支持向量机(上篇)

杨航锋

支持向量机，英文全名为Support Vector Machine故经常简称为SVM，和感知机算法一样它通常用于二分类样本数据的分类，不同之处在于感知机算法要求样本数据线性可分否则算法不收敛，而支持向量机算法通过引入核函数巧妙的解决了这个问题。支持向量机的基本思想是在样本数据中训练出一个间隔最大化的线性分类器，其学习策略为间隔最大化，最终转化为求解一个凸二次规划问题。[更多文章见GitHub地址]

函数间隔和几何间隔

假设输入空间是 $X \in \mathbb{R}^n$,输出空间是 $Y \in \{+1, -1\}$ ，不妨假设含有m个样本数据($x^{(1)}$,$y^{(1)}$)、($x^{(2)}$,$y^{(2)}$)、$\cdots$、($x^{(m)}$,$y^{(m)}$)，其中$x^{(i)} \in X$、$y^{(i)} \in Y$ 。在感知机算法中我们尝试寻找一个超平面 $\omega^T x + b = 0$ 从而把样本数据分隔开来，在超平面上侧定义 $y^{(i)} = +1$ 在超平面下侧定义 $y^{(i)} = -1$ ，因此样本数据集的函数间隔 $\bar{\gamma}$ 可以定义为

$$\bar{\gamma} = \min_{\omega,b} \ \hat{\gamma}^{(i)}$$
$$\hat{\gamma}^{(i)} = y^{(i)}(\omega^T x^{(i)} + b)$$

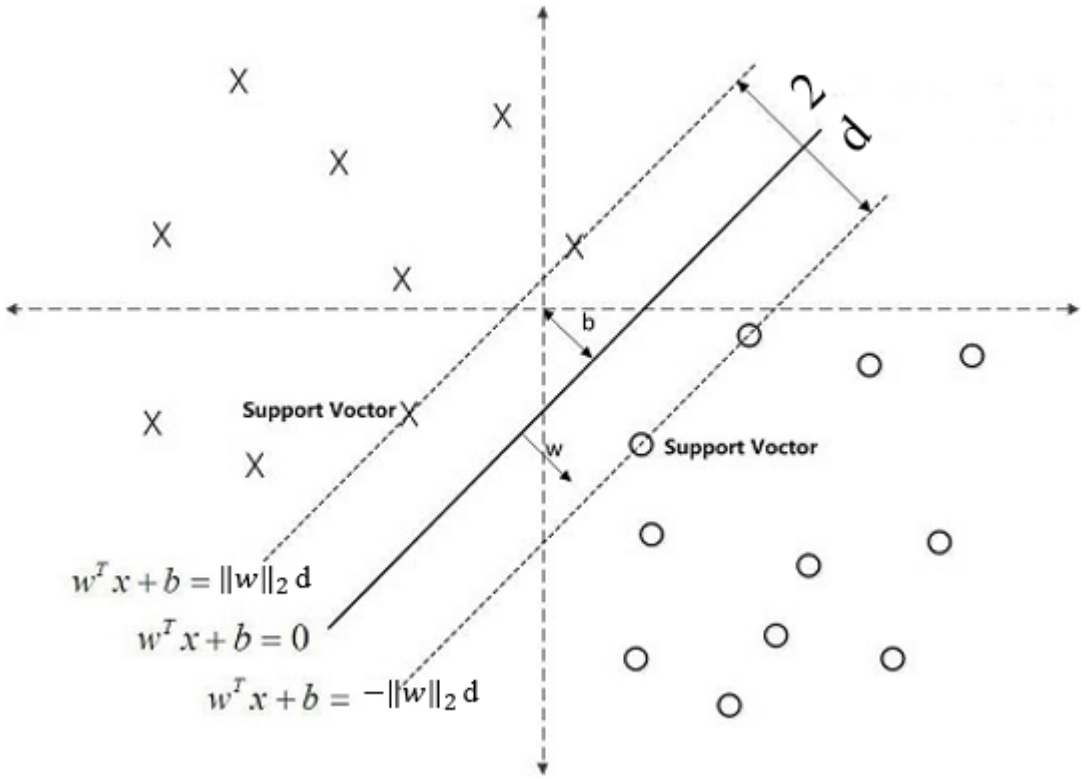
对于任意样本点 $(x^{(i)}, y^{(i)})$ 要使分类准确那么必然有 $y^{(i)}(\omega^T x^{(i)} + b) > 0$ ，所以函数间隔 $\bar{\gamma}$ 作用可以理解为使得样本数据点分类准确率最高，直观感觉就是函数间隔越大分类效果越好。但是函数间隔并不能反映点到超平面的距离，当人为成倍的改变 ω、b 的取值时，分类超平面是不变的($\omega^T x + b = 0 \iff n\omega^T x + nb = 0$)但是函数间隔会相应的增大，从而影响分类效果。为了统一度量，只需要对法向量 ω 标准化即可，这样我们就得到样本数据集的几何间隔 γ

$$\gamma = \min_{\omega,b} \ \gamma^{(i)}$$
$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{\omega}{\|\omega\|_2} \right)^T x^{(i)} + \frac{b}{\|\omega\|_2} \right)$$

支持向量与其所在的超平面方程

在感知机算法中分类超平面可能有多个，优化模型时也希望所有的数据点都远离分类超平面，但是实际上离分类超平面很远的点已经被正确分类了，再让它继续远离分类超平面将没有更大的意义，相反那些离分类超平面很近的数据点更容易被误分类，只要这些更近的数据点被正确分类那么离分类超平面更远的数据点自然也被正确分类，和原超平面平行的保持一定的几何距离的这两个超平面对应的向量，我们定义为**支持向量**。

如下图所示，分类超平面为 $\omega^T x + b = 0$ ，支持向量所在超平面与分类超平面的几何间隔为 d ，则可以求出支持向量所在的超平面方程。



对任意数据点 $(x^{(i)}, y^{(i)})$ 根据点到平面的距离关系有如下式子成立

$$\begin{cases} \dfrac{\omega^T x^{(i)} + b}{\|\omega\|_2} \geqslant d, & \forall y^{(i)} = 1 \\ \dfrac{\omega^T x^{(i)} + b}{\|\omega\|_2} \leqslant -d, & \forall y^{(i)} = -1 \end{cases}$$

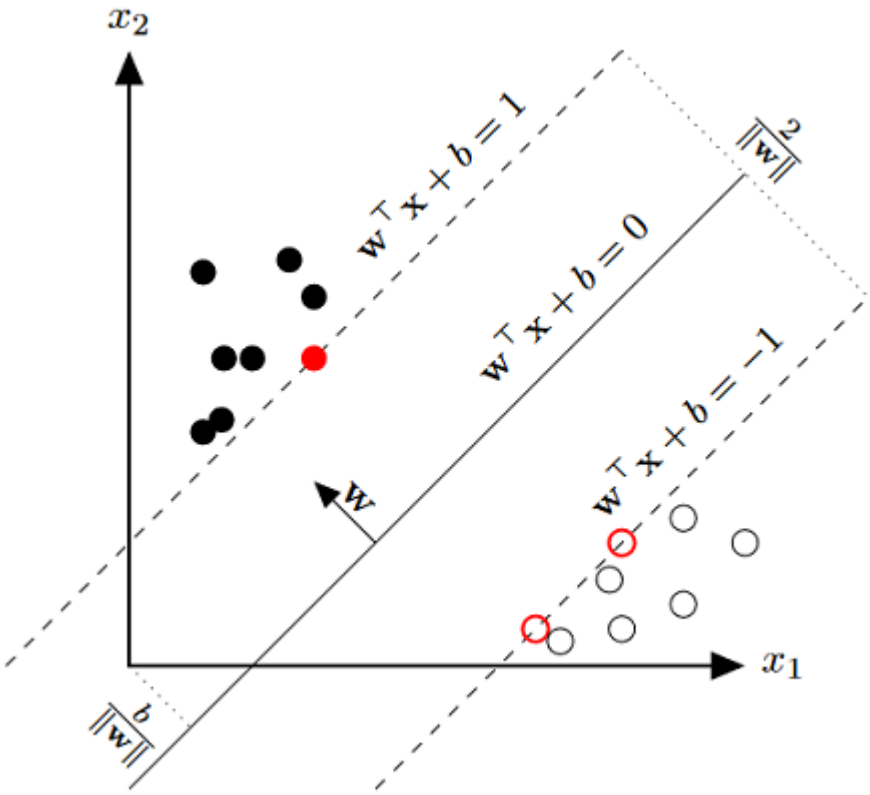
不等式的两边同时除以距离 d，并令 $\omega_d^T = \dfrac{\omega^T}{\|\omega\|_2 \cdot d}$、$b_d = \dfrac{b}{\|\omega\|_2 \cdot d}$ 可以化简为

$$\begin{cases} \dfrac{\omega^T x^{(i)} + b}{\|\omega\|_2 \cdot d} \geqslant 1, & \forall y^{(i)} = 1 \\ \dfrac{\omega^T x^{(i)} + b}{\|\omega\|_2 \cdot d} \leqslant -1, & \forall y^{(i)} = -1 \end{cases} \Rightarrow \begin{cases} \omega_d^T x^{(i)} + b_d \geqslant 1, & \forall y^{(i)} = 1 \\ \omega_d^T x^{(i)} + b_d \leqslant -1, & \forall y^{(i)} = -1 \end{cases}$$

所以支持向量所在的超平面可记为

$$\omega_d^T x + b_d = \pm 1$$

由于分类超平面方程为 $\omega^T x + b = 0$ 在等式的两端同时除以 $\|\omega\|_2 \cdot d$ 则可以写成 $\omega_d^T x + b_d = 0$，所以在同等放缩情况下分类超平面方程和支持向量所在的超平面方程一般写成下图所示，**这也解释了为什么很多博客、书籍可以直接把支持向量所在的超平面方程写成 $\omega^T x + b = \pm 1$ 。**



支持向量机的目标函数

训练支持向量机模型也就是求解下列二次规划问题，其中 $x^{(j)}$ 为支持向量

$$\max_{\omega,b} \ d = \frac{|\omega^T x^{(j)} + b|}{\|\omega\|_2}$$
$$s.t \begin{cases} \omega^T x^{(i)} + b \geqslant 1, & \forall y^{(i)} = 1 \\ \omega^T x^{(i)} + b \leqslant -1, & \forall y^{(i)} = -1 \end{cases}$$

根据支持向量所在的超平面方程可知 $|\omega^T x^{(j)} + b| = 1$ 故上述模型可以写成

$$\max_{\omega,b} \ \frac{1}{\|\omega\|_2}$$
$$s.t \quad y^{(i)}(\omega^T x^{(i)} + b) \geqslant 1$$

即

$$\min_{\omega,b} \ \frac{1}{2}\|\omega\|_2^2$$
$$s.t \quad y^{(i)}(\omega^T x^{(i)} + b) \geqslant 1$$

拉格朗日乘数与对偶问题

拉格朗日乘数法

一般的假设 $f(x), c_i(x), h_j(x)$ 是定义在 $\mathbb{R}^n$ 上的连续可微函数，考虑约束最优化问题：

$$\min_{x\in\mathbb{R}^n} f(x)$$
$$s.t.\begin{cases} c_i(x)\leqslant 0, & i=1,\cdots,k \\ h_j(x)=0, & j=1,\cdots,l \end{cases}$$

称为约束最优化问题的原始问题。

通过引入广义拉格朗日函数

$$\mathcal{L}(x,\alpha,\beta)=f(x)+\sum_{i=1}^{k}\alpha_i c_i(x)+\sum_{j=1}^{l}\beta_j h_j(x)$$
$$\text{其中}\quad \alpha_i\geqslant 0$$

现在如果把 $\mathcal{L}(x,\alpha,\beta)$ 看作是 α_i,β_j 的函数，要求其最大值，即

$$\max_{\alpha,\beta;\alpha_i\geqslant 0}\mathcal{L}(x,\alpha,\beta)$$

需要注意的是 $\mathcal{L}(x,\alpha,\beta)$ 是一个关于 α_i,β_j 的函数，优化就是确定 α_i,β_j 的值使得目标函数 $\mathcal{L}(x,\alpha,\beta)$ 取得最大值，确定了 α_i,β_j 的值，就可以得到 $\mathcal{L}(x,\alpha,\beta)$ 的最大值，因为 α_i,β_j 已经确定，显然 $\max\limits_{\alpha,\beta;\alpha_i\geqslant 0}\mathcal{L}(x,\alpha,\beta)$ 就是只关于 x 的函数，不妨定义这个函数为

$$\begin{aligned}\theta_p(x)&=\max_{\alpha,\beta;\alpha_i\geqslant 0}\mathcal{L}(x,\alpha,\beta)\\&=\max_{\alpha,\beta;\alpha_i\geqslant 0}[f(x)+\sum_{i=1}^{k}\alpha_i c_i(x)+\sum_{j=1}^{l}\beta_j h_j(x)]\end{aligned}$$

接下来通过 x 是否满足约束条件来分析这个函数，若 x 不满足原始约束条件时也就是当 $c_i(x)>0$ 时，则令 $\alpha_i\to+\infty$ 有 $\theta_p(x)=+\infty$；当 $h_j(x)\neq 0$ 时，则存在 $\beta_j h_j(x)\to+\infty$ 有 $\theta_p(x)=+\infty$。若 x 满足原始约束条件，则有 $\theta_p(x)=\max\limits_{\alpha,\beta;\alpha_i\geqslant 0}[f(x)]=f(x)$,注意上述最大化是确定 α_i,β_j 的过程，将 $f(x)$ 看成是一个常量，常量的最大值就是其本身。综上所述可以得到

$$\theta_p(x)=\begin{cases} f(x), & x\text{满足原始问题的约束条件} \\ +\infty, & \text{其他}\end{cases}$$

那么当 x 满足原始问题的约束条件下有

$$\min_x\theta_p(x)=\min_x\max_{\alpha,\beta;\alpha_i\geqslant 0}\mathcal{L}(x,\alpha,\beta)=\min_x f(x)$$

即 $\min\limits_x\theta_p(x)$ 与原始优化问题等价，所以 $\min\limits_x\theta_p(x)$ 常用来代表原始问题，下标 p 表示原始问题，定义原始问题的最优值为

$$p^*=\min_x\theta_p(x)$$

==通过拉格朗日乘数法重新定义了一个无约束问题这个无约束问题等价于原来的约束优化问题，从而将约束问题无约束化。==

对偶问题

定义关于 α,β 的函数 $\theta_D(\alpha,\beta)=\min\limits_x\mathcal{L}(x,\alpha,\beta)$ ，注意等式的右边是关于 x 的函数的最小化，确定 x 后最小值就只和 α,β 有关，所以是一个关于 α,β 的函数。因此原问题的对偶问题写作

$$\min_x\theta_p(x)=\min_x\max_{\alpha,\beta;\alpha_i\geqslant 0}\mathcal{L}(x,\alpha,\beta)\to\max_{\alpha,\beta;\alpha_i\geqslant 0}\min_x\mathcal{L}(x,\alpha,\beta)=\max_{\alpha,\beta;\alpha_i\geqslant 0}\theta_D(\alpha,\beta)$$

可以看出原始问题和它的对偶问题很对称，原始问题是先固定 $\mathcal{L}(x,\alpha,\beta)$ 中的 x ，优化出参数 α,β ，再优化出最优 x ，而对偶问题是先固定 α,β ，优化出最优 x ，然后再确定参数 α,β 。定义对偶问题的最优值为

$$d^* = \max_{\alpha,\beta;\alpha_i\geqslant 0} \theta_D(\alpha,\beta)$$

定理1：若原始问题与对偶问题都有最优值，则

$$d^* = \max_{\alpha,\beta;\alpha_i\geqslant 0} \min_x \mathcal{L}(x,\alpha,\beta) \leqslant \min_x \max_{\alpha,\beta;\alpha_i\geqslant 0} \mathcal{L}(x,\alpha,\beta) = p^*$$

也就是说原始问题的最优值不小于对偶问题的最优值，但是我们需要通过对偶问题来求解原始问题，就必须使得原始问题的最优值与对偶问题的最优值相等，于是给出下面的推论

推论：设 x^* 和 α^*,β^* 分别是原始问题和对偶问题的可行解，如果 $d^* = p^*$ ，那么 x^* 和 α^*,β^* 都是原始问题和对偶问题的最优解。

定理2：对于原始问题和对偶问题，假设函数 $f(x)$ 和 $c_i(x)$ 是凸函数， $h_j(x)$ 是仿射函数（即由一阶多项式构成的函数， $f(x) = Ax + b, A$是矩阵，x、b是向量）；并且假设不等式约束 $c_i(x)$ 是严格可行的，即存在 x ，对所有 i 有 $c_i(x) < 0$ ，则 x^* 和 α^*,β^* 分别是原始问题和对偶问题的最优解的充分必要条件是 x^* 和 α^*,β^* 满足下面的 KKT 条件：

$$\begin{aligned}
&\nabla_x \mathcal{L}(x^*,\alpha^*,\beta^*) = 0\\
&\nabla_\alpha \mathcal{L}(x^*,\alpha^*,\beta^*) = 0\\
&\nabla_\beta \mathcal{L}(x^*,\alpha^*,\beta^*) = 0\\
&\alpha_i^* c_i(x^*) = 0, i = 1,2,\cdots,k(KKT\text{对偶互补条件})\\
&c_i(x^*) \leqslant 0, i = 1,2,\cdots,k\\
&\alpha_i^* \geqslant 0, i = 1,2,\cdots,k\\
&h_j(x^*) = 0, j = 1,2,\cdots,l
\end{aligned}$$

支持向量机算法的对偶问题

回到支持向量机的优化模型

$$\begin{aligned}
&\min_{\omega,b} \ \frac{1}{2}\|\omega\|_2^2\\
&s.t. \quad y^{(i)}(\omega^T x^{(i)} + b) \geqslant 1
\end{aligned}$$

由于该模型满足 KKT 条件，因此可以通过拉格朗日乘数法将有约束的优化目标转化为无约束的优化函数

$$\begin{aligned}
&\mathcal{L}(\omega,b,\alpha) = \frac{1}{2}\|\omega\|_2^2 - \sum_{i=1}^m \alpha_i[y^{(i)}(\omega^T x^{(i)} + b) - 1]\\
&s.t. \ \ \alpha_i \geqslant 0
\end{aligned}$$

因此优化目标转化为

$$\min_{\omega,b} \max_{\alpha_i\geqslant 0} \mathcal{L}(\omega,b,\alpha)$$

该问题的对偶问题为

$$\max_{\alpha_i\geqslant 0} \min_{\omega,b} \mathcal{L}(\omega,b,\alpha)$$

从上式中可以观察到先求优化函数对于 ω,b 的极小值，然后再求拉格朗日乘子 α_i 的极大值。首先为了求 $\min\limits_{\omega,b} \mathcal{L}(\omega,b,\alpha)$,可以通过对 ω,b 分别求偏导数令其为 0 得到

$$\nabla_\omega \mathcal{L}(\omega, b, \alpha) = 0 \Rightarrow \omega = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\nabla_b \mathcal{L}(\omega, b, \alpha) = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

上面已经求出来的 ω, α 的关系，可以带入 $\mathcal{L}(\omega, b, \alpha)$ 消去 ω ，令 $\phi(\alpha) = \min\limits_{\omega, b} \mathcal{L}(\omega, b, \alpha)$ 则

$$\begin{aligned}
\phi(\alpha) &= \frac{1}{2}\|\omega\|_2^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(\omega^T x^{(i)} + b) - 1] \\
&= \frac{1}{2}\omega^T\omega - \sum_{i=1}^{m} \alpha_i y^{(i)} \omega^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= \frac{1}{2}\omega^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} \omega^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= \frac{1}{2}\omega^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \omega^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\omega^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)})^T (\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}) - b\sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - b\sum_{i=1}^{m} \alpha_i y^{(i)} + \sum_{i=1}^{m} \alpha_i \\
&= -\frac{1}{2}\sum_{i=1,j=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T \alpha_j y^{(j)} x^{(j)} + \sum_{i=1}^{m} \alpha_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}
\end{aligned}$$

对 $\phi(\alpha)$ 极大化的数学表达式如下：

$$\begin{aligned}
&\max_{\alpha} \ \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)}, x^{(j)}) \\
&s.t. \begin{cases} \sum\limits_{i=1}^{m} \alpha_i y^{(i)} = 0 \\ \alpha_i \geqslant 0, i = 1, 2, \cdots, m \end{cases}
\end{aligned}$$

等价于求解如下极小化问题

$$\begin{aligned}
&\min_{\alpha} \ \frac{1}{2}\sum_{i=1,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)}, x^{(j)}) - \sum_{i=1}^{m} \alpha_i \\
&s.t. \begin{cases} \sum\limits_{i=1}^{m} \alpha_i y^{(i)} = 0 \\ \alpha_i \geqslant 0, i = 1, 2, \cdots, m \end{cases}
\end{aligned}$$