支持向量机

支持向量机简介

什么是支持向量机？

支持向量机(Support Vector Machines, SVM) 被Vapnik与他的合作者提出于1995年,基础为统计学习理论和结构风险最小化原则.支持向量机具有完备的理论基础和出色的学习能力,是借助于最优化方法解决有限样本机器学习问题的数据挖掘出色方法之一.

支持向量机的原理是？

假设给定一个特征空间上的训练集:

$$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_i,y_i)\};x_i\in \boldsymbol{X}=\mathbb{R}^n,y_i\in \boldsymbol{Y}=\{+1,-1\};i=1,2,\dots,N$$

其中x_i为第i个特征向量,y_i为x_i的类标记,当$y_i=+1$时,x_i称为正例,当$y_i=-1$时,x_i称为负例.

线性可分SVM

假设训练数据集线性T可分,通过间隔最大化或等价求解相应凸二次规划问题得到分离超平面$w^*\cdot x+b^*=0$和相应的分类决策函数$f(x)=sign(w^*\cdot x+b^*)$为线性可分SVM.

为了量化分类的正确性和确信度,引入函数间隔的概念.

对于给定的训练数据集T和超平面(w,b),定义超平面关于样本点(x_i,y_i)的函数间隔为:

$$\hat{\gamma_i}=y_i(w\cdot x_i+b)$$

定义超平面(w,b)关于训练数据集T的函数间隔为关于T中所有样本点(x_i,y_i)的函数间隔的最小值:

$$\hat{\gamma}=\mathbf{min}_{i=1,\dots,N}\hat{\gamma_i}$$

为了取消成比例改变$w.b$导致函数间隔变化但超平面不变的问题,引入规范化$\|w\|$,此时函数间隔变为几何间隔.

对于给定的训练数据集T和超平面(w,b),定义超平面关于样本点(x_i,y_i)的几何间隔为:

$$\hat{\gamma_i}=y_i(\frac{w}{\|w\|}\cdot x_i+\frac{b}{\|w\|})$$

定义超平面(w,b)关于训练数据集T的函数间隔为关于T中所有样本点(x_i,y_i)的几何间隔的最小值:

$$\hat{\gamma}=\mathbf{min}_{i=1,\dots,N}\hat{\gamma_i}$$

可知函数间隔和几何间隔的关系为:

$$\gamma_i=\frac{\hat{\gamma_i}}{\|w\|},\gamma=\frac{\hat{\gamma}}{\|w\|}\qquad(1)$$

硬间隔最大化

通过寻找最大化几何间隔的分离超平面可以以充分大的确信度对训练数据进行分类,最大化几何间隔又称为硬间隔最大化,此时的约束最优化问题为:

$$\max_{w,b} \gamma$$

$$s.t.\ y_i(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|})s \geqslant \gamma, i = 1, 2, \ldots, N$$

即最大化超平面(w, b)关于训练数据集T的几何间隔γ

考虑(1),问题可改写为:

$$\max_{w,b} \frac{\hat{\gamma}}{\|w\|}$$

$$s.t.\ y_i(w \cdot x_i + b) \geqslant \hat{\gamma}, i = 1, 2, \ldots, N$$

基于与引入规范化$\|w\|$同样的原因,$\hat{\gamma}$的取值对结果没有影响,故取$\hat{\gamma} = 1$,考虑

$$\max_{w,b} \frac{1}{\|w\|} \Leftrightarrow \min_{w,b} \frac{1}{2}\|w\|^2$$

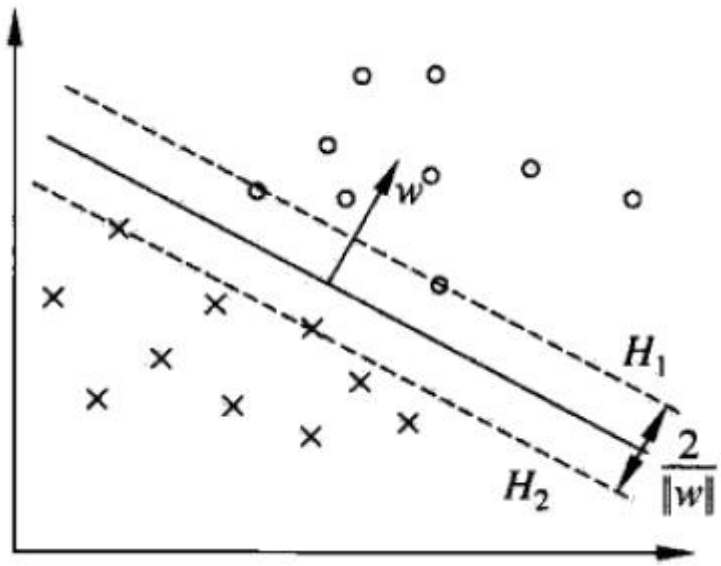
有以下线性可分SVM学习的最优化问题:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad (2)$$

$$s.t.\ y_i(w \cdot x_i + b) - 1 \geqslant 0, i = 1, 2, \ldots, N \quad (2)$$

该问题是一个凸二次规划问题.可以证明,若训练数据集T线性可分,那么可将训练数据集中的样本点完全正确分开的最大间隔超平面存在且唯一.

在线性可分的情况下,训练数据集的样本点与分离超平面距离最近的样本点的示例称为支持向量,如下图的H_1和H_2上的点:



H_1和H_2之间的距离称为间隔,为$\frac{2}{\|w\|}$,H_1和H_2称为间隔边界.

线性可分SVM对偶学习算法

通过运用拉格朗日对偶性,可以得到原始问题的对偶问题,对偶问题往往更容易求解,也便于引入核函数推广到非线性分类问题.拉格朗日对偶性在此不赘述,但有个重要定理需要特别指出:对于原始问题和对偶问题,在满足特定条件时,则$x^*.\alpha^*.\beta^*$分别为原始问题和对偶问题的解的充要条件为$x^*.\alpha^*.\beta^*$满足KKT条件.KKT条件表述如下:

$$\bigtriangledown_x L(x^*, \alpha^*, \beta^*) = 0$$

$$\alpha_i^* c_i(x^*) = 0, i = 1, 2, \ldots, k\text{(KKT的对偶互补条件)}$$

$$c_i(x^*) \leqslant 0, i = 1, 2, \ldots, k$$

$$\alpha_i^* \geqslant 0, i = 1, 2, \ldots, k$$

$$h_j(x^*) = 0, j = 1, 2, \ldots, l$$

向(2)引入拉格朗日乘子$\alpha_i \geq 0, i = 1, 2, \ldots, N$定义拉格朗日函数

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i \quad (3)$$

其中,$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$为拉格朗日乘子向量.

根据拉格朗日对偶性可以得到原始问题的对偶问题即极大极小问题:

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

先求$\min_{w,b} L(w, b, \alpha)$,分别让$(3)$对$w. b$的偏导等于0,有:

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

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

$$w = \sum_{i=1}^{N} \alpha_i x_i y_i \ (4)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad (4)$$

把(4)代入(3)整理可得:

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

接下来再求$\max_{\alpha} \min_{w,b} L(w, b, \alpha)$,其对偶问题为:

$$\max_{\alpha} -\frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i \quad (5)$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (5)$$

$$\alpha_i \geqslant 0, i = 1, 2, \ldots, N \quad (5)$$

(5)等价于下面的最优化问题:

$$\min_{\alpha} \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \ (6)$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (6)$$

$$\alpha_i \geqslant 0, i = 1, 2, \ldots, N \quad (6)$$

可以证明以下定理:设$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_i^*)^T$为对偶优化问题$(6)$的解,则$\exists j, \alpha_j^* > 0$且原始最优化问题可按下式求解$w^*. b^*$:

$$w^* = \sum_{i=1}^{N} \alpha_i^* x_i y_i$$

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

此时有分离超平面:

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

分类决策函数:

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

考虑原始最优化问题(2)和对偶最优化问题(6),将数据集中对应$\alpha_i^* > 0$的样本点(x_i, y_i)的实例$x_i \in \mathbb{R}^n$称为支持向量.可证明支持向量x_i一定在间隔边界上.

线性可分SVM学习算法

构造求解约束最优化问题

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

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

$$\alpha_i \geqslant 0, i = 1, 2, \ldots, N$$

求得最优解$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_i^*)^T$

计算$w^* = \sum_{i=1}^{N} \alpha_i^* x_i y_i$并选取$\alpha_j^* > 0$计算$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$.

求得分离超平面:$w^* \cdot x + b^* = 0$

分类决策函数:$f(x) = sign(w^* \cdot x + b^*)$

软间隔最大化线性SVM

假设训练数据集T线性不可分且存在特异点,除特异点以外的数据线性可分.

为了使得线性不可分而不满足(2)中约束条件的的训练数据集T可被训练,对每个样本点(x_i, y_i)引入松弛变量$\xi \geq 0$,并支付相应的距离代价ξ_i,使得(2)变为:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

$$s.t. y_i (w \cdot x_i + b) \geqslant 1 - \xi_i, i = 1, 2, \ldots, N$$

其中$C > 0$,称为惩罚系数,取值由实际问题决定.这个思路称为软间隔最大化.

此时线性不可分SVM的学习问题变为如下的凸二次规划原始问题:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \qquad (7)$$

$$s.t. y_i (w \cdot x_i + b) \geqslant 1 - \xi_i, i = 1, 2, \ldots, N \quad (7)$$

$$\xi_i > 0, i = 1, 2, \ldots, N \qquad (7)$$

可以证明$\exists(w,b,\xi)$为(7)的解,且w唯一,b可能不唯一且存在于一个区间中.

(7)的对偶问题是:

$$\min_\alpha \frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N\alpha_i\alpha_jy_iy_j(x_i\cdot x_j)-\sum_{i=1}^N\alpha_i \quad (8)$$

$$s.t.\sum_{i=1}^N\alpha_iy_i=0 \quad (8)$$

$$0\leq\alpha_i\leq C,i=1,2,\dots,N \quad (8)$$

线性SVM学习算法

选择惩罚参数$C>0$,构造并求解凸二次规划问题

$$\min_\alpha \frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N\alpha_i\alpha_jy_iy_j(x_i\cdot x_j)-\sum_{i=1}^N\alpha_i$$

$$s.t.\sum_{i=1}^N\alpha_iy_i=0$$

$$C\geqslant\alpha_i\geqslant 0,i=1,2,\dots,N$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\dots,\alpha_i^*)^T$

计算$w^*=\sum_{i=1}^N\alpha_i^*x_iy_i$并选取$C>\alpha_j^*>0$计算$b^*=y_j-\sum_{i=1}^N\alpha_i^*y_i(x_i\cdot x_j)$.

求得分离超平面:$w^*\cdot x+b^*=0$

分类决策函数:$f(x)=sign[\sum_{i=1}^N\alpha_i^*y_i(x\cdot x_i)+b^*]$

非线性SVM

假设训练数据集T能用$\mathbb{R}^n$中的一个超曲面将正负例分开,这种分类问题被称为非线性可分问题.

对于这种问题,需要应用SVM的核技巧如下:

通过非线性变换将输入空间(欧式空间$\mathbb{R}^n$或离散集合)对应于一个特征空间(希尔伯特空间$\mathbb{H}$)

输入空间$\mathbb{R}^n$中的超曲面模型对应于特征空间$\mathbb{H}$中的超平面模型(SVM)

通过求解特征空间$\mathbb{H}$中的线性SVM完成非线性可分分类问题的学习任务.

设$\boldsymbol{X}$为输入空间(欧式空间$\mathbb{R}^n$的子集或离散集合),又称$\mathbb{H}$为特征空间(希尔伯特空间),若$\exists\phi(\boldsymbol{x}):\boldsymbol{X}\to\mathbb{H}$使得$\forall x.z\in\boldsymbol{X},K(x,z)=\phi(x)\cdot\phi(z)$,则称$K(x,z)$为核函数,$\phi(x)$称为映射函数.$\phi(x)\cdot\phi(z)$称为$\phi(x)$和$\phi(z)$的内积.

用核函数$K(x,z)$代替对偶问题(8)中的目标函数和分类决策函数的内积,有新的目标函数:

$$W(\alpha)=\frac{1}{2}\sum_{j=1}^N\sum_{i=1}^N\alpha_i\alpha_jy_iy_jK(x_i,x_j)-\sum_{i=1}^N\alpha_i$$

和新的分类决策函数:

$$f(x)=sign[\sum_{i=1}^N\alpha_i^*y_iK(x_i,x)+b^*]$$

若使用多项式核函数$K(x,z)=(x\cdot z+1)^p$,对应的SVM为p次多项式分类器,分类决策函数成为:

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

若使用高斯核函数$K(x,z)=exp(-\frac{\|x-z\|^2}{2\sigma^2})$,对应的SVM为高斯径向基函数(RBF)分类器,分类决策函数成为:

$$f(x)=sign[\sum_{i=1}^{N}\alpha_i^*y_iexp(-\frac{\|x-z\|^2}{2\sigma^2})+b^*]$$

非线性SVM学习算法

选取适当核函数$K(x,z)$和适当的惩罚参数$C>0$,构造并求解最优化问题

$$\min_\alpha \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_jy_iy_jK(x_i,x_j)-\sum_{i=1}^{N}\alpha_i \qquad (9)$$

$$s.t.\sum_{i=1}^{N}\alpha_iy_i=0 \qquad (9)$$

$$C\geqslant\alpha_i\geqslant 0,i=1,2,\dots,N \qquad (9)$$

求得最优解$\alpha^*=(\alpha_1^*,\alpha_2^*,\dots,\alpha_i^*)^T$

并选取$C>\alpha_j^*>0$计算$b^*=y_j-\sum_{i=1}^{N}\alpha_i^*y_iK(x_i,x_j)$.

构造分类决策函数:$f(x)=sign[\sum_{i=1}^{N}\alpha_i^*y_iK(x,x_i)+b^*]$

可以证明,当$K(x,z)$为正定核函数时,(9)为凸二次规划问题并有解.

附录

B. Gram矩阵

定义:n维欧式空间任意$k(k\leq n)$个向量$\alpha_1,\alpha_2,\dots,\alpha_k$的内积组成的矩阵

$$\Delta(\alpha1,\alpha2,\dots,\alpha_k)=\begin{vmatrix}(\alpha1,\alpha1) & (\alpha1,\alpha2) & \dots & (\alpha1,\alpha k)\\(\alpha2,\alpha1) & (\alpha2,\alpha2) & \dots & (\alpha2,\alpha k)\\\dots & \dots & \dots & \dots\\(\alpha k,\alpha1) & (\alpha k,\alpha2) & \dots & (\alpha k,\alpha k)\end{vmatrix}$$

称为k个向量$\alpha_1,\alpha_2,\dots,\alpha_k$的Gram矩阵,行列式$G(\alpha_1,\alpha_2,\dots,\alpha_k)=\Delta(\alpha1,\alpha2,\dots,\alpha_k)$成为Gram行列式

参考文献

【1】李航.《统计学习方法》.清华大学出版社.2012年3月

Thanks

Chilam

Ben