支持向量机(SVM)

解决问题：给定训练样本 $D=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\cdots,(\vec{x}_m,y_m)\}$，$y_i\in\{-1,+1\}$，
在样本空间中找到一个划分超平面。定义该超平面为
$\vec{w}^T\vec{x}+b=0$，其中 $\vec{w}=(w_1;w_2;\ldots w_d)$ 为法向量，决定了该超平面的方向；b 为位移项，决定超平面和原点的距离。

(1) 线性~~可分~~SVM（适用线性可分训练数据）

假设超平面可将样本正确分类，我们继续定义：

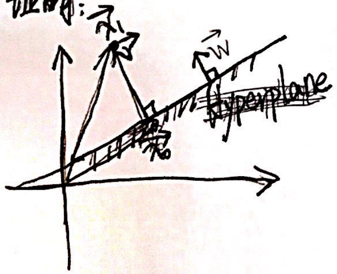
若 $y_i=1$，则有 $\vec{w}^T\vec{x}_i+b>1$；若 $y_i=-1$，则有 $\vec{w}^T x_i+b<1$.

$$\begin{cases}\vec{w}^T\vec{x}_i+b\geq 1, & y_i=1\\ \vec{w}^T\vec{x}_i+b\leq -1, & y_i=-1\end{cases}\quad(1)$$

注意：等号成立时，满足条件的样本被称为支持向量(support vector)

两个异类支持向量到超平面的距离之和为 $\frac{2}{\|\vec{w}\|}$，被称为"间隔"(Margin)
（上述定义最后影响的是 $\vec{w}$ 的值而已）

证明：



令 $\vec{x}_1$ 为支持向量，$\vec{x}_0$ 为 $\vec{x}_1$ 在超平面上的投影。
$\|\vec{w}\|$：L2范数

$\vec{x}_1=\vec{x}_0+\vec{r}$ $\quad(\vec{r}=r\cdot\frac{\vec{w}}{\|\vec{w}\|})$

$\vec{w}^T\vec{x}_0+b=0$

$\Rightarrow \vec{w}^T(\vec{x}_1-\vec{r})+b=\vec{w}^T(\vec{x}_1-r\cdot\frac{\vec{w}}{\|\vec{w}\|})+b=0$

$\Rightarrow \vec{w}^T x_1+b=\frac{\vec{w}\cdot\vec{w}^T}{\|\vec{w}\|}\cdot r=\frac{\|\vec{w}\|^2}{\|\vec{w}\|}\cdot r$

$\Rightarrow r=\frac{|\vec{w}^T x_1+b|}{\|w\|}$ （r 需为正）

$=\frac{1}{\|w\|}$

$\frac{2}{\|\vec{w}\|}$ 被称为间隔，我们的目标是在满足(1)式的情况下，
$\max\frac{2}{\|\vec{w}\|}$，提升对未知样本的泛化性能（未知样本可能在超平面附近）
所以线性可分支持向量机学习的优化问题是：

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

s.t. $y_i(\vec{w}^T\vec{x}_i+b)-1\geq 0$，$i=1,2,\cdots m$

式(2)本身是个凸二次规划问题，可用现成的优化包计算求解，

但我们有更高效的方法：通过求解对偶问题得到原始问题的最优解。

好处：(1)更容易求解；(2)自然引入核函数，进而推广到非线性SVM

方法：(1)利用拉格朗日乘子法得到原始问题的对偶问题(最大最小问题)

$$L(\vec{w},b,\vec{\alpha})=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\vec{w}^T\vec{x}_i+b)),\quad \alpha_i\geqslant 0 \qquad (3)$$

$$\text{目标：}\max_{\vec{\alpha}}\min_{\vec{w},b}L(\vec{w},b,\vec{\alpha}) \qquad (4)$$

$\min f(x)$

s.t. $c_i(x)\leq 0,\ i=1,2,\cdots k$

$h_j(x)=0,\ i=1,2,\cdots l$

⇓拉格朗日

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

其中：f,c,h为R^n上的连续可微函数

(2)先求$\min_{\vec{w},b}L(\vec{w},b,\vec{\alpha})$，令$L$对$\vec{w}$和$b$的偏导为0：

$$\vec{w}=\sum_{i=1}^{m}\alpha_i y_i\vec{x}_i$$

$$0=\sum_{i=1}^{m}\alpha_i y_i$$

代入(3)(4)，对偶问题为 $\max_{\vec{\alpha}}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\vec{x}_i^T\vec{x}_j$

$$\text{s.t.}\quad \sum_{i=1}^{m}\alpha_i y_i=0,\ \alpha_i\geqslant 0,\ i=1,2,\ldots,m$$

(3)解出$\vec{\alpha}$后，令$\vec{\alpha}^*$为对偶问题的最优解，

需满足KKT条件，w^*,b^*才是原始问题的最优解：

$$\nabla_w L(\vec{w}^*,b^*,\vec{\alpha}^*)=w^*-\sum_{i=1}^{m}\alpha_i^* y_i x_i=0 \qquad (5)$$

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

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

$$\alpha_i^*(y_i(\vec{w}^*\cdot x_i+b^*)-1)=0,\ i=1,2,\ldots m \qquad (6)$$

$\rightarrow \alpha_i^* c_i(x^*)=0$

$$y_i(\vec{w}^*\cdot\vec{x}_i+b^*)-1\geqslant 0,\ i=1,2,\ldots m$$

$c_i(x^*)\leq 0$

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

$\alpha_i^*\geqslant 0.$

由此得 $w^*=\sum_{i=1}^{n}\alpha_i^* y_i x_i$，存在$\alpha_j^*\neq 0$，否则由(5)得$w^*=0$，不符合原始问题的解。

则由(6)得 $y_j(w^*\cdot\vec{x}_j+b^*)=1$，$y_j(\sum_{i=1}^{n}\alpha_i^* y_i\vec{x}_i)\vec{x}_j+y_j b^*=1=y_j^2$

$$b^*=y_j-\sum_{i=1}^{n}\alpha_i^* y_i(\vec{x}_i\cdot\vec{x}_j)\quad \#$$

(2) 线性SVM + 软间隔最大化

前面(1)节假设训练样本在特征空间线性可分，而现实我们更多存在~~的~~的是非线性不可分数据，这就需要把硬间隔修改为软间隔。这种方法允许一些样本在分类上出错，即允许

(Hard Margin)　(Soft Margin)

某些样本不满足 $y_i(\vec{w}^T\vec{x}_i+b)\geq 1$，于是，(2)式的优化目标可以写成

(其中 ε_i 为 hinge 损失 $\max(0, 1-z)$)

$$\min_{\vec{w},b,\vec{\varepsilon}} \frac{\|\vec{w}\|^2}{2} + C\sum_{i=1}^{m}\varepsilon_i$$

$$\text{s.t.}\quad y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1-\varepsilon_i,\ i=1,2,\cdots,m$$

$$\varepsilon_i\geq 0,\ i=1,2,\cdots,m$$

(其中，ε_i 为松弛变量(slack variable)；$C>0$ 为惩罚参数，C 较大时表示更加注重分类准确性；较小更注重最小化间隔。)

(1) 对偶问题：$L(\vec{w},b,\vec{\varepsilon},\vec{\alpha},\vec{u})=\frac{1}{2}\|\vec{w}\|^2+C\sum_{i=1}^{m}\varepsilon_i-\sum_{i=1}^{m}\alpha_i(y_i(\vec{w}\cdot\vec{x}_i+b)-1+\varepsilon_i)-\sum_{i=1}^{m}u_i\varepsilon_i$

其中 $\alpha_i\geq 0$，$u_i\geq 0$ 为拉格朗日乘子。(7)

(2) 令(7)对 $\vec{w}, b$ 的偏导为0：

$$\nabla_{\vec{w}}L=\vec{w}-\sum_{i=1}^{m}\alpha_i y_i\vec{x}_i=0 \Rightarrow \vec{w}=\sum_{i=1}^{m}\alpha_i y_i\vec{x}_i$$

$$\nabla_b L=-\sum_{i=1}^{m}\alpha_i y_i=0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i=0$$

$$\nabla_{\varepsilon_i}L=C-\alpha_i-u_i=0 \Rightarrow C=\alpha_i+u_i$$

代入(7)得：

$$L(\vec{w},b,\vec{\varepsilon},\vec{\alpha},\vec{u})=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(\vec{x}_i\cdot\vec{x}_j)-\sum_{i=1}^{m}\alpha_i$$

对偶问题可转化为：$\max_{\vec{\alpha}}\ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(\vec{x}_i\cdot\vec{x}_j)$

$$\text{s.t.}\quad \sum_{i=1}^{m}\alpha_i y_i=0 \ \text{and}\ 0\leq\alpha_i\leq C,\ i=1,2,\cdots,m$$

(3) 若存在 $\vec{\alpha}^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_m^*)$ 是对偶问题的解，需满足以下KKT条件：

$\cancel{\nabla L_w L =}$

$\nabla_w L = w^* - \sum_{i=1}^{m} a_i^* y_i \vec{x_i} = 0 \Rightarrow w^* = \sum_{i=1}^{m} a_i^* y_i \vec{x_i}$

$\nabla_b L = -\sum a_i^* y_i = 0 \Rightarrow \sum a_i y_i = 0$

$\nabla_{\xi} L = C - a_i^* - u_i^* = 0 \Rightarrow C = a_i^* + u_i^*$

$a_i^* (y_i(\vec{w}^* \cdot \vec{x_i} + b^*) - 1 + \xi_i^*) = 0$ （存在y_j，使$y_j(\vec{w}^* \vec{x_j} + b^*) = 1 - \xi_j$）

进而求b^*，令$\xi_j = 0$

求超平面 $\vec{w}\vec{x} + b = 0$

其中 $b^* = y_j - \sum_{i=1}^{m} y_i a_i^* (\vec{x_i} \cdot \vec{x_j})$

$u_i \xi_i^* = 0$

$y_i(w^* \cdot \vec{x_i} + b^*) - 1 + \xi_i^* \geq 0$

$\xi_i^* \geq 0$

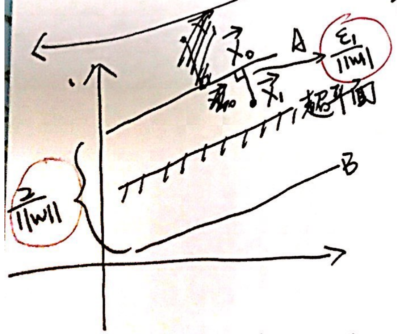
$a_i^* \geq 0$

$u_i^* \geq 0, \quad i = 1, 2, \cdots, m$

从上述：$a_i = 0$ 或 $y_i f(\vec{x_i}) = 1 - \xi_i$

该样本对f(x)无影响

软间隔支持向量



A. B: ~~最~~ 硬间隔支持向量（最大间隔）

假设软间隔的支持向量$\vec{x_i}$，在最大间隔投影为$\vec{x_0}$

参考之前：求出$\vec{x_i}$到同类的最大间隔的距离：

$|\vec{w}^T \vec{x_i} + b| = \left| \frac{\vec{w} \cdot \vec{w}^T}{\|\vec{w}\|} \cdot r + 1 \right|$

$r = \frac{|w^T \vec{x_i} + b - 1|}{\|\vec{w}\|} = \frac{|1 - \xi_i - 1|}{\|\vec{w}\|} = \boxed{\frac{\xi_i}{\|\vec{w}\|}}$

若考虑ξ_i不同取值，我们可观察到其分布情况

$a_i > 0$，$y_i f(\vec{x_i}) = 1 - \xi_i$，样本$x_i$为支持向量

(1) $a_i < C$，$u_i > 0$，$\xi_i = 0$，~~样~~ 支持向量在最大间隔边界上

(2) $a_i = C$，$u_i = 0$，$0 < \xi_i < 1$，支持向量在最大间隔和超平面之间

$\xi_i = 1$，支持向量在超平面上

$\xi_i > 1$，支持向量在超平面误分另一侧。