支持向量机（二）：算法详细解析

原创 石头 机器学习算法那些事 2018-11-20

前言

我们在工作或学习中遇到的训练数据集常分为三种，线性可分数据集，近似线性可分数据集和非线性可分数据集，对应的支持向量机算法分别为线性可分支持向量机，线性支持向量机和非线性支持向量机。这三类支持向量机构建最优模型的思想和步骤相似，因此本篇详细介绍了硬间隔支持向量机算法，其他两种算法作相关的文字说明，然后，简单介绍了支持向量的含义和分布。

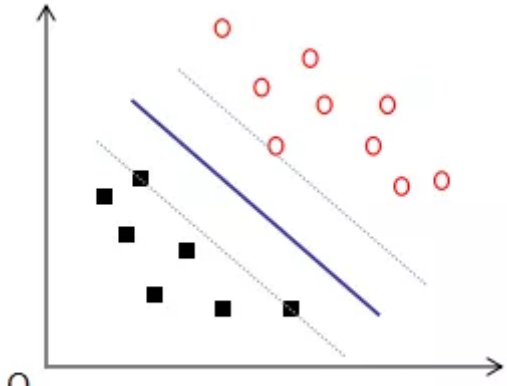
目录

1. 支持向量机种类
2. 支持向量机算法
3. 支持向量浅析
4. 总结

支持向量机种类

1. 硬间隔支持向量机

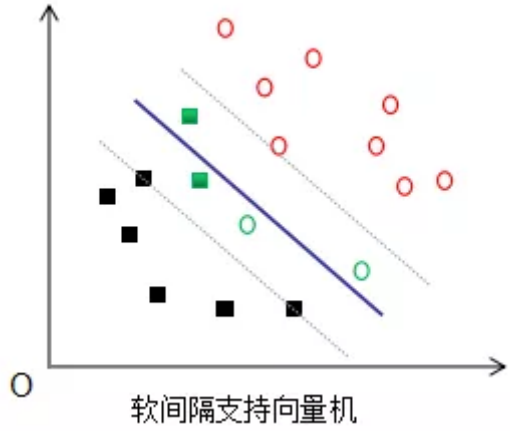
当支持向量机处理的数据集是线性可分时（如下图），称为硬间隔支持向量机。



硬间隔支持向量机

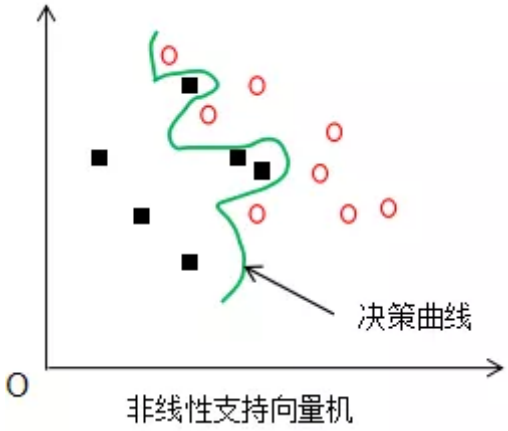
2. 软间隔支持向量机

当支持向量机处理的数据近似线性可分时，通过对每一个样本点增加松弛变量 ζ （$\zeta \geq 0$），构建学习模型，称为软间隔支持向量机。



3. 非线性支持向量机

支持向量机构建最优模型的目标函数可以表示为输入变量的内积形式，于是，很方便的通过核函数进行非线性分类。

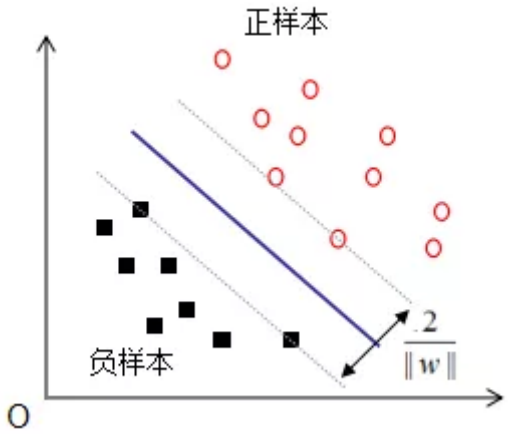


支持向量机算法

上一节把支持向量机处理的数据集分为三种，线性可分数据集，近似线性可分数据集以及非线性可分数据集，本节循序渐进的讲解这三类算法。

1. 线性可分支持向量机学习算法

对于线性可分的数据集（如下图），只要使正负样本边界函数的间隔值达到最大（硬间隔最大化），就是最优模型。



支持向量机（一）引出了线性可分数据集的最优化问题：

$$\text{目标函数：}\ \min_{w,b}\ \frac{1}{2}\|\mathrm{w}\|^2 \qquad (2.1)$$

约束条件：

$$y_i(w\cdot x_i+b)-1\geq 0 \qquad (2.2)$$

对（2.2）式进行如下变换：

$$-y_i(w\cdot x_i+b)+1\leq 0 \qquad (2.3)$$

变换的作用是使目标函数转换为凸优化问题。

首先构建拉格朗日函数，对2.3式每一个不等式引进拉格朗日乘子$\alpha_i(\alpha_i\geq 0)$得：

$$L(w,b,a)=\frac{1}{2}\|\mathrm{w}\|^2+\alpha_i\sum_{i=1}^{N}[-y_i(w\cdot x_i+b)+1]$$

$$L(w,b,a)=\frac{1}{2}\|\mathrm{w}\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i \qquad (2.4)$$

$$\because \alpha_i\geq 0$$

$$\therefore \max_{\alpha} L(w,b,a)=\frac{1}{2}\|\mathrm{w}\|^2$$

因此，目标函数等价于：

$$\min_{w,b}\max_{\alpha} L(w,b,a)$$

根据拉格朗日对偶性，原始问题的对偶问题是：

$$\max_{\alpha}\min_{w,b} L(w,b,a) \qquad (2.5)$$

(1)求$\min_{w,b} L(w,b,\alpha)$

将拉格朗日函数$L(w,b,\alpha)$分别对w,b求偏导并令其等于0，即可求得最小值。

$$\nabla_w L(w,b,\alpha) = w - \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\nabla_b L(w,b,\alpha) = \sum_{i=1}^{N} \alpha_i y_i = 0$$

得：

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (2.6)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (2.7)$$

将2.6式代入2.4式，并利用式2.7，得：

$$L(w,b,\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i y_i ((\sum_{j=1}^{N} \alpha_i y_j x_j) \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i$$

即：

$$\min_{w,b} L(w,b,\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \qquad (2.8)$$

（2）求$\min_{w,b} L(w,b,\alpha)$对α的极大，得：

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \qquad (2.9)$$

$$s.t \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

负数的最大化等价于正数的最小化，因此2.9式等价于：

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \qquad (2.10)$$

$$s.t \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad (2.11)$$

$$\alpha_i \geq 0, \quad i = 1,2,...,N \qquad (2.12)$$

由于式2.10，式2.11，式2.12都只包含变量α，因此通过*SMO*算法求解α

假设α的解为$\alpha^* = (\alpha_1^*, \alpha_2^*, ..., \alpha_N^*)$，且2.4式符合*KKT*条件，2.4式对w，b求偏导即得：

$$\nabla_w L(w,b,\alpha^*) = w^* - \sum_{i=1}^{N}\alpha_i^* y_i x_i \qquad (2.13)$$

$$\nabla_b L(w,b,\alpha^*) = -\sum_{i=1}^{N}\alpha_i^* y_i = 0$$

$$\alpha_i^*(y_i(w^* \cdot x_i + b^*) - 1) = 0, i = 1,2,...,N \qquad (2.14)$$

$$y_i(w^* \cdot x_i + b^*) - 1 \geq 0, i = 1,2,...,N$$

$$\alpha_i^* \geq 0$$

由2.13得：

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i \qquad (2.15)$$

$\because$存在分类的超平面

$\therefore w^* \neq 0$，由(2.15)可知，至少有一个$\alpha_i^* > 0$

$\because \alpha_i^* > 0$，由(2.14)可知，存在j满足：

$$y_i(w^* \cdot x_i + b^*) - 1 = 0 \qquad (2.16)$$

式2.16两边各乘以y_i，且$y_i^2 = 1$

得：

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i(x_i \cdot x_j)$$

因此，分离超平面可写成：

$$\sum_{i=1}^{N}\alpha_i^* y_i(x \cdot x_i) + b^* = 0 \qquad (2.17)$$

分类决策函数为：

$$f(x) = sign(\sum_{i=1}^{N}\alpha_i^* y_i(x \cdot x_i) + b^*) \qquad (2.18)$$

上式是硬间隔最大化对应的分类模型，软间隔最大化和非线性分类的最优模型参数求解过程与硬间隔最大化相同。

2. 线性支持向量机学习算法

若训练数据集有一些特异点（outlier），将这些特异点除去后，剩下的大部分样本点是线性可分的，

为了使特异点不影响最优模型的构建，引进一个松弛变量ζ（$\zeta \geq 0$），约束条件变为：

$$y_i(w \cdot x_i + b) \geq 1 - \zeta_i$$

对于每个松弛变量ζ_i，相应的目标函数增加一个代价ζ_i，目标函数为：

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\zeta_i$$

其中，C>0称为惩罚参数，若C值很大时，误分类产生的惩罚变大；C值很小时，误分类产生的惩罚变小。C的意义类似于正则化参数λ，C控制着模型的间隔和误分类点的重要程度，C是支持向量机比较重要的参数，实际调参中不能忽视它。

因此，近似线性不可分的线性支持向量机的学习问题：

$$\begin{aligned} \min_{w,b,\zeta} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\zeta_i \\ s.t. \quad & y_i(w \cdot x_i + b) \geq 1 - \zeta_i, i = 1,2,...N \\ & \zeta_i \geq 0 \end{aligned}$$

软间隔最大化的求解过程与上一小节一样，这里不再介绍了（若有不懂的细节请微信我）

3. 非线性支持向量机学习算法

如式2.10，线性可分支持向量机的目标函数可以写成对偶形式的表达式，因此，通过核函数进行非线性转换，实现非线性支持向量机分类算法。

支持向量浅析

本节通过公式去理解支持向量的含义和分布。

1. 线性可分支持向量机的支持向量

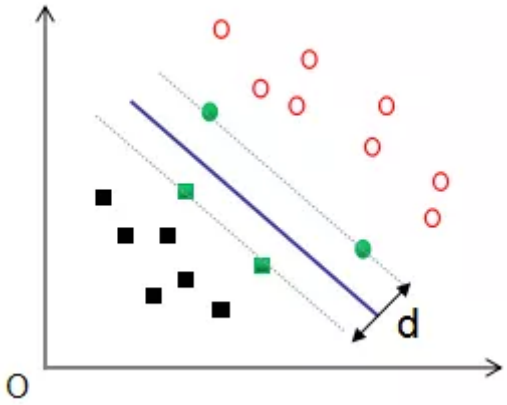
为方便阅读，重写2.14式：

$$\alpha_i^*(y_i(w^* \cdot x_i + b^*) - 1) = 0 \qquad (2.14)$$

若$\alpha_i^* > 0$，由2.14式可得：

$$y_i(w^* \cdot x_i + b^*) - 1 = 0$$

上式的含义是函数间隔等于1，对应于支持向量，如下图：



边界直线的绿色点就是支持向量，间隔最大化只与支持向量相关，因此，当忽视黑色和红色点的样本数据时，并不影响间隔d的大小，用2.14式解释黑色点和红色点不影响模型的原因：

由图可知，黑色点和红色点的函数间隔：

$$y_i(w^* \cdot x_i + b^*) - 1 > 0$$

由2.14式可得：

$$\alpha_i^* = 0$$

因此，当拉格朗日乘子为0时，由2.6式可知，对模型参数w不产生影响，因此，支持向量机的模型只受到支持向量的影响。

2. 线性支持向量机的支持向量

线性支持向量机相比于线性可分支持向量机增加了松弛变量，支持向量的含义在上一小节已经说明，本节通过公示推导线性支持向量机的支持向量的分布情况。

支持向量的表达式：

$$y(w^* \cdot x + b^* - 1 + \zeta) = 0 \qquad (3.1)$$

$$其中, 0 \le \zeta \le C$$

(1) 若 ζ_i =0，由（3.1）式可得支持向量落在间隔边界。

(2) 若0< ζ_i <1，由（3.1）式可得支持向量落在间隔边界和分离超平面之间。

(3) 若 ζ_i =1，由（3.1）式可得支持向量落在分离超平面上。

(4) 若 ζ_i >1，由（3.1）式可得支持向量落在分离超平面的误分类一侧。

总结

文章推导了支持向量机的求解算法步骤，并通过公式去理解支持向量的含义和分布。在实际项目中推荐使用含有核函数和松弛变量的支持向量机模型。下一篇讲解与算法相关的数学推导过程。

参考

李航 《统计学习方法》

推荐阅读文章

一起学习支持向量机（一）：支持向量机的分类思想

浅析感知机学习算法

-END-