅仅与支持向量有关。

4.对偶问题的求解

很显然，式子(7)的对偶问题最大化可以通过二次规划算法进行求解，但是因为求解的问题大小取决于数据样本数量，在实际中会有很大的时间开销。SMO算法可以将对偶问题高效的求解。

SMO的思路是是固定 α_i 之外的所有参数，然后求解出 α_i 的极值。但是由于约束 $\sum_{i=1}^{m}\alpha_i y_i=0$ 可以直接导出 α_i，所以每次选择两个变量 α_i 和 α_j，并固定其他参数。接下来，SMO算法不断的迭代优化参数 α_i 和 α_j 直至他们收敛，迭代过程如下：

1. 选取一对参数 α_i 和 α_j；

2. 固定 α_i 和 α_j 以外的参数，求解式子(7)并更新 α_i 和 α_j。

在参数选择的过程中，我们首先选择一个违背KKT条件程度最大的变量，因为违背的程度越大，参数更新后目标函数的增幅就越大。第二个参数我们选择距第一个样本最远的样本，一种解释是两个具有大间距的样本会带来较大的目标函数的更新。

使用SMO算法求解出参数 α_i 后，即可直接求解得到 $\boldsymbol{w}$。注意到对于任意的支持向量 $(\boldsymbol{x}_i, y_i)$ 都有 $y_i f(\boldsymbol{x}_i) = 1$，即

$$b = \frac{1}{y_i} - \sum_{j \in S} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i. \tag{8}$$

其中 $S = \{j | \alpha_j \geq 0, j = 1, 2, ..., m\}$为所有支持向量集合。为了使 b 的值更鲁棒，使用所有支持向量的平均值求解 b

$$b = \frac{1}{|S|} \sum_{i \in S} (\frac{1}{y_i} - \sum_{j \in S} \alpha_j y_j \boldsymbol{x}_j^T \boldsymbol{x}_i). \tag{9}$$

5.核函数与核SVM

SVM假设所有的训练数据是线性可分的，即模型可以找到一个超平面来将异类样本正确分类。然而很多情况下原始的样本空间并不存在可以将样本正确划分的超平面。如同刚开始我们叙述的天使分类球的问题，当我们讲样本投影到更高维度的特征空间时，样本有较大的概率是线性可分的。所以我们将在原始样本空间线性不可分的数据投影到高维空间，在高维空间中找到划分超平面来将样本正确分类。

令$\phi(\boldsymbol{x})$ 表示将 $\boldsymbol{x}$ 映射后的特征向量，于是，在特征空间中划分超平面所对应的SVM模型可以表示为

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b, \tag{10}$$

其中 $\boldsymbol{w}$ 和 b 是模型的参数，同式子(4)我们可以得到如下优化公式：

$$\min_{\boldsymbol{w}, b} \frac{1}{2} ||\boldsymbol{w}||^2 \tag{11}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + b) \geq 1, i = 1, 2, ..., m.$$

其对偶问题为

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \tag{12}$$

$$
\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0, \\
& \alpha_i \geq 0, i = 1, 2, ..., m.
\end{aligned}
$$

其中 $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ 为样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 映射到特征空间之后的内积。但是当特征空间的维度很高时，计算 $\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ 通常是很复杂的，所以我们进行一个假设，假设存在一个函数 κ:

$$
\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) =< \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) >= \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j). \tag{13}
$$

即 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在特征空间的内积等于他们在原始样本空间通过函数 κ 计算的结果。

由上面的假设，我们可以将式子(12)写成包含函数 κ 的形式

$$
\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{14}
$$

$$
\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_i = 0, \\
& \alpha_i \geq 0, i = 1, 2, ..., m.
\end{aligned}
$$

求解后得到

$$
\begin{aligned}
f(\boldsymbol{x}) &= \boldsymbol{w}^T \phi(\boldsymbol{x} + b) \\
&= \sum_{i=1}^{m} \alpha_i y_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) + b \\
&= \sum_{i=1}^{m} \alpha_i y_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b
\end{aligned} \tag{15}
$$

这里的函数 κ 就是核函数。

因为在实际应用中，我们希望样本可以在特征空间里面线性可分，因此特征空间的好坏对采用核函数的SVM的性能至关重要。于是采用一个合适的核函数在算法的应用中是非常必要的。常见的核函数有如下形式：

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j)^d$	$d \geq 1$为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2})$	$\sigma > 0$为高斯核的核宽

名称	表达式	参数
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma})$	$\sigma > 0$
$Sigmoid$核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(\beta \boldsymbol{x}_i^T \boldsymbol{x}_j + \theta)$	$tanh$为双曲正切函数，$\beta > 0, \theta > 0$

此外核函数还可以通过函数组合得到：

1. 若κ_1和κ_2为核函数，则对于任意正数γ_1和γ_2，其线性组合$\gamma_1 \kappa_1 + \gamma_2 \kappa_2$也是核函数；
2. 若κ_1和κ_2为核函数，则核函数的直积$\kappa_1 \bigotimes \kappa_2(\boldsymbol{x}, \boldsymbol{z}) = \kappa_1(\boldsymbol{x}, \boldsymbol{z}) \kappa_2(\boldsymbol{x}, \boldsymbol{z})$也是核函数；
3. 若κ_1为核函数，则对任意函数$g(\boldsymbol{x})$,$\kappa(\boldsymbol{x}, \boldsymbol{z}) = g(\boldsymbol{x}) \kappa_1(\boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{z})$也是核函数。

5.软间隔与正则化

在前面的介绍中，我们假设样本在样本空间或者特征空间中必定是线性可分的，即存在一个超平面可以将样本完全正确的划分开来。然而在实际数据中总存在这噪声或者离群点，很难在样本空间或者特征空间中找到合适的超平面来使的数据被完全正确分类，并且在优化的过程中，很容易出现数据的过拟合问题。为了缓解这些问题，我们在SVM中允许一部分的样本出错，即引入软间隔。

在前面的介绍中，我们要求超平面必须将所有的样本分类正确，这叫做硬间隔；软间隔就是允许一部分样本不满足约束

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1. \tag{16}$$

于是SVM的优化目标可以写成

$$\min_{\boldsymbol{w}, b} \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \iota_{0/1}(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1), \tag{17}$$

其中 $C > 0$ 是一个常数，$\iota_{0/1}$ 是 $0/1$ 损失函数

$$\iota_{0/1} = \begin{cases} 1, & \text{if } z < 0; \\ 0, & \text{otherwise.} \end{cases}$$

显然在这个优化问题中，C 越大，允许违背约束的样本数量越少，当 C 为无穷大时，迫使所有样本都满足约束，即退化到硬间隔。

但是 $\iota_{0/1}$ 非凸、非连续，数学性质不好，不易直接求解，所以通常使用其他函数进行替代。替代函数通常是凸的连续函数，并且是 $\iota_{0/1}$ 的上界：

1. $hinge$ 损失：$\iota_{hinge}(z) = max(0, 1 - z)$;
2. 指数损失：$\iota_{exp}(z) = \exp(-z)$;
3. 对率损失：$\iota_{log}(z) = \log(1 + \exp(-z))$.

引入松弛变量 $\xi \geq 0$ 代表损失函数，式子(17)被重写为

$$\min_{\boldsymbol{w},b,\xi_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{18}$$

$$\begin{aligned} \text{s.t.} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0,\ i = 1, 2, ..., m \end{aligned}$$

这就是软间隔支持向量机。其中 ξ_i 为第 i 个样本不满足约束(16)的程度。通过拉格朗日乘子法，我们可以得到

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ &= \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i, \end{aligned} \tag{19}$$

其中 $\alpha_i \geq 0, \mu_i \geq 0$。令 $L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$ 对 $\boldsymbol{w}, b, \xi_i$ 的偏导为零可得

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i,$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i,$$

$$C = \alpha_i + \mu_i.$$

同理可以得到式子(18)的对偶问题

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_i \boldsymbol{x}_i^T\boldsymbol{x}_j \tag{20}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C,\ i = 1, 2, ...m. \end{aligned}$$

将式子(20)和硬间隔下的对偶问题式子(7)对比可以看出，两者的唯一区别就是对变量 α_i 的约束不同，于是可以使用求解式子(7)的方法求解式子(20)，同理可以得到引入核函数后的软间隔对偶问题。

类似于硬间隔方法，我们可以得到软间隔方法的KKT条件

$$\begin{cases} \alpha_i \geq 0, \quad \mu_i \geq 0, \\ y_i f(\boldsymbol{x}_i) - 1 + \xi_i \geq 0, \\ \alpha_i (y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0, \xi_i \geq 0, \quad \mu_i \xi_i = 0. \end{cases} \tag{21}$$

可以看出，对于任意训练样本$(\boldsymbol{x}_i, y_i)$，总有 $\alpha_i = 0$ 或者 $y_i f(\boldsymbol{x}_i) = 1 - \xi_i$。若 $\alpha_i = 0$，则该样本不会对$f(\boldsymbol{x})$有任何影响；若 $\alpha_i > 0$，则必然有 $y_i f(\boldsymbol{x}_i) = 1 - \xi_i$，则该样本是支持向量。

由 $C = \alpha_i + \mu_i$ 可知，若 $\alpha_i < C$，则 $\mu_i > 0$，进而有 $\xi_i = 0$，则该样本恰在最大间隔边界上；若 $\alpha_i = C$，则有 $\mu_i = 0$，此时若 $\xi_i \leq 1$ 则样本落在最大间隔内部，若 $\xi_i > 1$ 则该样本被错误分类。