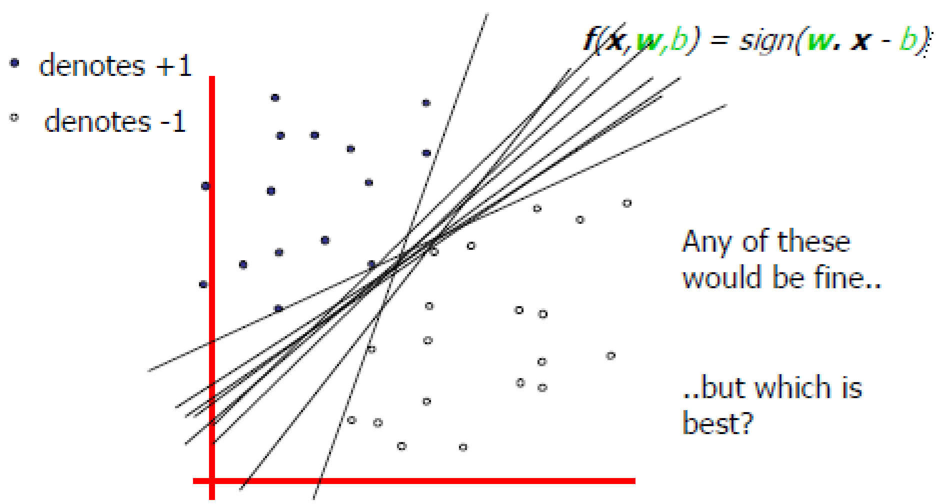
支持向量机

是一个非常常用的算法，在各个领域都有涉及

- > 支持向量的理解
  > 支持向量指的就是超平面距离最近的点；每个点都是一样向量；
  > 以前错误的理解向量一定是要两个点连接起来；此处就认为是将最近的两个点连起来就是支持向量；
  > 本质上来说，空间中的任何一点都是一个向量；只不过是以原来为起点；

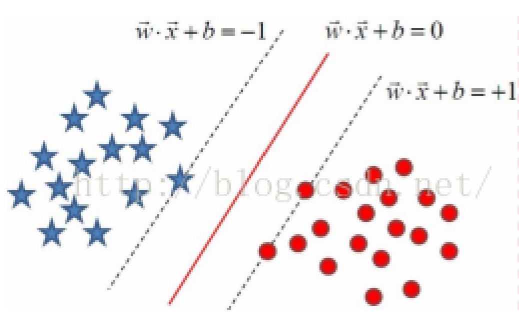
一、线性分类的SVM

1.1 线性分类



1.2 函数间隔和几何间隔

函数间隔



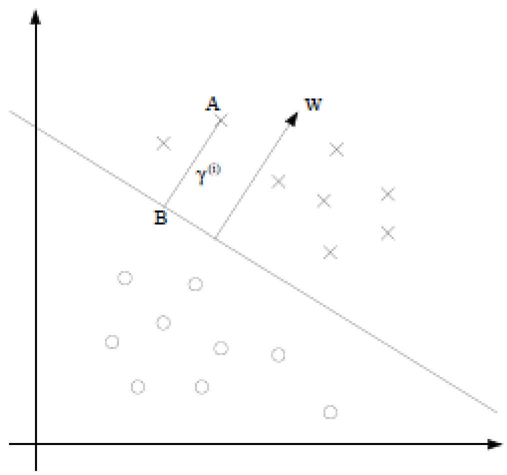
$$\hat{\gamma}^{(i)} = y^{(i)}\left(w^T x^{(i)} + b\right)$$

求最小的间隔

$$\hat{\gamma} = min_{i=1,2,\ldots m}\ \hat{\gamma}^{(i)}$$

求解需要最大化这个最小化间隔

几何间隔



如图所示，样本点A的坐标为 x (i)， A 到超平面的垂直距离为 γ(i)， 超平面的单位法向量是 ω / || ω ||， B点是超平面上一点且是A点的投影，则 B 点的坐标 x(i) - γ(i) ω / || ω ||， 将其带入超平面方程 f (x) = ωT x + b =0 .

$$w^T(x^{(i)} - \gamma^{(i)} \frac{w}{||w||}) + b = 0\text{；解得：}\gamma^{(i)} = \frac{(w^T x^{(i)} + b)}{||w||}$$

如果分类正确，则 y(i) 与 f(x(i)) 符号相同，几何间隔是：

$$\gamma^{(i)} = y^{(i)} \left(\frac{w^T x^{(i)} + b}{||w||} \right)$$

超平面与样本集的几何间隔是：

$$\gamma = min_{i=1,2,\ldots m}\ \gamma^{(i)}$$

如上所述，函数间隔与几何间隔的关系有：

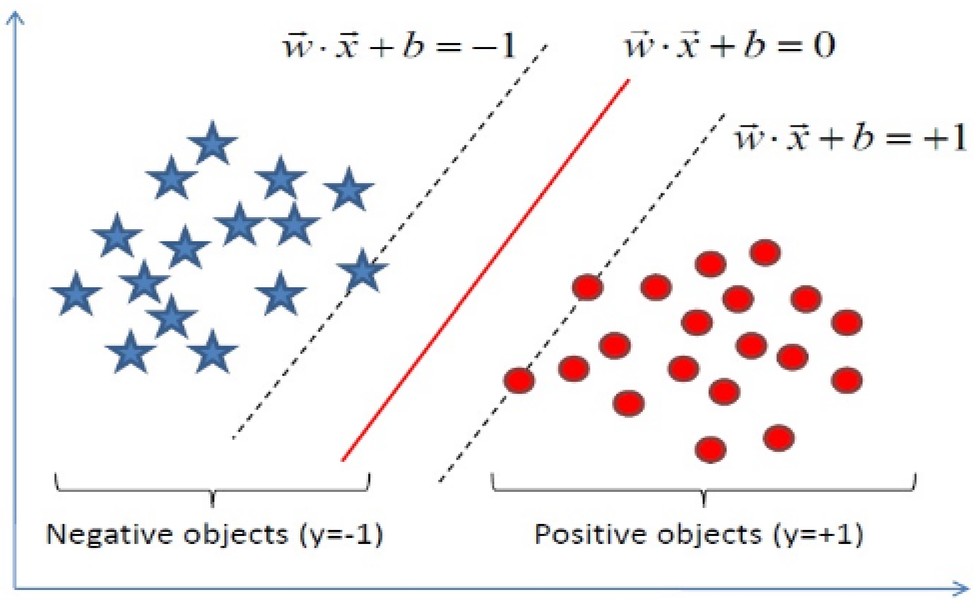
$$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{||w||}\ ;\ \gamma = \frac{\hat{\gamma}}{||w||}$$

1.3 间隔最大化

点到平面的距离公式

$$d = \frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}}$$

接下来再介绍一个概念：支持向量。如下图所示，假设中间的红线是超平面，那么虚线上的样本点就是支持向量。



这样一来，我们就可以定义间隔最大化：

$$\max\ \gamma = \frac{\hat{\gamma}}{||w||}$$

需要满足的条件是：

$$y_i(w^T x_i + b) = \hat{\gamma}_i \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

间隔的有效的样本点是支持向量的点，即是**符合函数间隔等于 1 的点**。那么就可以对间隔的最大化可以为：

$$\max \frac{1}{||w||}, \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

1.4 对偶问题

对偶问题^2d2e84
由上面的式子经过变化，得到凸优化问题，即 将max 1/ || ω || 转化成求 min 1/2 || ω || 2
，两者是等价的。

$$\min \frac{1}{2}\|w\|^2 \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1, \ldots, n$$

转化到这个形式后，我们的问题成为了一个凸优化问题，或者更具体的说，因为现在的目标函数是二次的，约束条件是线性的，所以它是一个凸二次规划问题。这个问题可以用任何现成的QP的优化包进行求解，归结为一句话即是：**在一定的约束条件下，目标最优，损失最小。**
除了用解决QP问题的常规方法之外，还可以通过求解对偶问题得到最优解，这就是线性可分条件下支持向量机的对偶算法，这样做的优点在于：一者对偶问题往往更容易求解；二者可以自然的引入核函数，进而推广到非线性分类问题。
现在要对凸优化进行求解，现在给出的经验求解方法是，通过求解对偶问题得到最优解。对偶问题是什么?这就要引出拉格朗日函数，将约束条件融合到目标函数中，通过给每个约束条件加上一个拉格朗日乘子 α 。

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i\Big(y_i(w^T x_i + b) - 1\Big)$$

转为对偶问题

.原问题是极小极大问题

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

原问题的对偶问题是极大极小问题 $\max_{\alpha}\min_{w,b} L(w,b,\alpha)$

原问题为：

$$\min_{w,b}\theta(w) = \min_{w,b}\max_{\alpha_i\geq 0}\mathcal{L}(w,b,\alpha) = p^*$$

对这个公式的理解：在满足约束条件（即样本点为支持向量）下，最大的几何间隔中取最小的值。max L(w,b,a) 是得到支持向量的样本点，然后再这些样本点中找到min 1/2 || w ||2 的最小值（最大间隔）。

注意了，对偶问题出来了（p *是原来的问题，* d 是对偶问题），将最小最大的位置交换，得到下面的式子：

$$\max_{\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w,b,\alpha) = d^*$$

交换之后的问题就不等价了，有以下关系：d $<=p$。但在满足某些条件下（KKT条件），两者是相等的。现在我们把对偶问题得出了，但怎么对偶问题进行求解呢？继续向下看，最终的目标式子就要出来了。

1.5 对偶问题求解

上述可知，原来问题与对偶问题的等价条件是满足某些条件，条件就是KKT 。不知道KKT条件没关系，现在你只需要明白原来问题是满足KKT条件，KKT 条件在核函数中讲解。而求解这个对偶学习问题，分为3个步骤：1、让L（ω, b, α）关于 ω 和 b 的最小化；2、求 α 的极大；3、利用SMO算法求解对偶因子（α）。这里主要讲1 和 2， 步骤 3 在第二部分讲。

二、非线性可分

若数据线性不可分，则增加松弛因子

$$\xi_i \geq 0$$

，使函数间隔加上松弛变量大于等于1，则约束条件变成

$$y_i(wx_i + b) \geq 1 - \xi_i$$

目标函数

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

（这里是为了保证松弛因子不至于过大）
此时的凸优化为

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \\ &s.t. \quad y_i(wx_i + b) \geq 1 - \xi_i, \quad i = 1,2,\cdots,n \\ &\qquad\quad \xi_i \geq 0, \ i = 1,2,\cdots,n \end{aligned}$$

拉格朗日函数为

$$L(w,b,\xi,\alpha,u)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(wx_i+b)-1+\xi_i)-\sum_{i=1}^{n}u_i\xi_i$$

对w, b, ξ求偏导

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i\Phi(x_n)$$

$$\frac{\partial L}{\partial b}=0\Rightarrow 0=\sum_{i=1}^{n}\alpha_i y_i$$

$$\frac{\partial L}{\partial \xi}=0\Rightarrow C-\alpha_i-u_i=0$$

将三式代入L中，得到

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

对上式求关于α的极大，得到：

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

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

$$\left.\begin{array}{l}C-\alpha_i-u_i=0\\ \alpha_i\geq 0\\ u_i\geq 0,\ i=1,2,\cdots,n\end{array}\right\}\Rightarrow 0\leq\alpha_i\leq C$$

整理，得到对偶问题的最优化问题

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

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

$$0 \le \alpha_i \le C,\ \ i = 1,2,\cdots,n$$

求得最优解

$$\alpha^*$$

计算

$$w^* = \sum_{i=1}^{n} a_i^* y_i x_i$$

$$b^* = \frac{\max_{i:y_i=-1} w^* x_i + \min_{i:y_i=1} w^* x_i}{2}$$

实践中往往取支持向量的所有值取平均，作为b*

求得分离超平面

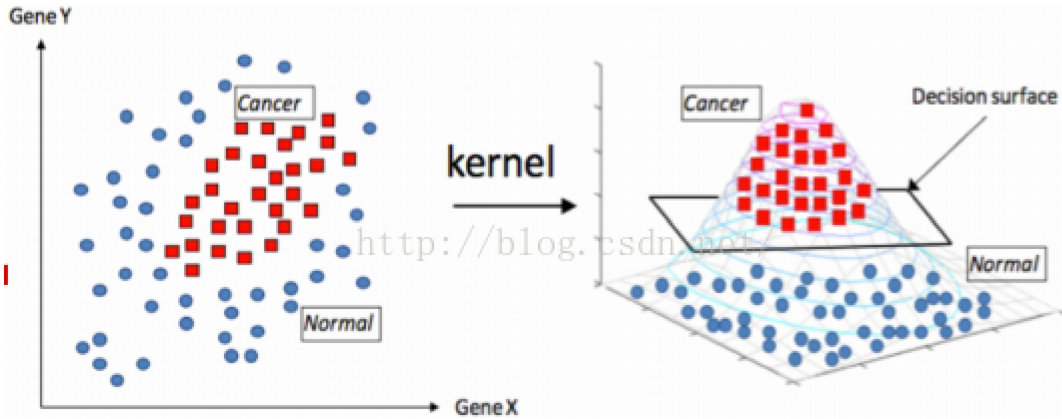
$$w^* x + b^* = 0$$

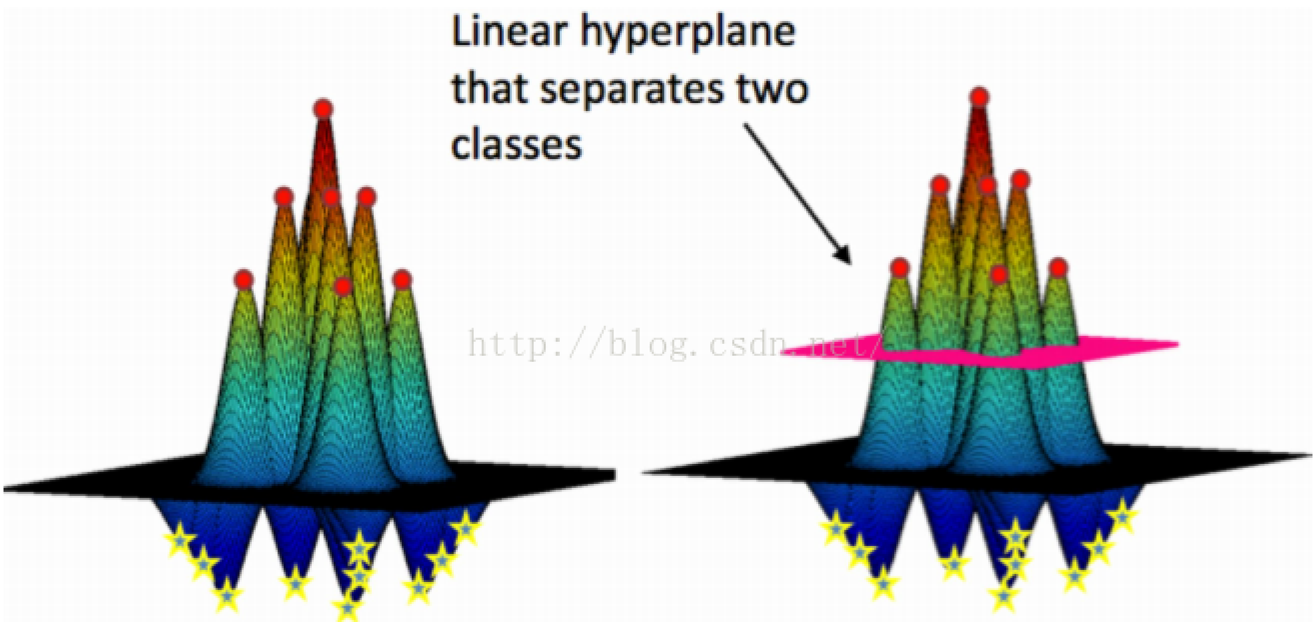
分类决策函数为

$$f(x) = sign(w^* x + b^*)$$

三、核函数

可以使用核函数，将原始输入空间映射到新的特征空间，从而使得原本线性不可分的样本可在核空间可分。
在实际应用中，往往依赖先验领域知识或交叉验证等方案才能选择有效的核函数。没有更多先验信息，则使用高斯核函数。





高斯核是无穷维的，因为

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + R_n$$

知识库

[知乎回答对偶问题]()