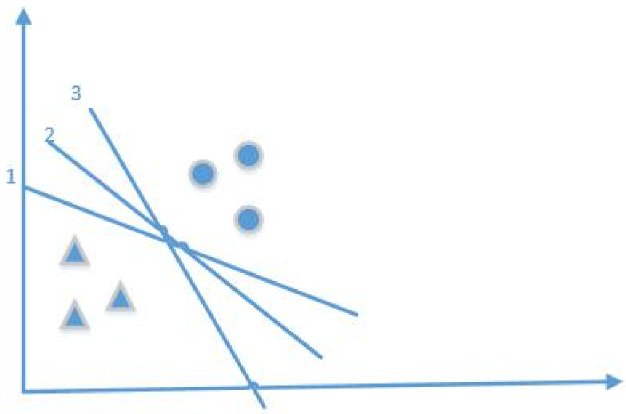
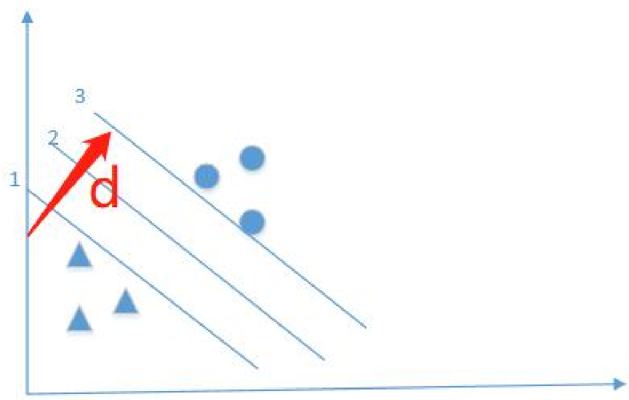
SVM 算法

1、线性模型

定义：可用一条线性直线，将数据集分开的模型，称之为线性模型。



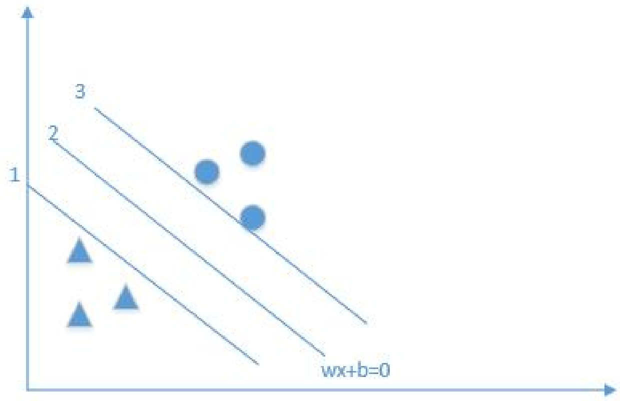
从上图看出，显然 2 号线划分得最好。那如何定义 2 号线呢？



2 号线在 1、3 号线的中间，很自然，d 越大，所找线越优。
d: 间隔 margin
中间 2 号线：决策边界

引来的问题是：如何寻找最大化间隔 d?

2、决策边界概述



wx+b＞0，规定在 2 号线上方，属于+1 类；
wx+b＜0，规定在 2 号线下方，属于-1 类；

SVM 有 3 种分类：

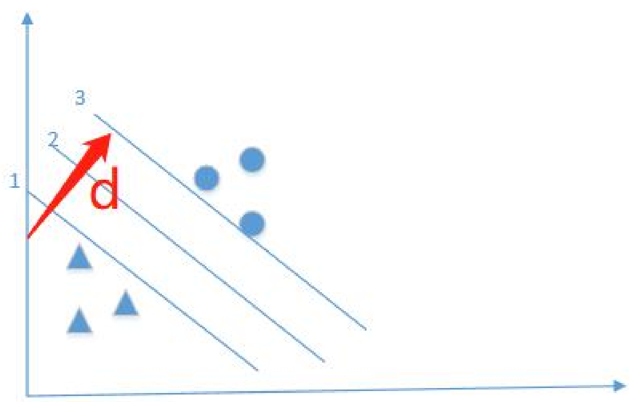
- 数据线性可分：硬间隔 SVM
- 数据近似线性可分：软间隔 SVM
- 数据不可线性分：核技巧

3、线性可分 SVM

最优化问题导出：最大化间隔 d

1) 二维数据：寻找一条直线 wx+b=0，可将数据切分开；
2) 三维空间：寻找平面：ax+by+c=0，可将数据切分开；
3) 高维空间：寻找超平面；

以二维数据为例：



直线 2：wx+b=0　　　②

因为 1、3 号线是 2 平移得到的，所以：

1 号线：wx+b=k　　　①

3 号线：wx+b=-k　　③

此时，对①③做简单处理，两边除以 k：

$$\begin{cases} \frac{\mathrm{w}}{k}x+\frac{b}{k}=1 \\ \frac{w}{k}x+\frac{b}{k}=-1 \end{cases} \qquad \mathrm{II}$$

对②，两边乘以 $\frac{1}{\mathrm{k}}$

$$\frac{\mathrm{w}}{\mathrm{k}}x+\frac{b}{k}=0$$

令 $\frac{\mathrm{w}}{k}=w^{new}, \frac{b}{k}=w^{new}$，带入II中

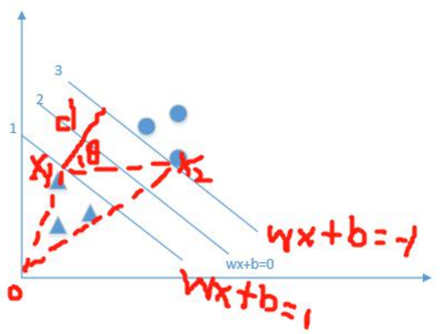
$$w^{new}\mathrm{x}+b^{new}=1 \qquad ①$$

$$w^{new}\mathrm{x}+b^{new}=-1 \qquad ③$$

$$w^{new}\mathrm{x}+b^{new}=0 \qquad ②$$

综合上述所得，可以写成：

$$\begin{cases} \mathrm{wx}+b=1 \\ \mathrm{wx}+b=-1 \\ \mathrm{wx}+b=0 \end{cases}$$



$$\begin{cases} \mathrm{wx}_1 + b = 1 \\ \mathrm{wx}_2 + b = -1 \end{cases}$$

两式相减：

$$\mathrm{w(x1\text{-}x2)=2}$$

$$\| \mathrm{w} \| \cdot \| \mathrm{x}_1 - \mathrm{x}_2 \| \cdot \cos\theta = 2$$

又因为由上图得出：

$$\cos\theta = \frac{\mathrm{d}}{\| x_1 - x_2 \|}$$

所以得：

$$\| \mathrm{w} \| \cdot \mathrm{d} = 2$$

$$\mathrm{d} = \frac{2}{\| \mathrm{w} \|}$$

因为要求 d 最大，所以||w||要最小

$$\max_d <==> \min \frac{1}{2} \| w \|^2$$

因此问题转化：最大间隔 d　<===>　最小化 $\frac{1}{2} \| \mathrm{w} \|^2$

又因为：

$\mathrm{w}x + b \geq 1$　+1 类　y=+1

$\mathrm{w}x + b \leq -1$　-1 类　y=-1

$==> \mathrm{y}_i (wx + b) \geq 1$

所以 SVM 的最优化问题：

$$\begin{cases} \min \frac{1}{2} \| w \|^2 \\ s \cdot t \quad y_i (wx + b) \geq 1 \end{cases}$$

3.1 SVM 算法最优化问题求解

SVM 优化问题求解：

$$\begin{cases} \min \frac{1}{2} \| w \|^2 \\ s \cdot t \quad y_i (wx + b) \geq 1 \end{cases}$$

在带有约束条件的问题，要求解该问题的最优解，一般采用的是拉格朗日乘数法；

举个简单例子：求 $\mu = \mathrm{xy} + 2yz$ ，在条件 $\mathrm{x}^2 + y^2 + z^2 = 10$ 下的极值

1）构建拉格朗日函数：

$$\mathrm{L}(x, y, z, \partial) = zy + 2yz + \partial(x^2 + y^2 + z^2 - 10)$$

2）求解

$$\begin{aligned} &\mathrm{L}_{\mathrm{x}}' = y + 2\partial x = 0 \\ &L_y' = x + 2z + 2\partial y = 0 \\ &L_z' = 2y + 2\partial z = 0 \\ &L_\partial' = x^2 + y^2 + z^2 \text{-} 10 = 0 \end{aligned}$$

求解得： $\mu_{\max} = 5\sqrt{5}, \mu_{\min} = -5\sqrt{5}$

对 SVM 优化问题求解：

$$\begin{aligned} \mathrm{L}(\mathrm{w}, \partial) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \partial_i [1 - y_i(wx + b)] \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \partial_i \text{-} \sum_{i=1}^{\mathrm{n}} \partial_i y_i (wx + b) \end{aligned}$$

$$\begin{aligned} &\frac{\partial \mathrm{L}}{\partial w} = w - \sum_{i=1}^{n} \partial_i (y_i x_i) = 0 \\ &\frac{\partial \mathrm{L}}{\partial b} = -\sum_{i=1}^{n} \partial_i y_i = 0 \end{aligned}$$

3.2 补充一个知识点（原问题与对偶问题）

假如**原问题**：

$$\min f(w)$$
$$s \cdot t \begin{cases} g_i(w) \le 0 \\ h_i(w) = 0 \end{cases}$$

则原问题的拉格朗日函数：

$$\begin{aligned} \mathrm{L}(w, \partial, \beta) &= f(w) + \sum_{i=1}^{n} \partial_i g_i(w) + \sum_{i=1}^{n} \beta_i h_i(w) \\ &= f(w) + \partial^T g(w) + \beta^T h(w) \end{aligned}$$

对偶问题：

$$\theta(\partial, \beta) = \inf L(w, \partial, \beta)$$

意思就是：

1）先以 w 为参数，求 $L(w,\partial,\beta)$ 的最小值

2）再以 ∂,β 为参数，求 $L(w,\partial,\beta)$ 的最大值

KKT 条件：

$$\begin{cases} \partial_{\mathrm{i}},\beta_i > 0 \\ g(w) \le 0 \\ h(w) = 0 \end{cases}$$

在满足 KKT 条件后，我们则可以将原始的最优化问题转换成求对偶函数的最优化问题；

3.3 求解 SVM 最优化问题

$$\begin{aligned} \mathrm{L}(\mathrm{w},\partial) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\partial_i[1-y_i(wx+b)] \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\partial_i - \sum_{i=1}^{\mathrm{n}}\partial_i y_i(wx+b) \end{aligned}$$

$$\frac{\partial \mathrm{L}}{\partial w} = w - \sum_{i=1}^{n}\partial_i(y_i x_i) = 0$$

$$\frac{\partial \mathrm{L}}{\partial b} = -\sum_{i=1}^{n}\partial_i y_i = 0$$

所以得：

$$w = \sum_{i=1}^{n}\partial_i(y_i x_i)$$

$$\sum_{i=1}^{n}\partial_i y_i = 0 \qquad \mathrm{III}$$

将其带回拉格朗日函数中得：

先以 w 为参数，求最小值：

$$
\begin{aligned}
\mathrm{L}(\mathrm{w},\mathrm{b},\partial)_{\min} &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\partial_i[1-y_i(wx+b)] \\
&= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}\partial_i - \sum_{i=1}^{\mathrm{n}}\partial_i y_i(wx+b) \\
&= \frac{1}{2}[\sum_{i=1}^{n}\partial_i(y_i x_i)]^2 + \sum_{i=1}^{n}\partial_i - \sum_{i=1}^{n}\partial_i y_i[(\sum_{i=1}^{n}\partial_i y_i x_i)x_j + b] \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\partial_j y_i y_j x_i x_j + \sum_{i=1}^{n}\partial_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\partial_j y_i y_j x_i x_j - \sum_{i=1}^{n}\partial_i \mathrm{y}_\mathrm{i}\mathrm{b} \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\partial_j y_i y_j x_i x_j + \sum_{i=1}^{n}\partial_i
\end{aligned}
$$

先以∂为参数，求最大值：

$$\max[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\partial_j y_i y_j x_i x_j + \sum_{i=1}^{n}\partial_i]$$

等价于求：

$$\min[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\partial_i\partial_j y_i y_j x_i x_j - \sum_{i=1}^{n}\partial_i]$$

最后通过 SMO 算法求解得：

$$\partial^* = (\partial_1{}^*, \partial_2{}^*, \partial_3{}^* \ldots, \partial_k{}^*)$$

所以，又由III式可知，求得 w:

$$w^* = \sum_{i=1}^{n}\partial_i{}^*(y_i x_i)$$

又因为对 SVM 的 KKT 条件是：

$$
\begin{cases}
1-\mathrm{y}_i(wx+b)=0\text{或}\partial_i{}^*=0 \\
\dfrac{\partial \mathrm{L}}{\partial w} = w - \sum_{i=1}^{n}\partial_i(y_i x_i) = 0 \\
\dfrac{\partial \mathrm{L}}{\partial b} = \sum_{i=1}^{n}\partial_i y_i = 0
\end{cases}
$$

所以由 $\mathrm{y_i}(wx_i+b)-1=0$，两边乘以 $\mathrm{y_i}$，所以得 $\mathrm{y_i}^2(wx_i+b)-\mathrm{y_i}=0$

又因为 $\mathrm{y_i}$ 是+1，或-1，所以 $\mathrm{y_i}^2$=1

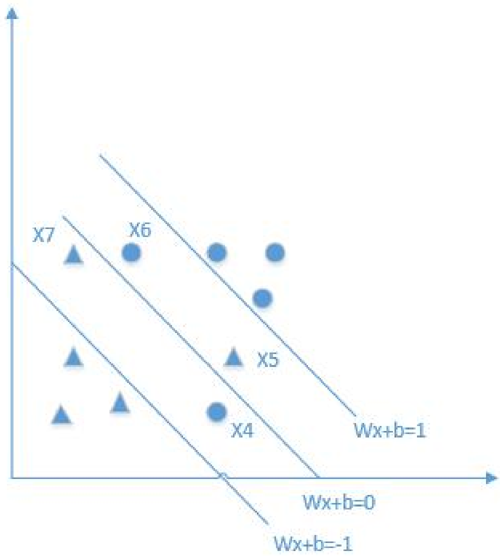
所以：$wx_i + b - y_i = 0$

所以得：$b^* = y_i - w^* x_i$，再将 w*带入求得：

$$b^* = y_i - \sum_{i=1}^{n} \partial_i^* y_i x_i x_j$$

4、近似线性可分：最大软间隔

近似线性可分：在线性可分的情况下，加入一些噪声数据点，噪声数据点满足的约束条件是：$0 < y_i(wx+b) < 1$



噪声点：X6,X7

$$\begin{aligned} 0 < wx_6 + b < 1 \qquad y = +1 \\ -1 < wx_7 + b < 0 \qquad y = -1 \end{aligned}$$

错误点：X4,X5

总结：

正确分类的点满足：

$$y_i(wx+b) \geq 1$$

错误分类的点满足：

$$y_i(wx+b) \leq 0$$

因此对于正确分类的点：

$$y_i(wx+b)\geq 1-\varepsilon_i \quad \varepsilon_i\text{是松弛因子}$$

希望ε_i越小越好；

- 当$\varepsilon_i=0$时，$y_i(wx_i+b)\geq 1$
- 当$0<\varepsilon_i<1$时，$y_i(wx_i+b)\geq 1-\varepsilon_i$，则此时存在噪声点；
- 当$\varepsilon_i>1$时，$y_i(wx_i+b)\geq 1-\varepsilon_i$，则此时存在错误分类的点；

SVM 的最优化问题:

$$\begin{cases} \min \dfrac{1}{2}\|w\|^2+C\cdot\sum_{i=1}^{n}\varepsilon_i \\ s\cdot t \quad \begin{cases} y_i(wx_i+b)\geq 1-\varepsilon_i \\ \varepsilon_i\geq 0 \end{cases} \end{cases}$$

- C↑，ε_i↓，即噪声少，超平面向内侧相互移动，减少间隔距离，从而减少噪声，此时，支持向量减少；
- C↓，ε_i↑，即噪声多，超平面向外侧相互移动，增加间隔距离，从而增加噪声，此时，支持向量减多；

5、数据不可线性分(非线性 SVM-核技巧)

核技巧：低维非线性=>映射到高维空间中，即可实现线性分离；

常见的核函数：

1）线性核函数：$k(x_i,x_j)=x_i\cdot x_j$

2）多项式核函数：$k(x_i,x_j)=(\gamma(x_i\cdot x_j)+\gamma)^p$

3）高斯核函数：$k(x_i,x_j)=\mathrm{e}^{-\gamma(|x_i-x_j|)^2}$

4）Sigmoid 核函数：$k(x_i,x_j)=\tanh h\gamma[(x_i-x_j)+\gamma]$