支持向量机

1. SVM基本型

平面的描述方法:

$$w^T x + b = 0$$

平面的法向量:

$$w = (w_1; w_2; \cdots; w_n)$$

证明:(w, b)表示平面

$$A = (x_0, y_0, \cdots), B = (x_1, y_1, \cdots) \in (w, b)$$

$$\overrightarrow{AB} = (x_1 - x_0, y_1 - y_0, \cdots)$$

$$\begin{cases} w^T A + b = 0 \\ w^T B + b = 0 \end{cases} => w^T \overrightarrow{AB} = 0 => w \cdot \overrightarrow{AB} = 0$$

任一点Q到平面的距离, P是平面(w, b)上的点

$$r = |PQ| \cdot cos\theta \cdot \frac{|\overrightarrow{n}|}{|\overrightarrow{n}|} = \frac{\overrightarrow{PQ} \cdot \overrightarrow{n}}{|\overrightarrow{n}|} = (\overrightarrow{OQ} - \overrightarrow{OP}) \cdot \frac{\overrightarrow{n}}{|\overrightarrow{n}|} = \frac{|w^T x + b|}{||w||}$$

超平面:
对于n维空间, n-1维的平面就是"超平面"
支持向量(SV)
距离超平面最近的向量
SVM算法基本型
考虑二分类问题
优化目标:
对于超平面左右的支持向量:$x_1, x_2, \cdots$应该使:

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

$$\text{s.t.} \ \frac{y_i(w^T x_i + b)}{||w||} \geq r$$

其中y_i用于标记两个类

$$y_i = \begin{cases} +1, \ (w^T x_i + b) > 1 \\ -1, \ (w^T x_i + b) < 1 \end{cases}$$

上面的优化目标等价于:

$$\max_{w,b} \ \frac{r}{||w||}$$

$$\text{s.t.} \ y_i(w^T x_i + b) \geq r$$

容易看出:
当w, b成倍扩大时, r也会成倍扩大, $||w||$也会成倍扩大, 不会改变最大值, 所以应该固定r, 使得分母最小

所以优化目标等价于:

$$\min_{w,b} \frac{1}{2}||w||^2 = \frac{1}{2}w^T \cdot w$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \qquad \text{(SVM基本型)}$$

PS: 为什么选择$\frac{1}{2}||w||^2$? 方便计算系数

2. 由SVM基本型求参数

考虑Lagrange乘数法
将约束条件纳入到目标函数中:

$$L(w,b,\alpha) = \frac{1}{2}||w||^T + \sum \alpha_i(1 - y_i(w_i^T + b))$$

对w b求偏导

$$w = \sum \alpha_i y_i x_i$$

$$0 = \sum \alpha_i y_i$$

对偶问题即变成:

$$\max_{\alpha} \left(\sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

$$\text{s.t. } \sum \alpha_i y_i = 0, \alpha_i \geq 0$$

若分别考虑约束条件取严格等号和严格不等号的情况:
α应该满足KKT条件:

$$\begin{cases} \alpha_i \geq 0; \\ y_i(w_i^T + b) - 1 \geq 0; \\ \alpha_i(y_i(w_i^T + b) - 1) = 0 \end{cases}$$

以上问题可以通过二次规划算法解决, 但是由于算法复杂度过高, 一般选择SWO算法
基本思路:
首先固定除α_i以外的其他参数, 然后求α_i满足对偶函数最大化, 但是由于$\sum \alpha_i y_i = 0$, 这样就确定了全部的α, 所以每次选择两个α_i和α_j, 并固定其他参数, 逐级迭代.
算法如下:
每次选择两个α_i和α_j, 并固定其他参数, 反复执行一下步骤:

- 选取一对需更新的变量α_i和α_j
- 固定α_i和α_j以外的参数, 求解对偶问题的极值, 获取更新后的α_i和α_j

如何选择α_i***和***α_j***?***
α_i和α_j中至少要有一个满足KKT条件, 为了增加迭代速度, 要求另一个使得目标函数下降最快.
优化过程是怎样的?
固定α_i和α_j, 约束条件可以重写为:

$$\alpha_i y_i + \alpha_j y_j = c, \alpha_i \geq 0, \alpha_j \geq 0$$

其中:

$$c = -\sum_{k \neq i,j} \alpha_k y_k$$

并且满足约束条件消去α_j带入目标函数即可

如何确定偏移项b？

注意到支持向量(x_s, y_s)(满足$\alpha_i > 0$), 满足$y_s f(x_s) = 1$, 即(S为支持向量集):

$$\sum_{i \in S} \alpha_i y_i x_i^T x_s + b = y_s$$

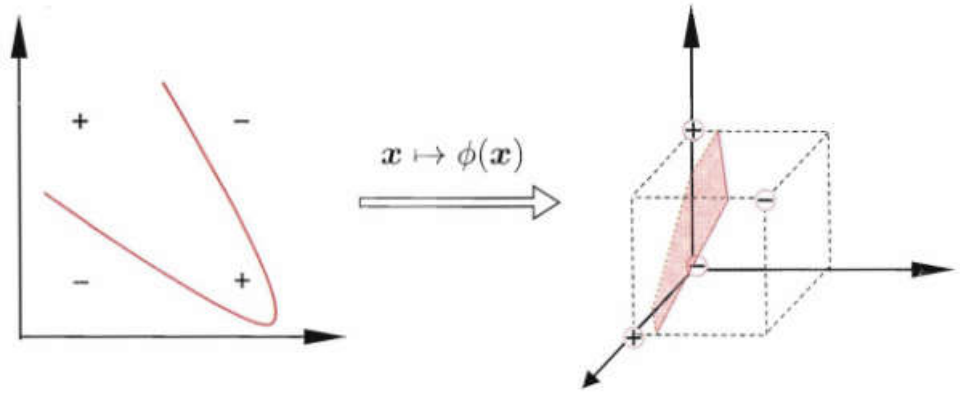
为保证鲁棒性, 对所有支持向量求平均值:

$$b = \frac{1}{|S|} \sum_{s \in S} \left(y_s - \sum_{i \in S} \alpha_i y_i x_i^T x_s \right)$$

3. 核函数

为什么引入核函数?

为了解决数据不可被超平面简单分割的问题, 将低维的数据映射到高维使之线性可分.如下图:



令$\phi(x)$为向量x的特征向量, 所以在高维空间的模型如下所示:

$$f(x) = w^T \phi(x) + b$$

其中w, b是模型参数,

$$\min_{w,b} \ \frac{1}{2} ||w||^2 = \frac{1}{2} w^T \cdot w$$

$$\text{s.t. } y_i (w^T \phi(x_i) + b) \geq 1$$

类似地得到对偶问题如下:

$$\max_{\alpha} \left(\sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \right)$$

$$\text{s.t. } \sum \alpha_i y_i = 0, \alpha_i \geq 0$$

引入核函数

这样的ϕ很难求得，引入核函数$\mathcal{K}$来表示模型:

$$\mathcal{K}(x_i, x_j) =< \phi(x_i), \phi(x_j) >= \phi(x_i)^T \phi(x_j)$$

于是优化目标变成:

$$\max_{\alpha} \left(\sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathcal{K}(x_i, x_j) \right)$$

$$\text{s.t.} \sum \alpha_i y_i = 0, \alpha_i \geq 0$$

模型函数变成:

$$f(x) = \sum \alpha_i y_i \mathcal{K}(x_i, x) + b$$

核函数是否存在?

核函数存在定理:

令$\mathcal{X}$为输入空间, $\mathcal{K}$为定义在$\mathcal{X} \times \mathcal{X}$对称核函数, 则 $\mathcal{K}$是核函数等价于对于任意数据$D = \{x_1, x_2, \cdots, x_m\}$, 核函数矩阵$\mathbf{K}$总是半正定的:

$$\mathbf{K} = \begin{bmatrix} \mathcal{K}(x_1, x_1) & \mathcal{K}(x_1, x_2) & \cdots & \mathcal{K}(x_1, x_m) \\ \mathcal{K}(x_2, x_1) & \mathcal{K}(x_2, x_2) & \cdots & \mathcal{K}(x_2, x_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(x_m, x_1) & \mathcal{K}(x_m, x_2) & \cdots & \mathcal{K}(x_m, x_m) \end{bmatrix}$$

换句话说, 对于一个半正定的核矩阵, 总能找到一个与之对应的映射ϕ

核函数隐式地定义了映射后的特征空间, 所以核函数确定了映射后特征向量的分布直接决定了支持向量机的性能.

常用核函数

名称	表达式	参数
线性核	$\kappa(\vec{x_i}, \vec{x_j}) = \vec{x_i}^T \vec{x_j}$	
多项式核	$\kappa(\vec{x_i}, \vec{x_j}) = (\vec{x_i}^T \vec{x_j})^n$	$n \geq 1$为多项式的次数
高斯核(RBF)	$\kappa(\vec{x_i}, \vec{x_j}) = exp(-\frac{\|\vec{x_i} - \vec{x_j}\|^2}{2\sigma^2})$	$\sigma > 0$为高斯核的带宽
拉普拉斯核	$\kappa(\vec{x_i}, \vec{x_j}) = exp(-\frac{\|x_i - x_j\|}{\sigma})$	σ>0
Sigmoid核	$\kappa(\vec{x_i}, \vec{x_j}) = tanh(\beta \vec{x_i}^T \vec{x_j} + \theta)$	thah为双曲正切函数

核函数的性质

- 核函数的线性组合仍然是核函数
- 核函数的乘积仍然是核函数
- $\mathcal{K}_1$是核函数, 则对于任意的函数g, $\mathcal{K}(x, z) = g(x)\mathcal{K}(x, z)g(z)$仍是核函数

4. 软间隔与正则化

为什么要引入软间隔和正则化方法?

因为实际任务中数据很难完全线性可分, 而且即使线性可分也容易造成过分学习的过拟合问题, 所以要引入软间隔和正则化方法来避免过拟合

软间隔

什么是硬间隔?

硬间隔指的是对于支持向量满足$y_i(w^T x_i + b) = 1$

所以对于全部向量要满足$y_i(w^T x_i + b) \geq 1$, 这就是硬间隔

什么是软间隔?

允许某些样本不满足$y_i(w^T x_i + b) \geq 1$

软间隔下的优化目标?

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} l_{0/1}(y_i(w_i^T + b) - 1)$$

C为加权常数, 当$C \to \inf$时, 相当于硬间隔, $l_{0/1}$定义如下:

$$l_{0/1}(z) = \begin{cases} 1, \ if \ z < 0 \\ 0, \ otherwise \end{cases}$$

但是l不连续, 性质不好, 所以常采用下面三种损失函数来代替:

$$l_{hinge} = \max(0, 1 - z)$$

$$l_{exp}(z) = \exp(-z)$$

$$l_{log}(z) = \log(1 + \exp(-z))$$

为了便于分析, 引入松弛变量ξ, 将目标写作:

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

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi \geq 0$$

由Lagrange乘数法(具体过程略)得到对偶问题:

$$L = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i + \sum \alpha_i(1 - \xi_i - y_i(w^T x_i + b)) - \sum \mu_i \xi_i$$

$$\max_{\alpha} \left(\sum \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

$$\text{s.t. } \sum \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C$$

而且要满足KKT条件:

$$\begin{cases} \alpha_i \geq 0, \mu_i \geq 0, \xi_i \geq 0 \\ y_i f(x_i) - 1 + \xi_i \geq 0 \\ \alpha_i(y_i f(x_i) - 1 + \xi_i) = 0 \\ \mu_i xi_i = 0. \end{cases}$$

对α取值讨论:

$$\begin{cases} \alpha_i = 0, \text{对模型没有影响} \\ \alpha_i \in (0, C), \mu_i > 0 (\alpha + \mu = C \text{是取极值的条件}), \xi_i = 0 \text{向量在边界上} \\ \alpha_i = C, \mu_i = 0, \xi_i \leq 1 \text{在间隔边界内部或者出现在”错误”一侧} \end{cases}$$

由以上模型可知软间隔条件下，模型仍然只与支持向量有关，叫做模型的稀疏性，采用l_{hinge}作为损失可以保持稀疏性, 采用l_{log}可以给出概率但是不能仅根据支持向量建立模型.

什么是正则化?

正则化是一种罚函数的方法, 对不希望出现的结果施以惩罚使得优化过程来接近目标.

SVM方法的一般表示:

$$\min_f \Omega(f) + C\sum l(f(x_i), y_i)$$

第一项被称为结构风险, 表示的是模型对目标的惩罚(如: 目标是模型复杂度低), 也有助于削减假设空间, 防止过拟合; 第二项被称为经验风险, 表示的是模型与数据契合程度惩罚(如: 判错率尽可能低).

第一项也叫做正则化项, C反映模型和数据的折衷, 叫做正则化常数.

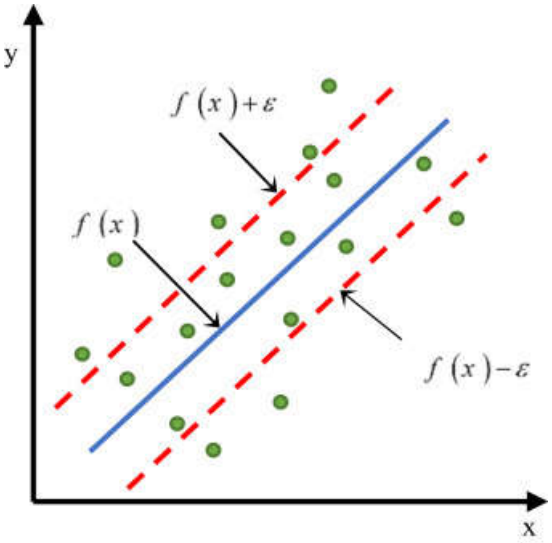
5. 支持向量回归

简称SVR

支持向量回归的基本思路?

普通的线性回归当预测值和实际值完全相同时, 损失函数为0, 而SVR**允许预测值和实际值有**$\pm\epsilon$**的偏差**来增强泛化能力, 本质上与软间隔的思想是一样的.

示意图:



数学模型

$$\min_{w,b}\ \frac{1}{2}||w||^2+C\sum_{i=1}^{m}l_\epsilon(f(x_i)-y_i)$$

C为加权常数, 当$C\to\inf$时, 相当于硬间隔, $l_{0/1}$定义如下:

$$l_\epsilon(z)=\begin{cases}0,\ if\ |z|\le\epsilon\\ |z|-\epsilon,\ otherwise\end{cases}$$

为了便于分析, 引入松弛变量$\xi,\hat{\xi}$,分别表示两侧的松弛程度, 将目标写作:

$$\min_{w,b,\xi}\ \frac{1}{2}||w||^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)$$

$$\text{s.t.}\begin{cases}f(x_i)-y_i\le\epsilon+\xi_i\\ y_i-f(x_i)\le\epsilon+\hat{\xi}_i\\ \xi_i\ge0,\hat{\xi}_i\ge0\end{cases}$$

由Lagrange乘数法

$$L=\frac{1}{2}||w||^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)+\sum\alpha_i(f(x_i)-y_i-\epsilon-\xi_i)+\sum\hat{\alpha}_i(y_i-f(x_i)-\epsilon-\hat{\xi}_i)-\sum\mu_i\xi_i-\sum\hat{\mu}_i\hat{\xi}_i$$

对w,b,ξ求偏导:

$$w=\sum(\hat{\alpha}_i-\alpha_i)x_i$$

$$0=\sum(\hat{\alpha}_i-\alpha_i)$$

$$C=\alpha_i+\mu_i=\hat{\alpha}_i+\hat{\mu}_i$$

代入L得到对偶问题:

$$\max_{\alpha}\left(\sum y_i(\hat{\alpha}_i-\alpha_i)-\epsilon(\hat{\alpha}_i+\alpha_i)-\frac{1}{2}\sum_i\sum_j(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)x_i^Tx_j\right)$$

$$\text{s.t.}\ \sum(\hat{\alpha}_i-\alpha_i)=0,\ 0\le\hat{\alpha}_i,\alpha_i\le C$$

而且要满足KKT条件:

$$\begin{cases}(C-\alpha_i)\xi_i=(C-\hat{\alpha}_i)\hat{\xi}_i=0\\ \alpha_i(f(x_i)-y_i-\epsilon-\xi_i)=\hat{\alpha}_i(y_i-f(x_i)-\epsilon-\hat{\xi}_i)=0\end{cases}$$

注意到向量不会同时出现在两侧, 所以应满足

$$\alpha_i\hat{\alpha}_i=0,\xi_i\hat{\xi}_i=0$$

最后的结果形式如下:

$$f(x)=\sum(\hat{\alpha}_i-\alpha_i)x_i^Tx+b$$

由KKT条件易知:
落在间隔带之外的点对于最后结果毫无影响因为他们对应的α和$\hat{\alpha}$均为0

如何求解 b?

当$\alpha_i\in(0,C)$时, 根据KKT条件, $\xi_i=0$, $f(x_i)-y_i-\epsilon-\xi_i=0$, 所以$b=y_i+\epsilon-\sum(\hat{\alpha}_i-\alpha_i)x_i^Tx$

PS: 若考虑核函数映射, $w=\sum(\hat{\alpha}_i-\alpha_i)\phi(x_i)$
最后的结果形式如下:

$$f(x)=\sum(\hat{\alpha}_i-\alpha_i)\mathcal{K}(x,x_i)+b$$

核方法

表示定理

对于核函数再生核希尔伯特空间$\mathbb{H}$, $||h||_{\mathbb{H}}$是$\mathbb{H}$的范数, 有核函数表示定理:
对于递增函数Ω和非负损失函数l, 优化问题

$$\min_{h\in\mathbb{H}}=\Omega(||h||_{\mathbb{H}})+l(h(x1),h(x2),\cdots,h(x_m))$$

的解总可以写成$\mathcal{K}(x,x_i)$的线性组合

核线性判别分析(KLDA)