12 机器学习 | 穷则变，变则通：支持向量机

2018-01-04 王天一

人工智能基础课 进入课程 >

讲述：王天一
时长 12:17 大小 5.63M

1963 年，在前苏联莫斯科控制科学学院攻读统计学博士学位的弗拉基米尔·瓦普尼克和他的同事阿列克谢·切尔沃宁基斯共同提出了支持向量机算法，随后几年两人又在此基础上进一步完善了统计学习理论。可受当时国际环境的影响，这些以俄文发表的成果并没有得到西方学术界的重视。直到 1990 年，瓦普尼克随着移民潮到达美国，统计学习理论才得到了它应有的重视，并在二十世纪末大放异彩。瓦普尼克本人也于 2014 年加入 Facebook 的人工智能实验室，并获得了包括罗森布拉特奖和冯诺伊曼奖章等诸多个人荣誉。

具体说来，**支持向量机是一种二分类算法，通过在高维空间中构造超平面实现对样本的分类。**最简单的情形是训练数据线性可分的情况，此时的支持向量机就被弱化为线性可分支持向量机，这可以视为广义支持向量机的一种特例。

线性可分的数据集可以简化为二维平面上的点集。在平面直角坐标系中，如果有若干个点全部位于 x 轴上方，另外若干个点全部位于 x 轴下方，这两个点集就共同构成了一个线性可分的训练数据集，而 x 轴就是将它们区分开来的一维超平面，也就是直线。

如果在上面的例子上做进一步的假设，假定 x 轴上方的点全部位于直线 $y=1$ 上及其上方，x 轴下方的点全部位于直线 $y=-2$ 上及其下方。如此一来，任何平行于 x 轴且在 (-2, 1) 之间的直线都可以将这个训练集分开。那么问题来了：在这么多划分超平面中，哪一个是最好的呢?

直观看来，最好的分界线应该是直线 $y=-0.5$，因为这条分界线正好位于两个边界的中间，与两个类别的间隔可以同时达到最大。当训练集中的数据因噪声干扰而移动时，这个最优划分超平面的划分精确度所受的影响最小，因而具有最强的泛化能力。

在高维的特征空间上，划分超平面可以用简单的线性方程描述

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0$$

式中的 n 维向量 $\mathbf{w}$ 为**法向量**，决定了超平面的方向；b 为**截距**，决定了超平面和高维空间中原点的距离。划分超平面将特征空间分为两个部分。位于法向量所指向一侧的数据被划分为正类，其分类标记 $y=+1$；位于另一侧的数据被划分为负类，其分类标记 $y=-1$。**线性可分支持向量机就是在给定训练数据集的条件下，根据间隔最大化学习最优的划分超平面的过程。**

给定超平面后，特征空间中的样本点 $\mathbf{x}_i$ 到超平面的距离可以表示为

$$r = \frac{\mathbf{w}^T \cdot \mathbf{x} + b}{||\mathbf{w}||}$$

显然，这个距离是个归一化的距离，因而被称为**几何间隔**。结合前文的描述，通过合理设置参数 $\mathbf{w}$ 和 b，可以使每个样本点到最优划分超平面的距离都不小于 -1，即满足以下关系

$$\mathbf{w}^T \cdot \mathbf{x}_i + b \geq 1, y_i = +1$$

$$\mathbf{w}^T \cdot \mathbf{x}_i + b \leq -1, y_i = -1$$

需要注意的是，式中的距离是非归一化的距离，被称为**函数间隔**。函数间隔和几何间隔的区别就在于未归一化和归一化的区别。

在特征空间中，距离划分超平面最近的样本点能让上式取得等号，这些样本被称为“**支持向量**”，两个异类支持向量到超平面的距离之和为 $2/||\mathbf{w}||$。因而对于线性可分支持向量机来说，其任务就是在满足上面不等式的条件下，寻找 $2/||\mathbf{w}||$ 的最大值。由于最大化 $||\mathbf{w}||^{-1}$ 等效于最小化 $||\mathbf{w}||^2$，因而上述问题可以改写为求解 $\frac{1}{2}||\mathbf{w}||^2$ 的最小值。

线性可分支持向量机是使硬间隔最大化的算法。在实际问题中，训练数据集中通常会出现噪声或异常点，导致其线性不可分。要解决这个问题就需要将学习算法一般化，得到的就是**线性支持向量机**(去掉了“可分”条件)。

线性支持向量机的通用性体现在将原始的硬间隔最大化策略转变为软间隔最大化。在线性不可分的训练集中，导致不可分的只是少量异常点，只要把这些异常点去掉，余下的大部分样本点依然满足线性可分的条件。

从数学上看，线性不可分意味着某些样本点距离划分超平面的函数间隔不满足不小于 1 的约束条件，因而需要对每个样本点引入大于零的松弛变量 $\xi \geq 0$，使得函数间隔和松弛变量的和不小于 1。这样一来，软间隔最大化下的约束条件就变成

$$\mathbf{w}^T\mathbf{x}_i + b \geq 1 - \xi_i, y_i = +1$$

$$\mathbf{w}^T\mathbf{x}_i + b \leq 1 - \xi_i, y_i = -1$$

相应地，最优化问题中的目标函数就演变为 $\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$，其中 $C > 0$ 被称为**惩罚参数**，表示对误分类的惩罚力度。这个最小化目标函数既使函数间隔尽量大，也兼顾了误分类点的个数。与要求所有样本都划分正确的硬间隔相比，软间隔允许某些样本不满足硬间隔的约束，但会限制这类特例的数目。

前文中涉及的分类问题都假定两类数据点可以用原始特征空间上的超平面区分开来，这类问题就是线性问题；如果原始空间中不存在能够正确划分的超平面，问题就演变成了非线性问题。在二维平面直角坐标系中，如果按照与原点之间的距离对数据点进行分类的话，分类的模型就不再是一条直线，而是一个圆，也就是**超曲面**。这个问题就是个非线性问题，与距离的平方形式相呼应。

不论是线性可分支持向量机还是线性支持向量机，都只能处理线性问题，对于非线性问题则无能为力。可如果能将样本从原始空间映射到更高维度的特征空间之上，在新的特征空间上样本就可能是线性可分的。如果样本的属性数有限，那么一定存在一个高维特征空间使样本可分。**将原始低维空间上的非线性问题转化为新的高维空间上的线性问题，这就是核技巧的基本思想。**

核技巧的例子可以用中国象棋来解释。开战之前，红黑两军各自陈兵，不越雷池一步，只需楚河汉界便可让不同的棋子泾渭分明。可随着车辚辚马萧萧激战渐酣，红黑棋子捉对厮杀，难分敌我，想要再造一条楚河汉界将混在一起的两军区分开已无可能。

好在我们还有高维空间。不妨把两军之争想象成一场围城大战，红帅率兵于高墙坚守，黑将携卒在低谷强攻。如此一来，不管棋盘上的棋子如何犬牙交错，只要能够在脑海中构造出一个立体城池，便可以根据位置的高低将红黑区分开来。这，就是棋盘上的核技巧。

当核技巧应用到支持向量机中时，原始空间与新空间之间的转化是通过非线性变换实现的。假设原始空间是**低维欧几里得空间**$\mathcal{X}$，新空间为**高维希尔伯特空间**$\mathcal{H}$，则从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射可以用函数 $\phi(x):\mathcal{X}\to\mathcal{H}$ 表示。核函数可以表示成映射函数内积的形式，即

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

核函数有两个特点。第一，其计算过程是在低维空间上完成的，因而避免了高维空间（可能是无穷维空间）中复杂的计算；第二，对于给定的核函数，高维空间 $\mathcal{H}$ 和映射函数 ϕ 的取法并不唯一。一方面，高维空间的取法可以不同；另一方面，即使在同一个空间上，映射函数也可以有所区别。

核函数的使用涉及一些复杂的数学问题，其结论是**一般的核函数都是正定核函数。**正定核函数的充要条件是由函数中任意数据的集合形成的核矩阵都是半正定的，这意味着任何一个核函数都隐式定义了一个成为“再生核希尔伯特空间”的特征空间，其中的数学推导在此不做赘述。**在支持向量机的应用中，核函数的选择是一个核心问题。**不好的核函数会将样本映射到不合适的特征空间，从而导致分类性能不佳。常用的核函数包括以下几种：

线性核：$K(\mathbf{X},\mathbf{Y})=\mathbf{X}^T\mathbf{Y}$

多项式核：$K(\mathbf{X},\mathbf{Y})=(\mathbf{X}^T\mathbf{Y}+c)^d$，$c$ 为常数，$d\geq 1$ 为多项式次数

高斯核：$K(\mathbf{X},\mathbf{Y})=\exp(-\frac{||\mathbf{X}-\mathbf{Y}||^2}{2\sigma^2})$，$\sigma>0$ 为高斯核的带宽

拉普拉斯核：$K(\mathbf{X},\mathbf{Y})=\exp(-\frac{||\mathbf{X}-\mathbf{Y}||}{\sigma})$，$\sigma>0$

Sigmoid 核：$K(\mathbf{X},\mathbf{Y})=\tanh(\beta\mathbf{X}^T\mathbf{Y}+\theta)$，$\beta>0,\theta<0$

核函数可以将线性支持向量机扩展为非线性支持向量机。非线性支持向量机的约束条件比较复杂，受篇幅所限，在此没有给出，你可以从相关书籍中查阅。非线性支持向量机是最普适的情形，前文中的两种情况都可以视为非线性支持向量机的特例，因而通常的支持向量机算法并不区分是否线性可分。

至此，我按照从简单到复杂的顺序，向你介绍了支持向量机三种模型的原理。在实际的应用中，对最优化目标函数的求解需要应用到最优化的理论，在这里对其思路加以简单说明。如果将支持向量机的最优化作为原始问题，应用最优化理论中的拉格朗日对偶性，可以通过求解其对偶问题得到原始问题的最优解。**三种模型的参数学习都可以转化为对偶问题的求解。**

在算法实现的过程中，支持向量机会遇到在大量训练样本下，全局最优解难以求得的尴尬。目前，高效实现支持向量机的主要算法是**SMO 算法**（Sequential Minimal Optimization，序列最小最优化）。支持向量机的学习问题可以形式化为凸二次规划问题的求解，SMO 算法的特点正是不断将原始的二次规划问题分解为只有两个变量的二次规划子问题，并求解子问题的解析解，直到所有变量满足条件为止。

支持向量机的一个重要性质是当训练完成后，最终模型只与支持向量有关，这也是“支持向量机”这个名称的来源。正如发明者瓦普尼克所言：**支持向量机这个名字强调了这类算法的关键是如何根据支持向量构建出解，算法的复杂度也主要取决于支持向量的数目。**

今天我和你分享了机器学习基本算法之一的支持向量机的基本原理，其要点如下：

线性可分支持向量机通过硬间隔最大化求出划分超平面，解决线性分类问题；

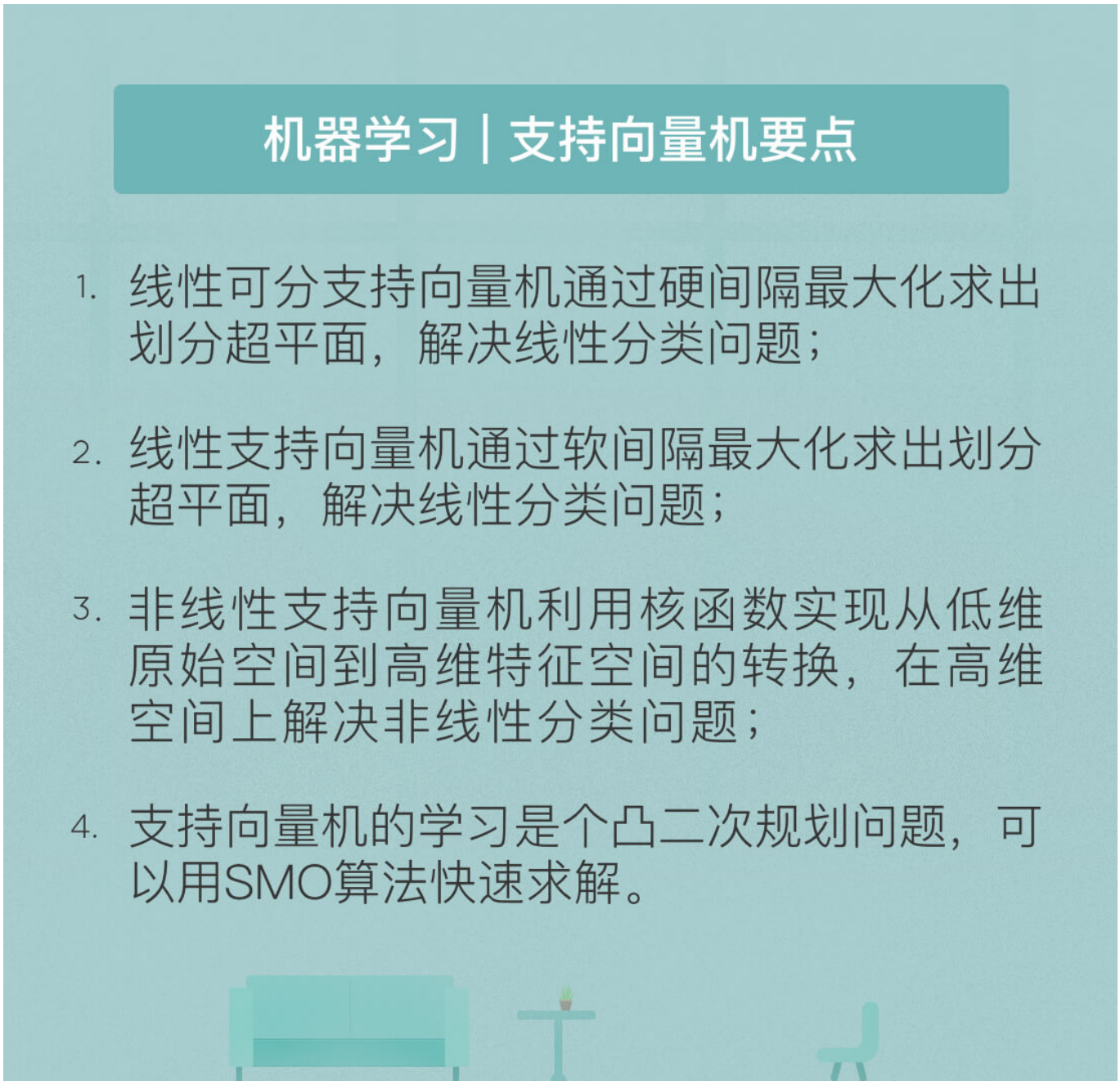
线性支持向量机通过软间隔最大化求出划分超平面，解决线性分类问题；

非线性支持向量机利用核函数实现从低维原始空间到高维特征空间的转换，在高维空间上解决非线性分类问题；

支持向量机的学习是个凸二次规划问题，可以用 SMO 算法快速求解。

支持向量机主要用于解决分类任务，那么它能否推而广之，用于解决回归任务呢？在回归任务中，支持向量又应该如何表示呢？

欢迎发表你的观点。


机器学习 | 支持向量机要点
1. 线性可分支持向量机通过硬间隔最大化求出划分超平面，解决线性分类问题；
2. 线性支持向量机通过软间隔最大化求出划分超平面，解决线性分类问题；
3. 非线性支持向量机利用核函数实现从低维原始空间到高维特征空间的转换，在高维空间上解决非线性分类问题；
4. 支持向量机的学习是个凸二次规划问题，可以用SMO算法快速求解。

上一篇　11 机器学习 | 步步为营，有章可循：决策树

下一篇　13 机器学习 | 三个臭皮匠，赛过诸葛亮：集成学习

精选留言 (8)

写留言

刘祯
2018-01-26

9

昨天还说要有抽象思维能力，今天的支持向量机就是直观的考验了。就像之前的同学说的，如果老师能够加上图其实就能够理解内涵了。

从低维到高维，这就是空间构建的方法。支持向量是最优分界线上的边缘样本，而机是机器学习的算法，全称为 Support Vectors Machines。...

展开

geoxs
2019-01-16

1

关于棋盘的例子，我觉得这样说更好一点:棋盘的棋子本来没有颜色，所以厮杀后就无法分

类了，这时候加上一个颜色维度，人类就可以看一眼就对棋子进行准确的分类

展开

lonelyand...

2018-06-06

1

软间隔最大化下的约束条件，第二个不等式，≤右侧的表达式是否应该为-1 + /xi_i?

作者回复: 留言里数学符号显示有问题，但你写的应该是对的。

wolfog

2018-02-08

1

王老师您第九段的那两个公式(WTX+B>=1).我在看其他资料的时候,假设函数间隔为1,所以就有了y(WTX+B)>=1(根据定义，函数间隔是y(WTX+B)的最小值，而y是结果标签，只能取1或者-1）所以根据不等式的运算WTX+B要么大于等于1或者小于等于-1吧

作者回复: 没错，少了个负号，谢谢你指正🙏小于-1意味着点到超平面的距离是大于1的

MJ小朋友

2018-01-06

1

老师讲的很不错，另外我又看了书上关于对偶拉格朗日的引入解参数w和b

展开

code2

2019-02-10

把公式重排一下版，显示出来的很杂乱，看不清楚。

展开

陈邓~cd

2018-06-25

“通过合理设置参数 w和 b
，可以使每个样本点到最优划分超平面的距离都不小于 -1，” 最末尾，距离是不小于1?

展开

作者回复: 是的，感谢细心指出

Colin.Tao

2018-01-19

感觉数学底子还不够，有点吃不太懂，得再学习学习...不过特别喜欢核函数这个东西，从低维升级到高维，解决线性不可分的问题。至少我先学习下这种解决问题的思路先，生活中也是，很多事情上升个维度思考，问题就很简单了。不过如果老师能再讲更细一些，比如画一些图之类的，然后简单推一下公式。如果还能加一个推导过程的视频就太好了，哈哈哈😂

展开

作者回复: 加上推导的话想要简单就不容易啦