机器学习

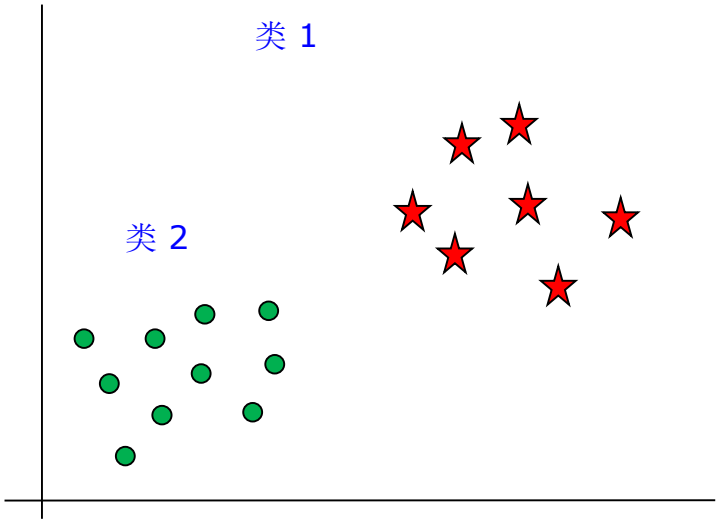
于元隆、朱丹红

福州大学计算机与大数据学院
Email: yu.yuanlong@fzu.edu.cn

线性判据
——支持向量机

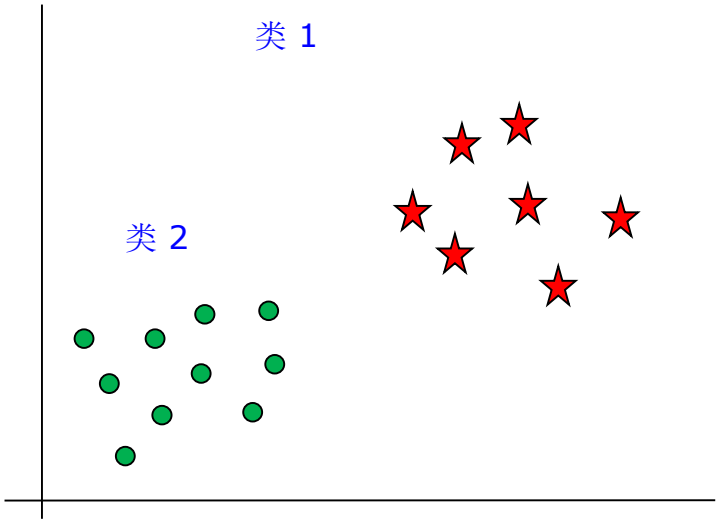
支持向量机：设计动机

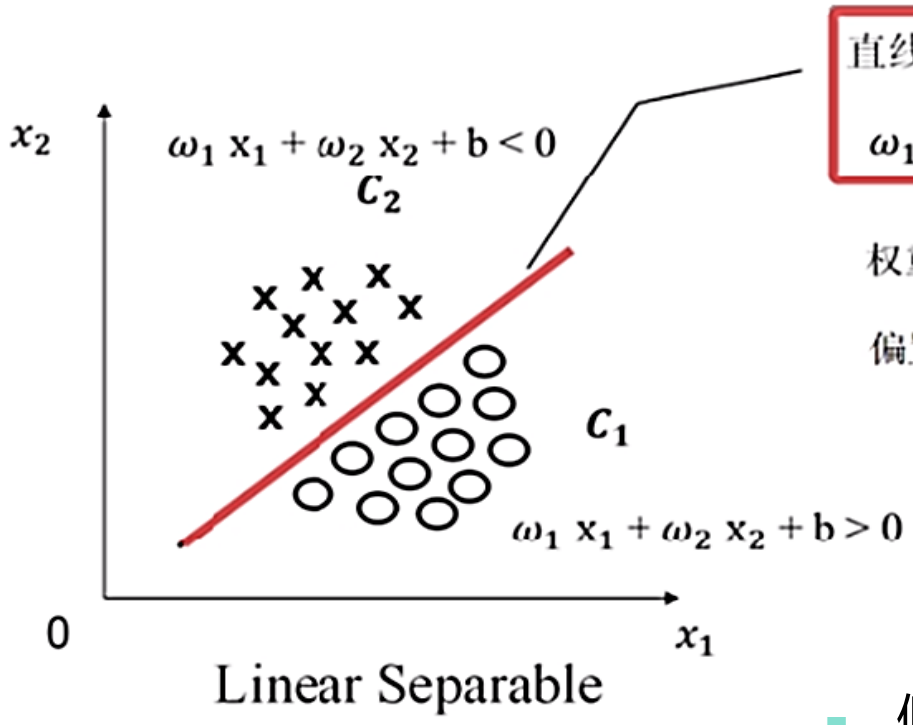
- 如果两个类是线性可分的，则存在着多个线性决策平面用作决策边界。
- 感知机从最小化分类误差角度来设计。
- Fisher线性判据首先降维到一维空间，从最大化类间距离同时最小化类内散度的角度来设计。
- 如何拥有更好的泛化性能?



支持向量机：设计动机

- 如果两个类是线性可分的，则存在着多个线性决策平面用作决策边界。
- 感知机从最小化分类误差角度来设计。
- Fisher线性判据首先降维到一维空间，从最大化类间距离同时最小化类内散度的角度来设计。
- 如何拥有更好的泛化性能?





- 假设：有N个训练样本和标签：

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_N, y_N)\}$$

其中 $X_i = [x_{i1}, x_{i2}]^T$

$y_i = \{+1, -1\}$

x_i属于$C1$　　x_i属于$C2$

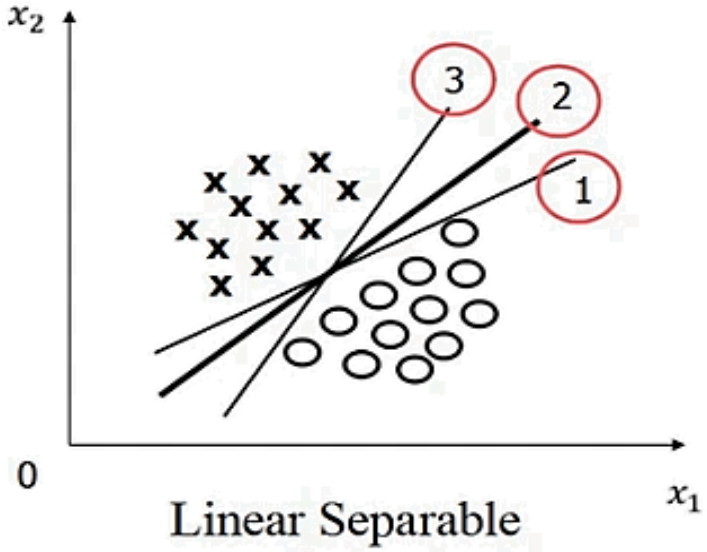
- 线性可分的向量形式定义：

(1) 若$y_i = +1$，则 $\omega^T X_i + \mathrm{b} > 0$

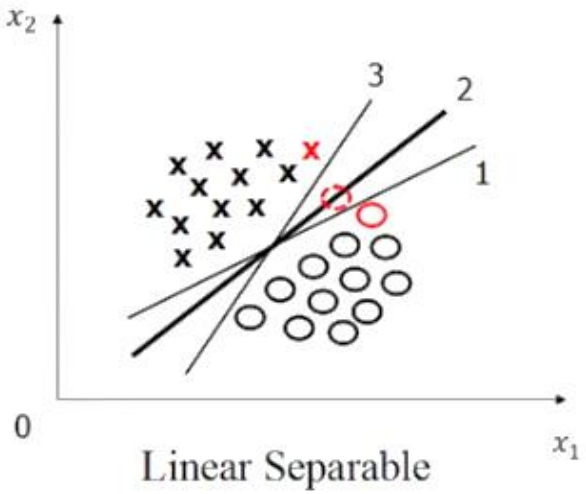
(2) 若$y_i = -1$，则 $\omega^T X_i + \mathrm{b} < 0$

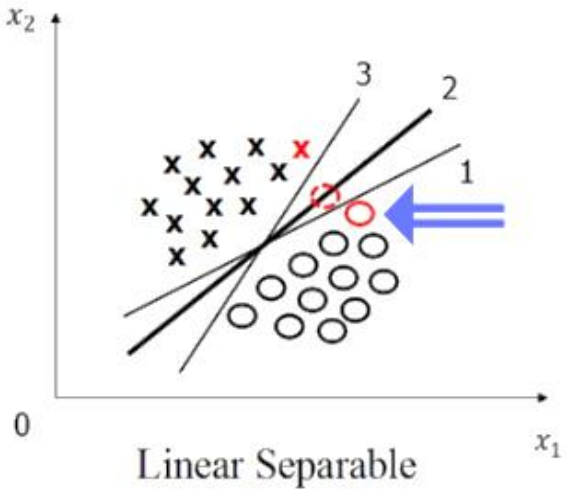
$$y_i\left(\omega^T X_i + b\right) > 0$$

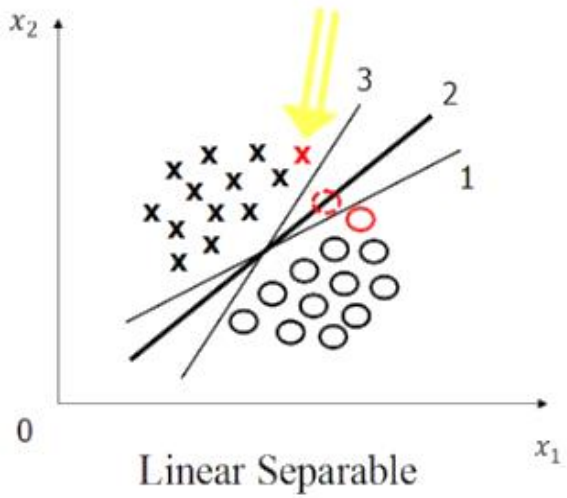
- 如果一个数据集是线性可分的，存在多少个超平面将各个类别分开？



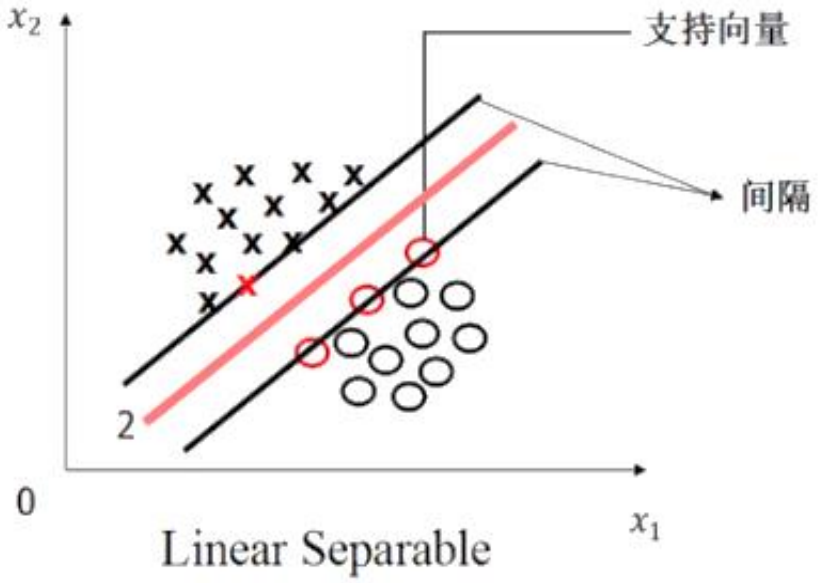
- 假设训练样本的位置在特征空间上有测量误差：







- 2号线更能抵御训练样本位置的测量误差。

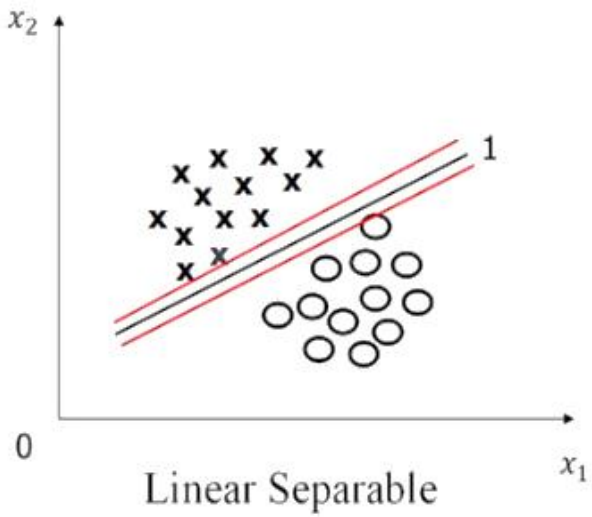


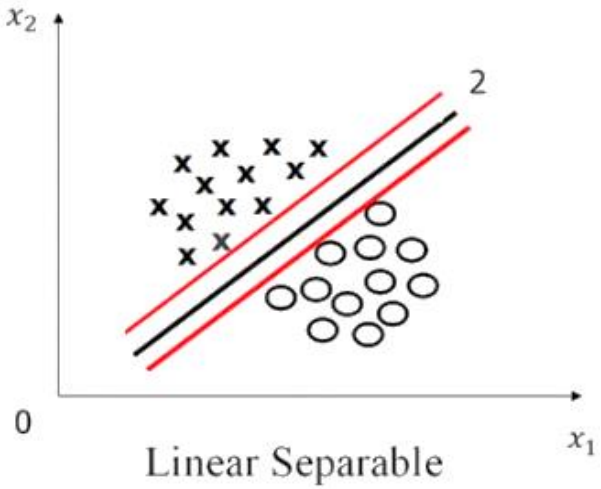
- 间隔（$Margin$）最大的是2号线。

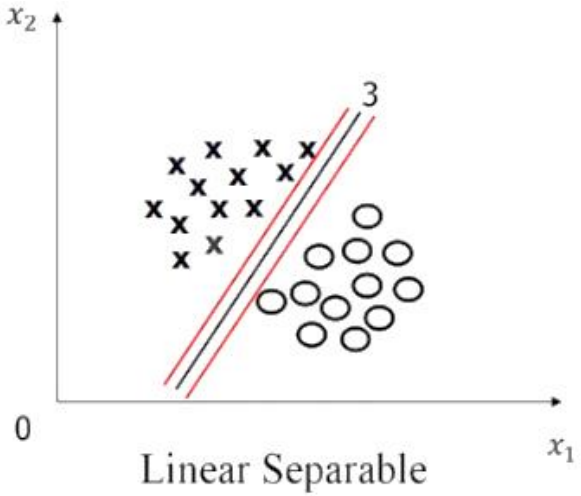
- 怎样画出2号线？基于最优化理论。

Vladimir Vapnik

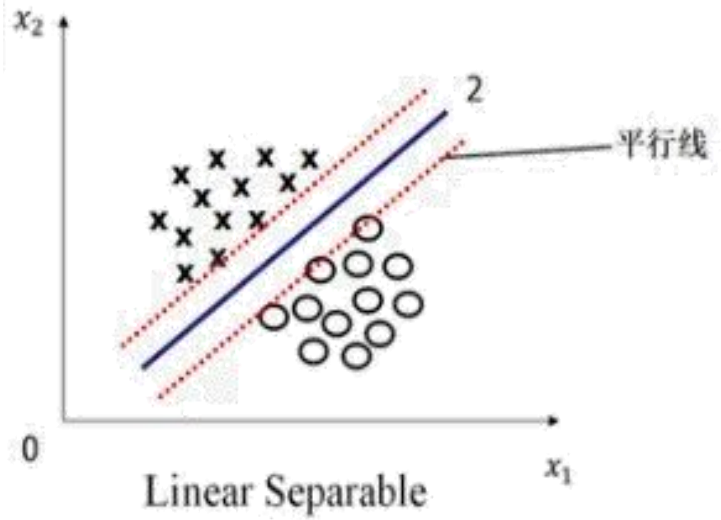
- 哪种方式得到的间隔最大?



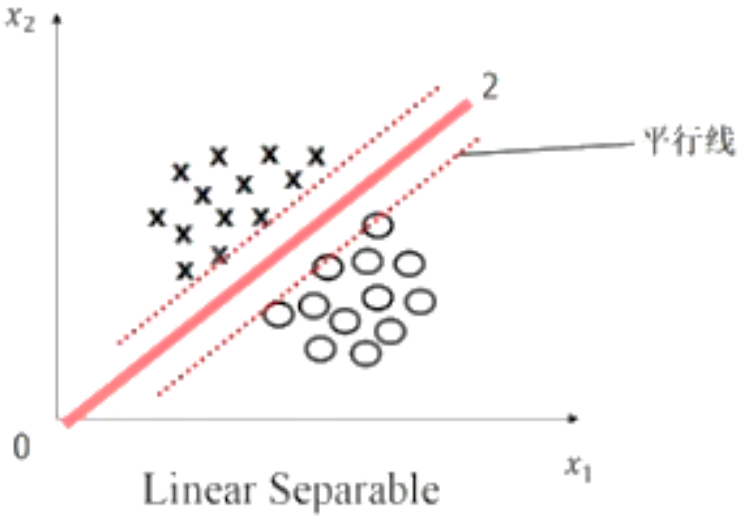




- 使用$Margin$最大还不能唯一确定一条直线。



- 这条线应在平行线的中间。



支持向量机：设计思想

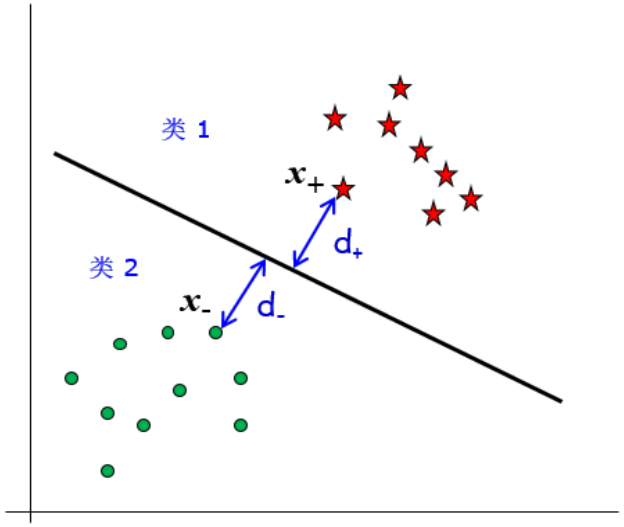
- 支持向量机寻找的最优分类直线（决策边界）应满足：

  （1）该直线分开了两类；

  （2）该直线最大号间隔；

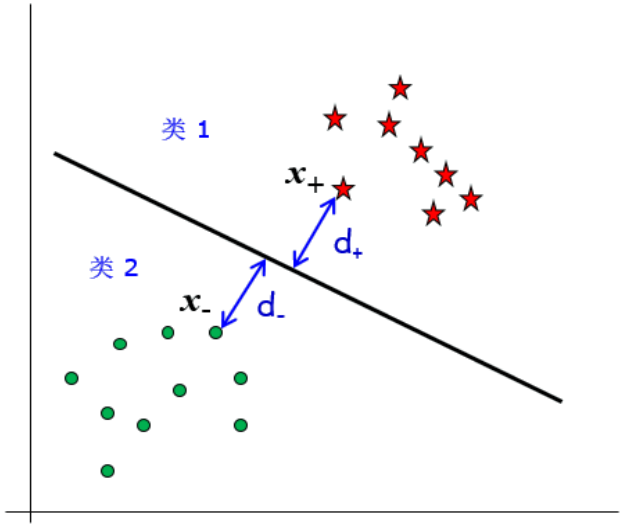
  （3）该直线处于间隔的中间，两个类中与决策边界最近的训练样本到决策边界之间的间隔最大。

- 高维空间中，直线==➔超平面
- 寻找最优分类超平面==➔最优化问题



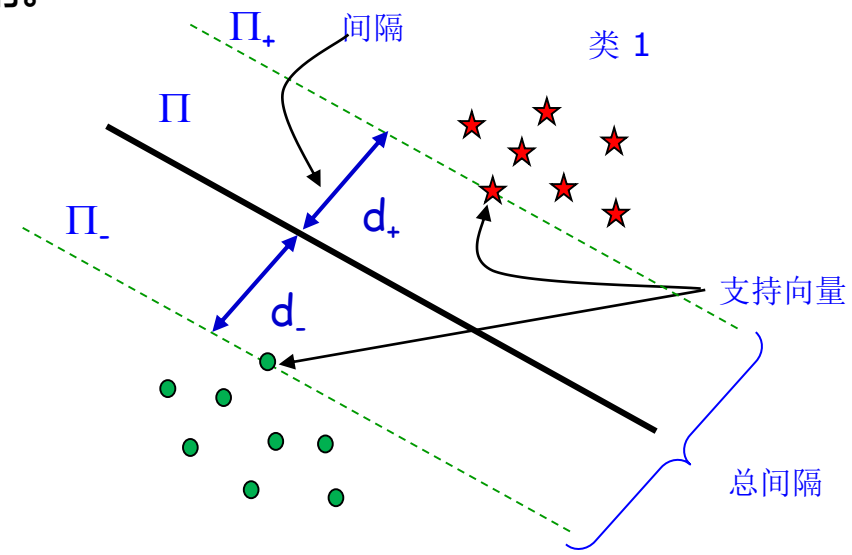
间隔的数学定义

- 间隔的数学定义：
- 在两个类的训练样本中，分别找到与决策边界最近的两个训练样本，记作$\boldsymbol{x}_+$和$\boldsymbol{x}_-$ 。
- $\boldsymbol{x}_+$和$\boldsymbol{x}_-$到决策边界的垂直距离叫作间隔，记作d_+和d_- 。



支持向量的概念

- 决策边界记作Π，平行于Π且分别通过x_+和x_-的两个超平面记作Π_+和Π_-，称为间隔边界。
- 没有任何训练样本落在这两个超平面中间的间隔区域。
- 位于超平面Π_+和Π_-上的样本被称为支持向量（Support vector）。
- 支持向量在确定决策边界Π中起到核心作用。



分类器重新表达

- 在支持向量机中，正负类训练样本输出真值分别用+1和-1来表达。
- 给定标记过的训练样本$\{(\boldsymbol{x}_n, t_n)\}$，线性分类器可以表达为：

$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 > 0 \quad t_n = +1, \quad Class\ C_1$$
$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 < 0 \quad t_n = -1, \quad Class\ C_2$$

- 加入间隔的概念，引入一个正常数Δ，分类器进一步表达为：

$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 \geqslant +\Delta \quad t_n = +1$$
$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 \leqslant -\Delta \quad t_n = -1$$

意味着没有训练样本落在±Δ间隔范围内。

- 加入间隔的概念，引入一个正常数Δ，分类器进一步表达为：

$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 \geqslant +1 \quad t_n = +1$$
$$\boldsymbol{w}^T\boldsymbol{x}_n + w_0 \leqslant -1 \quad t_n = -1$$
$$\longrightarrow \quad t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1 \geqslant 0, \quad \forall n$$

支持向量机：支持向量

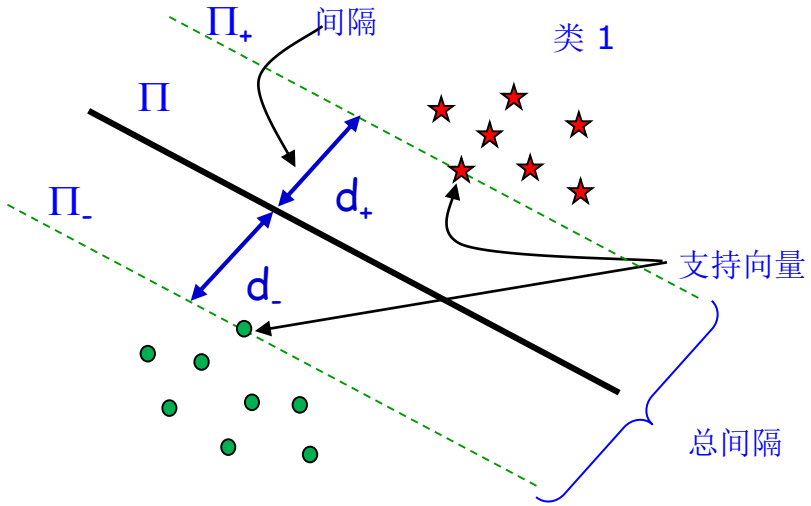
- 分类器新的表达中，什么时候等式成立呢?

$$t_n\left(\boldsymbol{w}^T\boldsymbol{x}_n + w_0\right) - 1 \geqslant 0, \quad \forall n$$

- 当$\boldsymbol{x}_n$是支持向量时，等式成立，例如， $\boldsymbol{x}_+$和$\boldsymbol{x}_-$。

$$\boldsymbol{w}^T\boldsymbol{x}_+ + w_0 = +1$$

$$\boldsymbol{w}^T\boldsymbol{x}_- + w_0 = -1$$



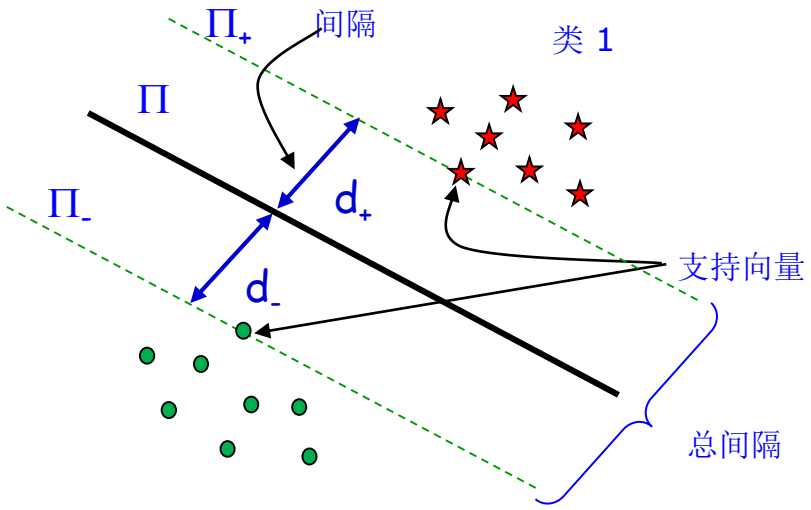
支持向量机：间隔计算

- 根据线性判据的几何含义，点$\boldsymbol{x}_+$到决策边界Π的距离为：

$$r = \frac{f(\boldsymbol{x}_+)}{||\boldsymbol{w}||}$$

$$\Rightarrow \ d_+ = |r| = \frac{|f(\boldsymbol{x}_+)|}{||\boldsymbol{w}||}$$

$$\Rightarrow \ d_- = |r| = \frac{|f(\boldsymbol{x}_-)|}{||\boldsymbol{w}||}$$



- 所以，总间隔为：

$$M = d_+ + d_- = \frac{|f(\boldsymbol{x}_+)| + |f(\boldsymbol{x}_-)|}{||\boldsymbol{w}||}$$

- 由于$\boldsymbol{x}_+$和$\boldsymbol{x}_-$是位于间隔边界的支持向量，可以得到：

$$M = d_+ + d_- = \frac{|f(\boldsymbol{x}_+)| + |f(\boldsymbol{x}_-)|}{||\boldsymbol{w}||}$$

$$= \frac{2}{||\boldsymbol{w}||}$$

支持向量机：目标函数

- 支持向量机（support vector machine, SVM）的目标：最大化总间隔。
- 最大化间隔，等价于最小化||$\boldsymbol{w}$||，所以目标函数设计为：

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||_2^2 \quad \Longrightarrow \quad \min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

- 同时满足如下约束条件：

$$s.t. \quad t_n(\mathbf{w}^T\boldsymbol{x}_n + w_0) - 1 \geqslant 0 \quad \forall n$$

当x_n位于超平面Π_+和Π_-上时，该约束条件等于0，其余情况都为大于0。

- 该目标函数是条件优化问题(Constrained Optimization)。
- 目标函数自变量为$\boldsymbol{w}$，是关于$\boldsymbol{w}$的二次型函数。约束条件是关于$\boldsymbol{w}$的仿射函数（线性函数）。

拉格朗日乘数法

条件优化问题

- 支持向量机的目标函数是一个条件优化问题(Constrained Optimization)。
- 拉格朗日乘数法（Lagrange Multiplier）是常用的解决该类问题的方法。

不等式约束优化问题

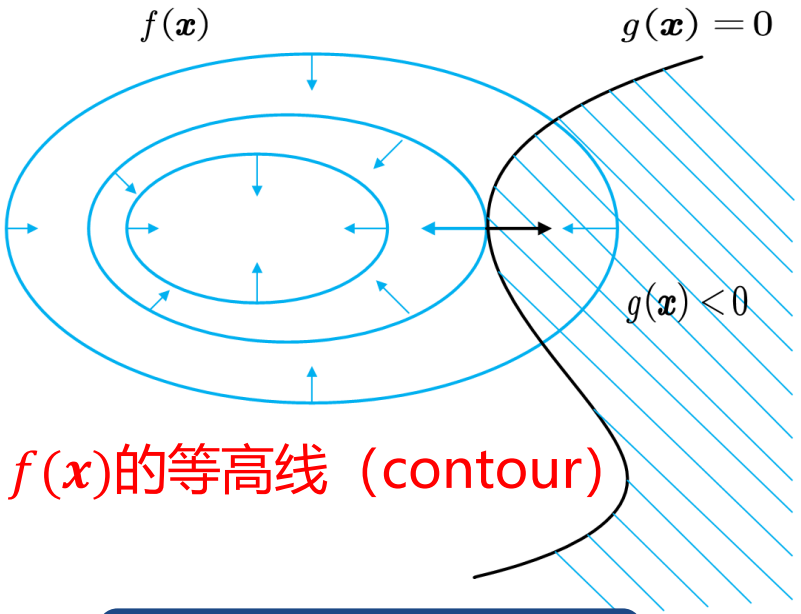
$$\min\ f(\boldsymbol{x})$$
$$s.t.\ \ g(\boldsymbol{x}) \leq 0$$

- 可行域（Feasible region）：$g(x) \leq 0$的区域。

等式约束优化问题

$$\min\ f(\boldsymbol{x})$$
$$s.t.\ \ g(\boldsymbol{x}) = 0$$

- 可行域：$g(x) = 0$的区域。



$f(\boldsymbol{x})$的等高线（contour）

条件优化求解

- 在可行域内寻找$f(\boldsymbol{x})$的最小值。

拉格朗日乘数法：等式约束

等式约束优化求解思路

- 函数在等高面上任意一点的梯度方向与其等高面（切线方向）正交，且朝向（即正方向）函数值较高方向。
- $f(\boldsymbol{x})$的极值点$\boldsymbol{x}^*$必须位于曲线$g(\boldsymbol{x})=0$上。
- 搜寻极值点$\boldsymbol{x}^*$：沿着$g(\boldsymbol{x})=0$的切线方向、向着$f(\boldsymbol{x})$负梯度方向移动。当出现沿着切线方向、无法再向$f(\boldsymbol{x})$负梯度方向移动时停止。

- 函数在等高面上任意一点的梯度方向与其等高面（切线方向）正交，且朝向（即正方向）函数值较高方向。
- $f(\boldsymbol{x})$的极值点$\boldsymbol{x}^*$必须位于曲线$g(\boldsymbol{x})=0$上。
- 搜寻极值点$\boldsymbol{x}^*$：沿着$g(\boldsymbol{x})=0$的切线方向、向着$f(\boldsymbol{x})$负梯度方向移动。当出现沿着切线方向、无法再向$f(\boldsymbol{x})$负梯度方向移动时停止。

蓝色箭头：$f(\boldsymbol{x})$负梯度方向，即梯度下降最快方向，$f(\boldsymbol{x})$取值沿着该方向不断减小。

等式约束：拉格朗日函数

拉格朗日函数

- 因此，存在一个$\lambda \neq 0$，使得：$f(\boldsymbol{x})$与$g(\boldsymbol{x})$的梯度记作$\nabla f(\boldsymbol{x})$和$\nabla g(\boldsymbol{x})$ 。

$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0,$$

- λ：拉格朗日乘子（可正可负），无符号限制。
- 由此，可以定义一个拉格朗日函数：

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

- 拉格朗日函数满足驻点(stationary point)条件和约束条件：

$$\frac{\partial L}{\partial \boldsymbol{x}} = 0 \Rightarrow \nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0$$

$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow g(\boldsymbol{x}) = 0$$

等式约束：等价优化问题

等价优化问题

- 因此，因等式约束问题可以转换为等价的不带约束的优化问题：

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x})$$

$$s.t. \quad g(\boldsymbol{x}) = 0$$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

拉格朗日乘数法：不等式约束

情况1：极值点落在可行域内

- 假设极值点$\boldsymbol{x}^*$落在了可行域内（不含边界），即极值点位于区域$g(\boldsymbol{x}) < 0$范围内。
- 这种情况下，约束条件不起作用。
- 直接通过$\nabla f(\boldsymbol{x}) = 0$获得极值点。
- 此时，在极值点$\boldsymbol{x}^*$上：

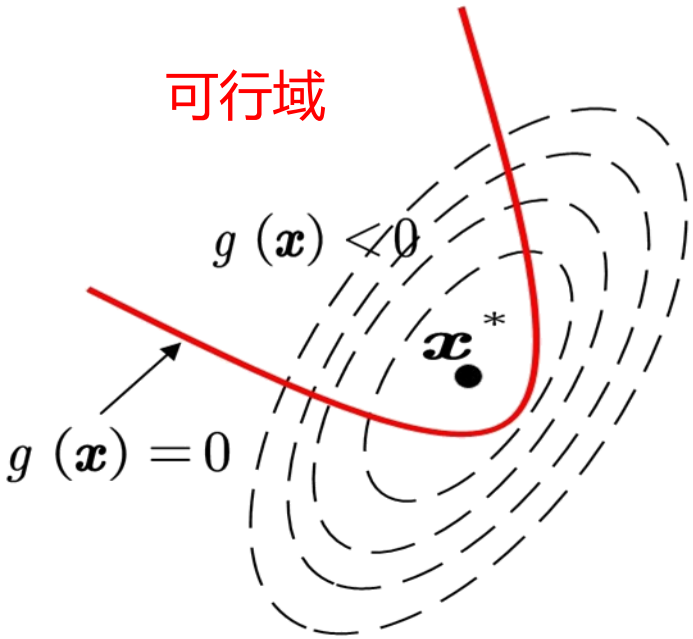
$$\begin{cases} \nabla f(\boldsymbol{x}^*) = 0 \\ g(\boldsymbol{x}^*) < 0 \end{cases}$$

- 这种情况相当于在拉格朗日函数中设置$\lambda = 0$：

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

- 从而满足驻点条件且$\lambda = 0$：

$$\frac{\partial L}{\partial \boldsymbol{x}} = 0 \Rightarrow \nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0$$



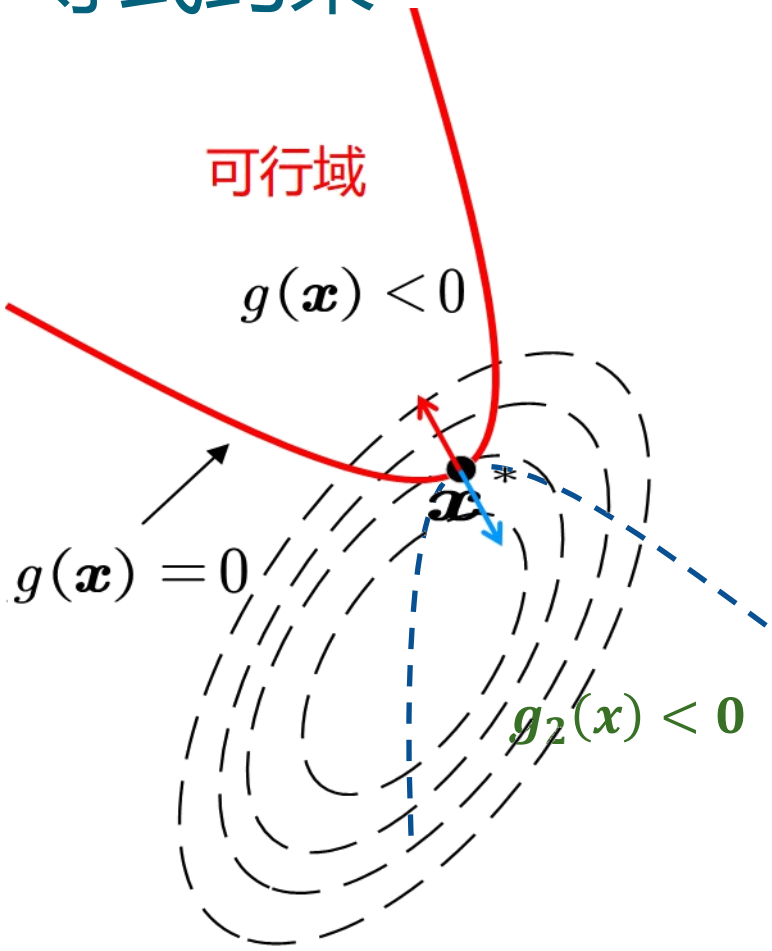
拉格朗日乘数法：不等式约束

情况2：极值点落在可行域边界

- 假设极值点落在了可行域边界，即极值点位于区域$g(\boldsymbol{x}) = 0$区域。
- 搜寻极值点$\boldsymbol{x}^*$：当出现沿着$g(\boldsymbol{x}) = 0$切线方向、无法再向$f(\boldsymbol{x})$负梯度方向移动时停止。在该点，$f(\boldsymbol{x})$等高线与$g(\boldsymbol{x}) = 0$相切，该点为$f(\boldsymbol{x})$的极值点$\boldsymbol{x}^*$。
- 对于不等式约束，在极值点$\boldsymbol{x}^*$，$f(\boldsymbol{x})$与$g(\boldsymbol{x})$的负梯度方向平行且相反。
- 梯度的幅值可能不同。
- 此种情况相当于存在一个$\lambda > 0$满足驻点条件：

$$\nabla f(\boldsymbol{x}^*) + \lambda \nabla g(\boldsymbol{x}^*) = 0$$

  - 注意：λ有符号限制。



如果在极值点$\boldsymbol{x}^*$，$f(\boldsymbol{x})$与$g(\boldsymbol{x})$的梯度方向相同，则说明在可行域内有更小的$f(\boldsymbol{x})$值存在。

拉格朗日乘数法：不等式约束

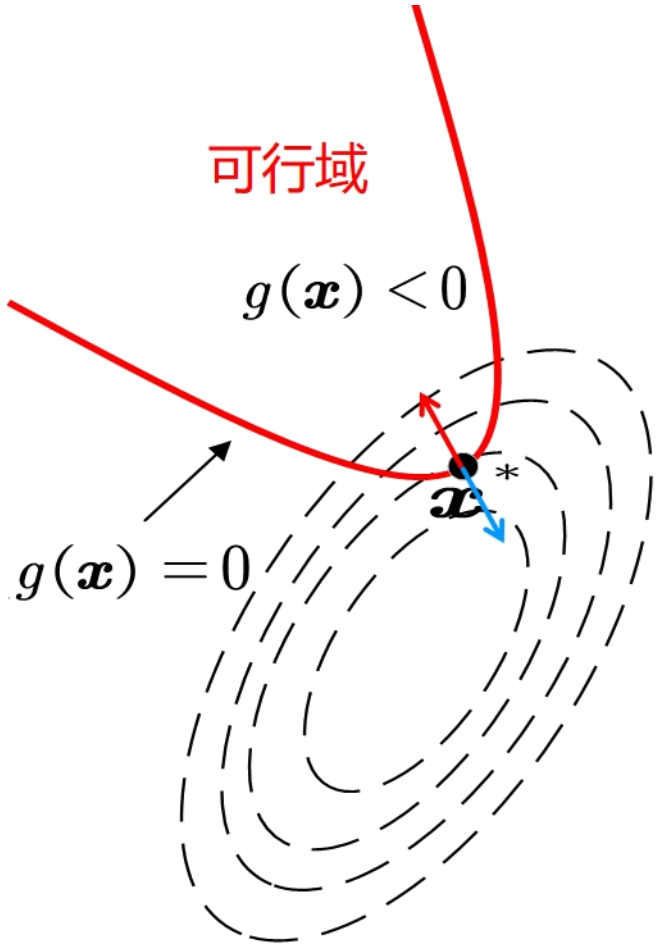
综合两种情况

- 无论是$g(\boldsymbol{x})<0$（$\lambda=0$）还是$g(\boldsymbol{x})=0$（$\lambda>0$）的约束情况，始终存在一个$\lambda\geq 0$（对偶可行性），满足：

$$\lambda g(\boldsymbol{x})=0$$

- 同时，始终存在一个$\lambda\geq 0$，满足驻点条件：

$$\nabla f(\boldsymbol{x}^*)+\lambda\nabla g(\boldsymbol{x}^*)=0$$



不等式约束：KKT条件

KKT条件&等价优化问题

- 在$g(\boldsymbol{x}) \leq 0$约束条件下最小化$f(\boldsymbol{x})$的问题，可以转化为如下约束条件（KKT条件）下的拉格朗日函数优化问题：

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x})$$
$$s.t. \quad g(\boldsymbol{x}) \leqslant 0$$

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

$$s.t. \begin{cases} g(\boldsymbol{x}) \leqslant 0 \\ \lambda \geqslant 0 \\ \lambda g(\boldsymbol{x}) = 0 \end{cases}$$

Karush-Kuhn-Tucker (KKT) 条件

拉格朗日乘数法：多个约束

KKT条件&等价优化问题

- 在多个约束条件下最小化$f(x)$的问题，转化为KKT条件下的拉格朗日函数优化问题：

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ s.t. \quad & h_j(\boldsymbol{x}) = 0 \\ & g_i(\boldsymbol{x}) \le 0 \end{aligned}$$

$$L(\boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \gamma_j h_j(\boldsymbol{x})$$

$$s.t. \begin{cases} g_i(\boldsymbol{x}) \leqslant 0 \\ \lambda_i \geqslant 0 \\ \lambda_i g_i(\boldsymbol{x}) = 0 \\ h_j(\boldsymbol{x}) = 0 \end{cases} \qquad \begin{aligned} \boldsymbol{\Lambda} &= [\lambda_1, \lambda_2, \ldots]^T \\ \boldsymbol{\Gamma} &= [\gamma_1, \gamma_2, \ldots]^T \end{aligned}$$

拉格朗日对偶问题

拉格朗日乘数法：主问题

主问题（Primal Problem）

- 根据原问题的约束条件 $h_j(\boldsymbol{x}) = 0$，对于γ_j的任意取值，拉格朗日函数第三项都可以消去。
- 根据原问题的约束条件$g_i(\boldsymbol{x}) \leqslant 0$，对于$\lambda_i$>0的任意取值，使得 $\lambda_i g_i(\boldsymbol{x}) \leqslant 0$
- 可见，拉格朗日函数$L(\boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma})$关于$\boldsymbol{\Lambda}, \boldsymbol{\Gamma}$的最大值就是$f(\boldsymbol{x})$。
- 主问题：带约束的原问题等价于如下（关于$\boldsymbol{x}$的）无约束问题。

$$\min_{\boldsymbol{x}} \max_{\lambda_i \geqslant 0, \gamma_j} L(\boldsymbol{x}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma})$$

注意： λ 的约束依然存在。

$$\lambda_i \geq 0$$

主问题如何求解?

针对不等式约束，主问题难以求解

$$\min_{\boldsymbol{x}} \max_{\lambda_i \geq 0, \gamma_j} f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \gamma_j h_j(\boldsymbol{x})$$

- λ_i求L偏导并设偏导等于0：

$$\frac{\partial L}{\partial \lambda_i} = g_i(\boldsymbol{x}) \qquad \frac{\partial L}{\partial \lambda_i} = 0 \Rightarrow g_i(\boldsymbol{x}) = 0$$

- 偏导中不含有λ_i，无法解析得到最优的λ_i^*
- 根据KKT条件，如果要求$g_i(x) = 0$，则$\lambda_i > 0$，但无法确定最优值。

$g_i(x) < 0$	$g_i(x) = 0$
$\lambda_i = 0$	$\lambda_i > 0$

主问题难以求解或者是NP难问题，如何解决?

拉格朗日对偶函数

- 取拉格朗日函数关于$\boldsymbol{x}$在其可行域内的最小值，记作L_D：

$$L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma}) = \min_{\boldsymbol{x}} L(\boldsymbol{x},\ \boldsymbol{\Lambda},\boldsymbol{\Gamma})$$

$$= \min_{\boldsymbol{x}} \left(f(\boldsymbol{x}) + \sum_i \lambda_i h_i(\boldsymbol{x}) + \sum_j \gamma_j g_j(\boldsymbol{x}) \right)$$

- 对偶函数L_D是关于$\boldsymbol{\Lambda} > 0$和$\boldsymbol{\Gamma}$的函数，与$\boldsymbol{x}$无关。
  - 对偶变量： $\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$。
  - 主变量：$\boldsymbol{x}$ 。

主问题最优值的下界

对偶函数是主问题的最优值下界

- 针对可行域R内的任意$\boldsymbol{x}$，对任意$\boldsymbol{\Lambda} > 0$和$\boldsymbol{\Gamma}$，都存在

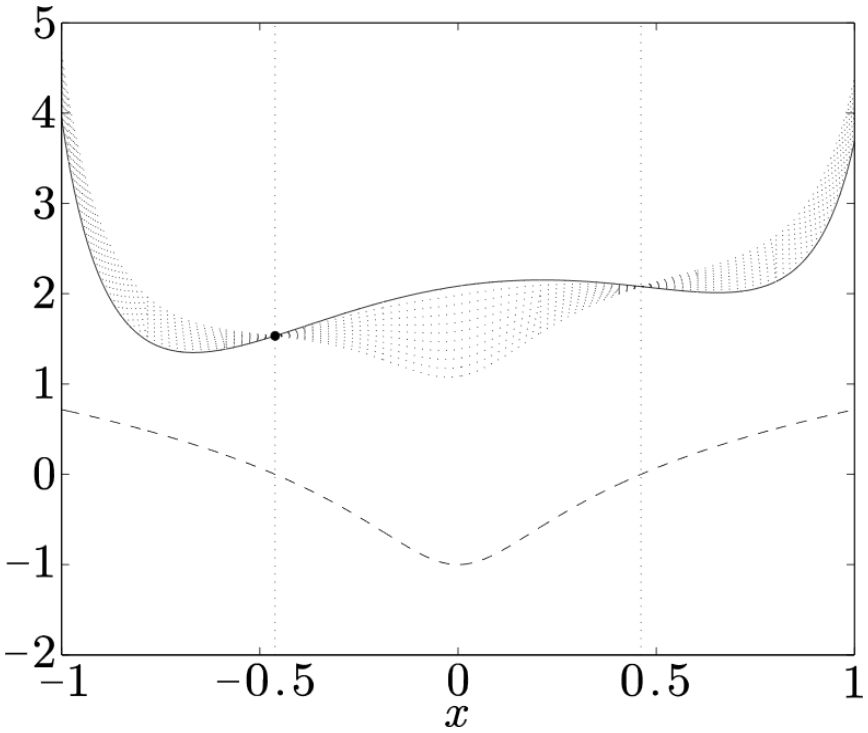
$$\sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \gamma_j h_j(\boldsymbol{x}) \leqslant 0$$

- 因此可以得到：$L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma}) \leqslant L(\boldsymbol{x},\ \boldsymbol{\Lambda},\boldsymbol{\Gamma}) \leqslant f(\boldsymbol{x})$
- 设主问题的最小值是$p^* = f(\boldsymbol{x}^*)$，则得到

$$L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma}) \leqslant L(\boldsymbol{x}^*,\boldsymbol{\Lambda},\boldsymbol{\Gamma}) \leqslant f(\boldsymbol{x}^*)$$

- 可见，对偶函数L_D给出了主问题最优值的下界。
- 该下界只跟对偶变量$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$有关，与$\boldsymbol{x}$无关。

主问题最优值的下界



- 实线：目标函数$f(x)$。
- 虚线：约束函数$g(x)$，x的可行域为[-0.46,0.46]，最优点为x^*=-0.46, p^*=1.54
- 点线：取不同λ值($\lambda = 0.1,0.2,\ldots,1.0$)的拉格朗日函数$L(x,\lambda)$。由于$L(x,\lambda) \leq f(x^*)$，每条曲线都有一个最小值小于$p^*$。

对偶问题

对偶问题（Dual Problem）

- 针对$\boldsymbol{\Lambda} > 0$和$\boldsymbol{\Gamma}$，最大化对偶函数L_D，得到主问题的对偶问题：

对偶问题

$$\max_{\boldsymbol{\Lambda},\boldsymbol{\Gamma}} L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma})$$

$$\max_{\boldsymbol{\Lambda},\boldsymbol{\Gamma}} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\Lambda},\boldsymbol{\Gamma})$$

$$\text{s.t. } \boldsymbol{\Lambda} \geq \mathbf{0}$$

主问题

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\Lambda},\boldsymbol{\Gamma}} L(\boldsymbol{x},\boldsymbol{\Lambda},\boldsymbol{\Gamma})$$

$$\text{s.t. } \boldsymbol{\Lambda} \geq \mathbf{0}$$

- 首先求取L关于$\boldsymbol{x}$的最小值（下界），再求取下界关于$\boldsymbol{\Lambda}, \boldsymbol{\Gamma}$的最大值。

对偶函数的凹凸性

对偶函数：分析

- 为什么要建立对偶问题?

$$L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma}) = \min_{\boldsymbol{x}} \left(f(\boldsymbol{x}) + \sum_i \lambda_i h_i(\boldsymbol{x}) + \sum_j \gamma_j g_j(\boldsymbol{x}) \right)$$

- 对偶函数是以$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$为自变量的、与$\boldsymbol{x}$无关。
- 因此，里面的拉格朗日函数L看做关于对偶变量$\boldsymbol{\Lambda}$和$\boldsymbol{\Gamma}$的仿射组合。
- 对偶函数L_D则是拉格朗日函数L的逐点（pointwise）最小值函数。

对偶函数的凹凸性

最小值函数的凹凸性

Given a min function $f(x) = \min_i x_i$, for $0 \leq \theta \leq 1$,

$$
\begin{aligned}
& f(\theta x + (1-\theta)y) \\
&= \min_i(\theta x_i + (1-\theta)y_i) \\
&\geq \theta \min_i x_i + (1-\theta)\min_i y_i \\
&= \theta f(x) + (1-\theta) f(y)
\end{aligned}
$$

- 可见，逐点（pointwise）最小值函数min是凹函数。

对偶函数的凹凸性

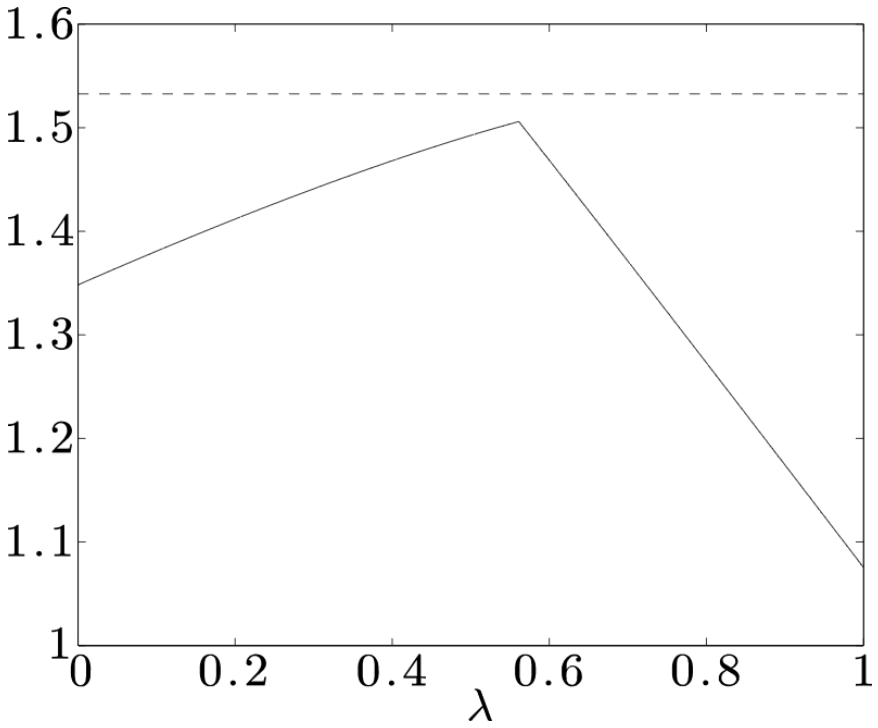
对偶函数是凹函数

Given $g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$, if the function f is concave, then g is concave.

$$L_D(\boldsymbol{\Lambda}, \boldsymbol{\Gamma}) = \min_{\boldsymbol{x}} \left(f(\boldsymbol{x}) + \sum_i \lambda_i h_i(\boldsymbol{x}) + \sum_j \gamma_j g_j(\boldsymbol{x}) \right)$$

- 由于逐点最大化函数是凹函数，且拉格朗日函数可以看做关于对偶变量的仿射组合，所以对偶函数L_D是凹函数。

对偶函数是凹函数



- 实线：对偶函数$L_D(\lambda)$是凹函数。由前图可知，$f(x)$和$g(x)$都是非凸的。
- 虚线：最优点p^*=1.54。

对偶问题：凸优化

对偶问题是凸还是非凸?

$$\max_{\boldsymbol{\Lambda},\boldsymbol{\Gamma}} L_D(\boldsymbol{\Lambda},\boldsymbol{\Gamma})$$

$$\text{s.t. } \boldsymbol{\Lambda} \geq \mathbf{0}$$

- 由于目标函数L_D是凹函数，约束条件是凸函数，所以对偶问题是凸优化问题。
- 无论主问题的凸性如何，对偶问题始终是一个凸优化问题。
- 凸优化的性质：局部极值点就是全局极值点。
- 所以，对偶问题的极值是唯一的全局极值点。
- 因此，对于难以求解的主问题（例如，非凸问题或者NP难问题），可以通过求解其对偶问题，得到原问题的一个下界估计。

对偶问题与主问题的差异?

弱对偶性（weak duality）

- 设对偶问题的最优值为d^*、主问题的最优值为p^*无论主问题的凸性如何，对偶问题始终是一个凸优化问题。
- 对于所有的优化问题都存在： $d^* \leq p^*$

强对偶性（strong duality）

- 强对偶性：$d^* = p^*$
- 如果强对偶性成立，则对偶问题获得主问题的最优下界。

强对偶性成立的条件（Slater条件）

$f(x)$是凸函数

$g_i(x)$是凸函数

$h_j(x)$是仿射函数

在可行域至少有一点使得不等式约束严格成立。

- 如果强对偶性成立，则对偶问题获得主问题的最优下界。

如何使用拉格朗日对偶法求解支持向量机呢?

支持向量机学习算法

支持向量机

线性判据

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

支持向量机目标函数

给定标记过的训练样本$\{(\boldsymbol{x}_n, t_n)\}_{n=1,\ldots,N}$：

$$\min_{\boldsymbol{w}} \ \frac{1}{2}||\boldsymbol{w}||_2^2$$

$$s.t. \ \ t_n(\boldsymbol{w}^T \boldsymbol{x}_n + w_0) - 1 \geqslant 0, \quad \forall n$$

目标函数如何求解

- 带不等式约束的优化问题，使用拉格朗日对偶法求解。

构建拉格朗日函数

拉格朗日函数

$$L(\boldsymbol{w}, w_0, \boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{n=1}^{N} \lambda_n [t_n(\boldsymbol{w}^T \boldsymbol{x}_n + w_0) - 1]$$

- 拉格朗日乘数向量： $\boldsymbol{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$
- KKT条件： $s.t. \begin{cases} -t_n(\boldsymbol{w}^T \boldsymbol{x}_n + w_0) + 1 \leq 0 \\ \lambda_n \geq 0 \\ \lambda_n [t_n(\boldsymbol{w}^T \boldsymbol{x}_n + w_0) - 1] = 0 \end{cases}$

构建对偶函数

对偶函数

$$L_D(\boldsymbol{\Lambda}) = \min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{\Lambda})$$

$$= \min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{n=1}^{N} \lambda_n [t_n(\boldsymbol{w}^T \boldsymbol{x}_n + w_0) - 1]$$

- L对参数w和w_0求导，并设偏导等于0：

$$\frac{\partial L(\boldsymbol{w},w_0\boldsymbol{\Lambda})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{n=1}^{N} t_n \lambda_n \boldsymbol{x}_n = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{n=1}^{N} t_n \lambda_n \boldsymbol{x}_n$$

$$\frac{\partial L(\boldsymbol{w},w_0\boldsymbol{\Lambda})}{\partial w_0} = \sum_{n=1}^{N} t_n \lambda_n = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} t_n \lambda_n = 0$$

构建对偶函数

在极值点计算$||w||_2^2$

$$L_D(\boldsymbol{\Lambda}) = \min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{n=1}^{N} \lambda_n[t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1]$$

$$\boldsymbol{w} = \sum_{n=1}^{N} t_n\lambda_n\boldsymbol{x}_n$$

$$\|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^T\boldsymbol{w} = (\sum_{n=1}^{N} t_n\lambda_n\boldsymbol{x}_n)^T(\sum_{n=1}^{N} t_n\lambda_n\boldsymbol{x}_n)$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M} t_n t_m \lambda_n\lambda_m \boldsymbol{x}_n^T\boldsymbol{x}_m$$

构建对偶函数

在极值点计算$||w||_2^2$

$$L_D(\boldsymbol{\Lambda}) = \min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{n=1}^{N} \lambda_n[t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1]$$

$$\boldsymbol{w} = \sum_{n=1}^{N} t_n\lambda_n\boldsymbol{x}_n$$

$$\sum_{n=1}^{N} t_n\lambda_n = 0$$

$$\sum_{n=1}^{N} \lambda_n[t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1]$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M} t_n t_m \lambda_n \lambda_m \boldsymbol{x}_n^T\boldsymbol{x}_m - \sum_{n=1}^{N} \lambda_n$$

构建对偶函数

在极值点得到L_D

$$L_D(\boldsymbol{\Lambda}) = \min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{n=1}^{N} \lambda_n[t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) - 1]$$

- 将上述偏导等于0的结果带入拉格朗日函数，消去$\boldsymbol{w}$和w_0：

$$L_D(\boldsymbol{\Lambda}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} t_n t_m \lambda_n \lambda_m \boldsymbol{x}_n^T \boldsymbol{x}_m + \sum_{n=1}^{N} \lambda_n$$

$$= -\frac{1}{2}\boldsymbol{\Lambda}^T \boldsymbol{H} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T \mathbf{1}$$

where $\boldsymbol{H}$ is Hessian matrix, $h_{nm} = t_n t_m \boldsymbol{x}_n^T \boldsymbol{x}_m$

构建对偶函数

对偶函数的约束条件

- 对偶函数是关于$\boldsymbol{\Lambda}$的函数，所以约束条件只需考虑$\boldsymbol{\Lambda}$的相关项。

- 对偶可行性：对于不等式约束，构建拉格朗日函数必须满足对偶可行性条件。

$$\boldsymbol{\Lambda} \geq 0$$

- 极值点上关于$\boldsymbol{\lambda}$的约束项：

$$\boldsymbol{\Lambda}^T \boldsymbol{t} = 0$$

$$\text{where } \boldsymbol{t} = [t_1, t_2, \ldots, t_N]^T$$

对偶问题

对偶问题

- 对偶函数是关于$\boldsymbol{\Lambda}$的函数，所以约束条件只需考虑$\boldsymbol{\Lambda}$的相关项。

$$L_D(\boldsymbol{\Lambda}) = -\frac{1}{2}\boldsymbol{\Lambda}^T \boldsymbol{H} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T \mathbf{1}$$

$$\text{s.t.} \quad \begin{cases} \boldsymbol{\Lambda} \geq 0 \\ \boldsymbol{\Lambda}^T \boldsymbol{t} = 0 \end{cases}$$

- 得到最优的$\boldsymbol{\Lambda}$。即可得到最优的参数$\boldsymbol{w}$和w_0。

对偶问题的求解

求解对偶问题

- 这是标准的关于λ的二次规划(quadratic programming)问题。
- 可以调用Matlab提供的quadprog函数来求解。

quadprog函数

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^T\boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^T\boldsymbol{x} + r \\ s.t. \quad & \boldsymbol{G}\boldsymbol{x} \leq \boldsymbol{h}, \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \end{aligned}$$

支持向量

求解支持向量

- 用二次规划求解得到最优的$\boldsymbol{\Lambda}^*$，包含N个最优的拉格朗日乘数。

$$\boldsymbol{\Lambda}^* = [\lambda_1^*, \lambda_2^*, \ldots, \lambda_N^*]^T$$

- 根据KKT条件可知：

$\lambda_n^* = 0$	$\lambda_n^* > 0$
$t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) > 1$	$t_n(\boldsymbol{w}^T\boldsymbol{x}_n + w_0) = 1$
该样本位于超平面Π_+和Π_-之外	该样本位于超平面Π_+或Π_-上
样本$\boldsymbol{x}_n$非支持向量	样本$\boldsymbol{x}_n$是支持向量

参数最优解：$\boldsymbol{w}$

w最优解

- 根据找到的支持向量$\boldsymbol{x}_n$以及对应的拉格朗日乘子λ_n^*构建$\boldsymbol{w}^*$：

$$\boldsymbol{w}^* = \sum_{n=1}^{N_s} t_n \lambda_n^* \boldsymbol{x}_n$$

- 其中，N_S表示支持向量的个数。

参数最优解：$\boldsymbol{w}$

w最优解

- 根据找到的支持向量$\boldsymbol{x}_n$以及对应的拉格朗日乘子λ_n^*构建$\boldsymbol{w}^*$：

$$\boldsymbol{w}^* = \sum_{n=1}^{N_s} t_n \lambda_n^* \boldsymbol{x}_n$$

- 其中，N_S表示支持向量的个数。

参数最优解：w_0

w_0最优解

- 根据支持向量机定义的约束条件，针对任意一个支持向量$\boldsymbol{x}_s$：

$$\lambda_s[t_s(\boldsymbol{w}^T\boldsymbol{x}_s + w_0) - 1] = 0$$

- 由此得到：$w_0^* = \frac{1}{t_s} - \boldsymbol{w}^{*T}\boldsymbol{x}_s$

- w_0通常由所有支持向量取均值得到：

$$w_0^* = \frac{1}{N_s}\sum_{n=1}^{N_s}(\frac{1}{t_n} - \boldsymbol{w}^{*T}\boldsymbol{x}_n)$$

支持向量机：决策过程

如何用于识别决策过程?

- 给定一个测试模式$\boldsymbol{x}_{test}$，支持向量机分类器可表达为：

$$y(\boldsymbol{x}_{test}) = sign\Big(\sum_{k=1}^{N_s} t_k \lambda_k \boldsymbol{x}_{test}^T \boldsymbol{x}_k + w_0^*\Big)$$

$$\begin{cases} \boldsymbol{x}_{test} \in C_1 & \text{if } y(\boldsymbol{x}_{test}) = 1 \\ \boldsymbol{x}_{test} \in C_2 & \text{if } y(\boldsymbol{x}_{test}) = -1 \end{cases}$$

- $\boldsymbol{w}$和w_0的学习过程实际上是从训练样本中选择一组支持向量，并将这些支持向量存储下来，用作线性分类器。