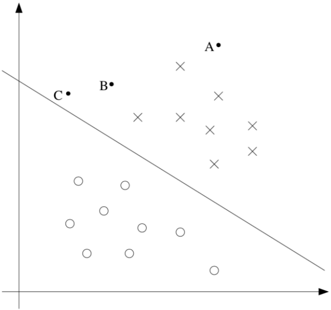
第十一章 凸优化与支持向量机

Margin

- 本章以支持向量机 (Support Vector Machine, SVM) 为例，介绍带限制条件的凸优化问题的一般解法
- Margin是 SVM 的一个重要概念，代表一种预测的 “信心”
  - Logistic 回归做预测时可以采用以下规则：如果 $P(Y=1 \mid \boldsymbol{x}, \boldsymbol{\theta}) \geq 0.5$，预测 $Y=1$，反之 $Y=0 \Leftrightarrow \boldsymbol{\theta}^\top \boldsymbol{x} \geq 0$，预测 $Y=1$，反之 $Y=0$
  - $\boldsymbol{\theta}^\top \boldsymbol{x}$ 越大，预测 $Y=1$ 越有信心；$\boldsymbol{\theta}^\top \boldsymbol{x}$ 越小，预测 $Y=0$ 越有信心



  - 当要预测的点越远离决策边界，对它的预测越有信心

Margin

在训练集 $S = \{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ 中，每个点 i 由一个特征向量 $\boldsymbol{x}_i$ 和一个标签 $y_i \in \{-1, 1\}$ 组成, 线性 SVM 分类器假设决策边界具有如下形式:

$$\boldsymbol{\omega}^\top \boldsymbol{x} + b = 0$$

- 决策规则为：$\boldsymbol{\omega}^\top \boldsymbol{x} + b \geq 0$, 预测 $y = 1$; 反之，预测 $y = -1$
- 如果将点 i 的 margin 定义为

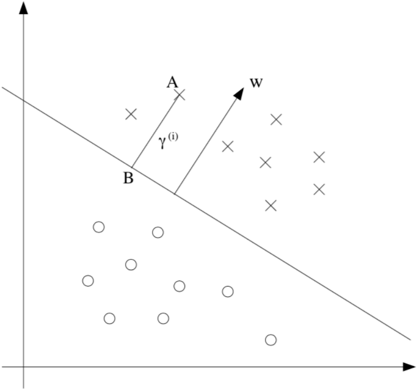
$$\gamma_i = y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) \tag{1}$$

  - $\gamma_i > 0$ 表明对点 i 的预测是正确的，同时较大的 γ_i 代表对预测值较大的信心
  - 如果将 ω 和 b 同时扩大 2 倍，决策边界不变，但是对预测的信心 γ_i 却扩大了 2 倍
- 为了保证 margin 可识别，需要对(1)中的系数加一些规范化条件, 比如令 $\|\boldsymbol{\omega}\|_2 = 1$ 或者令

$$\gamma_i = y_i\left(\frac{\boldsymbol{\omega}^\top \boldsymbol{x}_i + b}{\|\boldsymbol{\omega}\|_2}\right) \tag{2}$$

Margin

- Margin 的**几何意义**



由(2)定义的 margin γ_i 等于点 i 到决策边界的距离

Margin

假设训练集线性可分，SVM 希望训练集中的所有点都远离决策边界，令

$$\gamma = \min_{i} \ \gamma_i$$

SVM 的目标是寻找一条决策边界使最小的 margin γ 最大:

$$\max_{\boldsymbol{\omega},b} \ \gamma \ \ \text{s.t.} \ \ y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b)/\left\|\boldsymbol{\omega}\right\| \geq \gamma, \ \ i = 1,\ldots,n \tag{3}$$

在(3)中令 $\|\boldsymbol{\omega}\| = 1/\gamma$，则最大化最小 margin 的问题(3)可以转化为如下非常容易求解的优化问题:

$$\min_{\boldsymbol{\omega},b} \ \frac{1}{2}\left\|\boldsymbol{\omega}\right\|^2 \ \ \text{s.t.} \ \ -y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) + 1 \leq 0, \ \ i = 1,\ldots,n \tag{4}$$

- (4)可以用二次规划 (QP) 算法求解
- 在很多实际问题中，特征 $\mathbf{x}_i \in \mathbb{R}^d$ 是一个高维向量 ($d \gg n$), 如果将(4)转化为**拉格朗日对偶形式**求解, 会比直接使用 QP 更高效

凸优化理论

一般的凸优化问题：

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^d} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}) \le 0,\ i = 1,\ldots,m \\ & h_j(\boldsymbol{x}) = 0,\ j = 1,\ldots,p. \end{aligned} \tag{5}$$

其中函数 $f: \mathbb{R}^d \to \mathbb{R}$ 和 $g_i: \mathbb{R}^d \to \mathbb{R}$, $i = 1,\ldots,m$ 都是可导的凸函数，函数 $h_j: \mathbb{R}^d \to \mathbb{R}$, $j = 1,\ldots,p$ 都是仿射函数

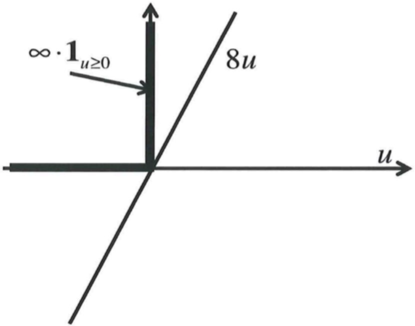
- (5)可以写为以下等价的无限制优化问题：

$$\min_{\boldsymbol{x}} \Theta_P(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) + \infty \sum_{i=1}^{m} \mathbf{1}\left(g_i(\boldsymbol{x}) > 0\right) + \infty \sum_{j=1}^{p} \mathbf{1}\left(h_j(\boldsymbol{x}) \neq 0\right) \tag{6}$$

称(6)为**原始优化问题 (primal optimization)**

凸优化理论

- (6)很难求解，考虑用某种可导函数替换惩罚函数 $\infty \cdot \mathbf{1}(u > 0)$，比如线性函数 αu. 当 $\alpha \geq 0$ 时，函数 αu 是 $\infty \cdot \mathbf{1}(u > 0)$ 的一个下界函数



类似地，函数 βu 总是 $\infty \cdot \mathbf{1}(u \neq 0)$ 的一个下界函数

凸优化理论

- 定义**拉格朗日函数 (Lagrangian)**：

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \beta_j h_j(\boldsymbol{x})$$

  其中 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ 和 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ 的元素称为**拉格朗日乘子 (Lagrange multipliers)**
- 可以证明

$$\Theta_P(\boldsymbol{x}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \ \text{ s.t. } \alpha_i \geq 0, \forall i$$

- 因此原始优化问题 (6)可以转化为以下目标函数可导的优化问题：

$$\min_{\boldsymbol{x}} \Theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \left[\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_i \geq 0, \forall i} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right] \tag{7}$$

- 如果点 $\boldsymbol{x}$ 满足所有限制条件，即 $g_i(\boldsymbol{x}) \leq 0, \forall i$ 且 $h_j(\boldsymbol{x}) = 0, \forall j$, 称点 $\boldsymbol{x}$ 为**原始可行的 (primal feasible)**
- 假设 $\Theta_P(\boldsymbol{x})$ 在 $\boldsymbol{x}^\star$ 处达到最小，最小值记为 $p^\star = \Theta_P(\boldsymbol{x}^\star)$

凸优化理论

- 交换(7)中 $\min$ 和 $\max$ 的顺序, 就得到了另一个不同的优化问题，称为(7)的**对偶问题 (dual problem)**：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i\geq 0,\forall i}\left[\min_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})\right]=\max_{\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_i\geq 0,\forall i}\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) \tag{8}$$

此处定义**对偶目标函数 (dual objective)** $\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})=\min_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$

- 如果点 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 满足 $\alpha_i\geq 0,\ \forall i$, 称点 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 为**对偶可行的 (dual feasible)**
- 假设 $\Theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})$ 在 $(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$ 处达到最大，最大值记为 $d^\star=\Theta_D(\boldsymbol{\alpha}^\star,\boldsymbol{\beta}^\star)$

定理

对任意一对原始和对偶问题 (7)和(8), 总有 $d^\star\leq p^\star$.

- 如果原始和对偶问题满足 $d^\star=p^\star$，称为**强对偶性 (strong duality)**
- 很多条件可以保证强对偶性成立，最常用的是 **Slater's condition**: 即优化问题(5)的解 $\boldsymbol{x}^\star$ 使所有不等式限制条件都严格成立，$g_i(\boldsymbol{x}^\star)<0,\ \forall i$

KKT 条件

- 对于带限制的优化问题(5), 找到满足 KKT 条件的解等价于找到全局最小值点 (global minimum)

引理 (互补松弛性 (Complementary Slackness))

如果强对偶性成立，那么 $\alpha_i^\star g_i(\boldsymbol{x}^\star) = 0,\ i = 1, \ldots, m.$

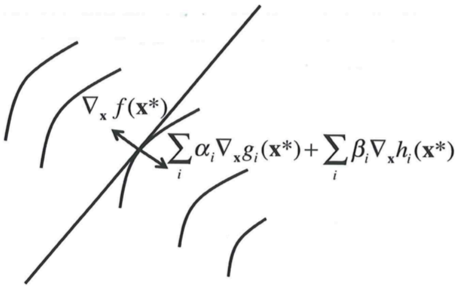
- 当强对偶性成立时，在原始/对偶问题的最优解 $(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 处有以下结论成立:
  - 如果某个 $\alpha_i^\star > 0$，则对应的 $g_i(\boldsymbol{x}^\star) = 0$，此时称该限制条件 g_i 为 active constraint 或 binding constraint
  - 如果某个 $g_i(\boldsymbol{x}^\star) < 0$, 则对应的 $\alpha_i^\star = 0$

KKT 条件

- 当强对偶性成立时，由上述引理的证明可得，$\boldsymbol{x}^\star$ 是凸函数 $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 的最小值点，因此满足梯度为零：

$$\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^\star) + \sum_{i=1}^{m} \alpha_i^\star \nabla_{\boldsymbol{x}} g_i(\boldsymbol{x}^\star) + \sum_{j=1}^{p} \beta_j^\star \nabla_{\boldsymbol{x}} h_j(\boldsymbol{x}^\star) = \mathbf{0} \quad (9)$$

  称等式(9)为**拉格朗日不动性 (Lagrangian stationarity)**
- (9)表明在最优解 $x^\star$ 处，目标函数 f 的梯度和限制函数的梯度方向相反，模长相等



KKT 条件

定理 (KKT 条件)

如果点 $\boldsymbol{x}^\star \in \mathbb{R}^d$, $\boldsymbol{\alpha}^\star \in \mathbb{R}^m$, $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ 满足以下条件:

- *(原始可行性)* $g_i(\boldsymbol{x}^\star) \leq 0,\ i = 1, \ldots, m$ 且 $h_j(\boldsymbol{x}^\star) = 0,\ j = 1, \ldots, p$
- *(对偶可行性)* $\alpha_i^\star \geq 0,\ i = 1, \ldots, m$
- *(互补松弛性)* $\alpha_i^\star g_i(\boldsymbol{x}^\star) = 0,\ i = 1, \ldots, m$
- *(拉格朗日不动性)* $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \mathbf{0}$.

则 $\boldsymbol{x}^\star$ 是原始问题最优解, $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 是对偶问题最优解. 如果强对偶性成立，则任何原始问题最优解 $\boldsymbol{x}^\star$ 及任何对偶问题最优解 $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ 必须满足以上条件.

- 如果强对偶性不成立，KKT 条件是找到优化问题(5)全局最优解的充分条件
- 如果强对偶性成立，KKT 条件是找到(5)全局最优解的充要条件

SVM: 最大化最小 margin

回到线性 SVM 分类模型，最佳决策边界是以下带限制的凸优化问题的解：

$$\begin{aligned} &\min_{\boldsymbol{\omega},b} \ \frac{1}{2}\|\boldsymbol{\omega}\|^2 \\ &\text{s.t.} \ -y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) + 1 \le 0, \ i = 1, \dots, n. \end{aligned} \tag{10}$$

下面列出(10)的最优解需要满足的 KKT 条件：

- 拉格朗日不动性. (10)的拉格朗日函数为

$$\mathcal{L}([\boldsymbol{\omega}, b], \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{d}\omega_j^2 + \sum_{i=1}^{n}\alpha_i\left[-y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) + 1\right]$$

  计算 $\mathcal{L}$ 关于 ω 和 b 的梯度并令其等于零：

$$\nabla_{\boldsymbol{\omega}}\mathcal{L} = \boldsymbol{\omega} - \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i = \mathbf{0} \implies \boldsymbol{\omega}^\star = \sum_{i=1}^{n}\alpha_i^\star y_i \boldsymbol{x}_i \tag{11}$$

$$\nabla_b\mathcal{L} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \implies \sum_{i=1}^{n}\alpha_i^\star y_i = 0 \tag{12}$$

SVM: 最大化最小 margin

- 对偶可行性: $\alpha_i^\star \geq 0, \forall i$
- 原始可行性: $-y_i(\boldsymbol{\omega}^{\star\top}\mathbf{x}_i + b^\star) + 1 \leq 0, \forall i$
- 互补松弛性: $\alpha_i^\star\left[-y_i(\boldsymbol{\omega}^{\star\top}\mathbf{x}_i + b^\star) + 1\right] = 0, \forall i$

将(11)和(12)代入拉格朗日函数，得到对偶目标函数在 $\boldsymbol{\alpha}^\star$ 处的值:

$$\Theta_D(\boldsymbol{\alpha}^\star) = \mathcal{L}([\boldsymbol{\omega}^\star, b^\star], \boldsymbol{\alpha}^\star) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i^\star\alpha_k^\star y_i y_k \mathbf{x}_i^\top \mathbf{x}_k + \sum_{i=1}^{n}\alpha_i^\star$$

SVM: 最大化最小 margin

考虑到 $\boldsymbol{\alpha}^\star$ 还需满足条件(12)和对偶可行性，$\boldsymbol{\alpha}^\star$ 是以下对偶优化问题的解：

$$
\begin{aligned}
&\max_{\boldsymbol{\alpha}} \Theta_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^\top \boldsymbol{x}_k \\
&\text{s.t. } \alpha_i \geq 0, \; i = 1, \ldots, n \\
&\qquad \sum_{i=1}^{n} \alpha_i y_i = 0
\end{aligned} \tag{13}
$$

- 当特征 $\boldsymbol{x}_i$ 的维数 $d \gg n$ 时，与(10)相比，(13)仅对应一个 n 维凸优化，此时可以使用 QP 算法求解，或者专门为 SVM 设计的 SMO 算法
- 假设已经解出 $\boldsymbol{\alpha}^\star$，由(11)可以得到原始问题最优解：

$$
\boldsymbol{\omega}^\star = \sum_{i=1}^{n} \alpha_i^\star y_i \boldsymbol{x}_i
$$

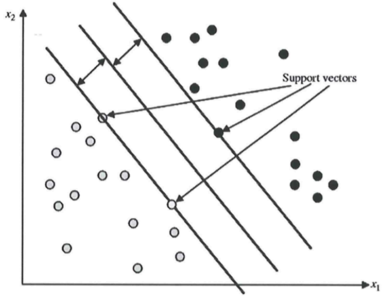
但是仍然不知道 $b^\star$ 的取值, 注意 KKT 条件中的 “原始可行性” 和 “互补松弛性” 条件还没有用到

支持向量

由互补松弛性条件 $\alpha_i^\star\left[-y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i+b^\star)+1\right]=0,\ \forall i$，可得:

$$\alpha_i^\star>0\Rightarrow y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i+b^\star)=1$$

即在 $\alpha_i^\star>0$ 对应的的点 $(\boldsymbol{x}_i,y_i)$ 处，不等式限制条件以等式成立，说明该点到决策边界的距离最小 (为 $\gamma=1/\|\boldsymbol{\omega}^\star\|$)，训练集中这样的点 $(\boldsymbol{x}_i,y_i)$ 被称为支持向量 (support vectors)，它们是最靠近决策边界的点



- 因此可以从解出的 $\boldsymbol{\alpha}^\star$ 中找到 $\alpha_i^\star>0$ 对应的支持向量, 再从任一支持向量 $(\boldsymbol{x}_i,y_i)$ 处利用等式 $y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i+b^\star)=1$ 计算出 $b^\star$

线性不可分情形

很多实际问题不存在线性决策边界 (超平面) 将训练集中的正负点区分开

因此需要对 SVM 模型(10)做一些修改以适用线性不可分情形 (nonseparable case)，修改后的模型将允许一些分类错误，但需要为错误付出一定代价

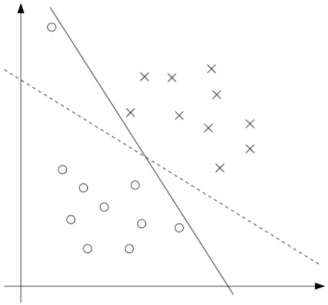
线性不可分情形

修改后的 SVM 求解的优化问题变为:

$$\begin{aligned} \min_{\omega,b,\xi}\ & \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.}\ & y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) \geq 1-\xi_i \\ & \xi_i \geq 0,\ i=1,\ldots,n \end{aligned} \tag{14}$$

- (14)在限制条件中加入了一些 “松弛” (slack) ξ_i
  - 如果观察点 i 满足 $y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) \geq 1$, 令 $\xi_i = 0$ 可以避免惩罚
  - 如果观察点 i 出现 $y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) = 1-\xi_i$ 且 $\xi_i > 0$，则需要付出代价 $C\xi_i$
- 参数 C 代表对实现以下两个目标的权衡: (i) 保证训练集中大部分观察点被正确分类 (ii) 使支持向量的 margin $\gamma = 1/\|\boldsymbol{\omega}\|$ 尽可能大

线性可分情形



- C 较大: 所有点都被正确分类，但支持向量的 margins 很小
- C 较小: 可以减小决策边界对异常点 (outliers) 的敏感性，通过付出一些分类错误的代价保证大多数点的 margins 较大

线性不可分情形

下面通过 KKT 条件求解(14)

- 建立拉格朗日函数

$$\mathcal{L}([\boldsymbol{\omega}, b, \boldsymbol{\xi}], \boldsymbol{\alpha}, \boldsymbol{r}) = \frac{1}{2}\boldsymbol{\omega}^\top \boldsymbol{\omega} + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \left[-y_i(\boldsymbol{\omega}^\top \boldsymbol{x}_i + b) + 1 - \xi_i\right] + \sum_{i=1}^{n} r_i(-\xi_i) \tag{15}$$

- 令 $\mathcal{L}$ 关于 ξ_i 的一阶偏导数等于 0 得

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - r_i = 0 \implies \alpha_i^\star = C - r_i^\star \tag{16}$$

  因此 $0 \le \alpha_i^\star \le C,\ i = 1, \dots, n$

- $\mathcal{L}$ 关于 ω 和 b 的梯度与(11) (12)相同，令其梯度等于零再代入 (15), 经过整理可得 $\boldsymbol{\alpha}^\star$ 是以下对偶问题的解：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \ & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^\top \boldsymbol{x}_k \\ \text{s.t. } & 0 \le \alpha_i \le C,\ \ i = 1, \dots, n \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{17}$$

线性不可分情形

- (17)与(13)的唯一区别是 α_i 的范围从 $\alpha_i \geq 0$ 变为 $0 \leq \alpha_i \leq C$

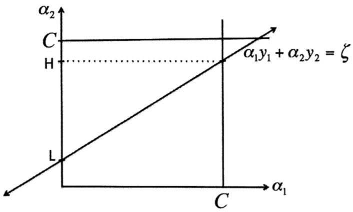
- 此时截距项 $b^\star$ 的计算与之前的方法不同，由互补松弛性条件可得：

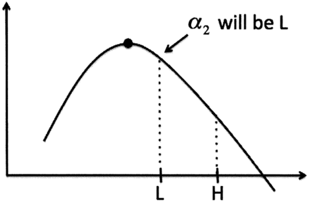
$$
\begin{aligned}
y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i + b^\star) > 1 \;&\Rightarrow\; \alpha_i^\star = 0 \\
y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i + b^\star) < 1 \;&\Rightarrow\; \xi_i^\star > 0 \;\Rightarrow\; r_i^\star = 0 \;\Rightarrow\; \alpha_i^\star = C \\
0 < r_i^\star < C \;&\Rightarrow\; 0 < \alpha_i^\star < C,\ \xi_i = 0 \;\Rightarrow\; y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i + b^\star) = 1
\end{aligned}
$$

所以只需找到 $0 < \alpha_i^\star < C$ 对应的观察点 $(\boldsymbol{x}_i, y_i)$，然后利用等式 $y_i(\boldsymbol{\omega}^{\star\top}\boldsymbol{x}_i + b^\star) = 1$ 解出 $b^\star$

SMO 算法

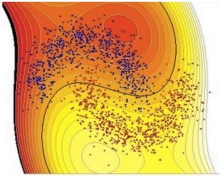
- Sequential Minimal Optimization (SMO) 是为求解 SVM 优化问题(17)设计的一个非常高效的算法，本质上是一种坐标下降算法
- 假设有一组 $\alpha_1, \ldots, \alpha_n$ 满足(17)中所有限制条件，如果固定 $\alpha_2, \ldots, \alpha_n$，通过调整 α_1 能使(17)的目标函数值上升吗?
  - 不能. 由(17)的限制条件 $\sum_{i=1}^n \alpha_i y_i = 0$ 可得当 $\alpha_2, \ldots, \alpha_n$ 固定时，α_1 也被固定了
- 考虑同时更新 $\boldsymbol{\alpha}$ 中的 2 个元素，比如固定 $\alpha_3, \ldots, \alpha_n$，如何调整 α_1, α_2 使(17)的目标函数值上升?
  - 首先由限制条件可得 (α_1, α_2) 只能位于正方形 $[0, C] \times [0, C]$ 内的一条线段上
  - 使用 α_2 表示 α_1，则目标函数可写为 α_2 的二次函数，易得 α_2 在区间 $[L, H]$ 上的最优解

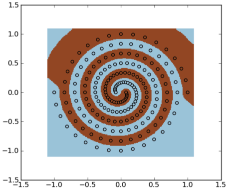




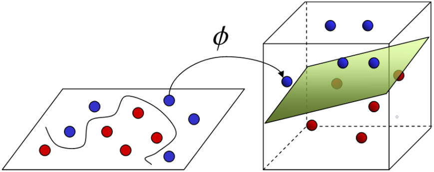
核函数

- SVM 与核函数 (kernels) 结合可以产生非常灵活的非线性决策边界或超曲面





- 当在 $\boldsymbol{x}$ 的特征空间 (feature space) 无法用线性决策边界将正负点区分时，一个解决办法是将 $\boldsymbol{x}$ 的特征空间升维到 $\phi(\boldsymbol{x})$ 所在的高维特征空间，使得在这个高维空间可以用线性超平面将正负点区分开，该超平面在原特征空间的投影是一条可区分正负点的曲线边界



核函数

SVM 的优化问题可以转化为求解以下对偶问题:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k y_i y_k \boldsymbol{x}_i^\top \boldsymbol{x}_k \\
\text{s.t.} \quad & 0 \le \alpha_i \le C, \; i = 1, \ldots, n \\
& \sum_{i=1}^{n} \alpha_i y_i = 0.
\end{aligned} \tag{18}
$$

- 注意到(18)只用到了特征的内积 $\boldsymbol{x}_i^\top \boldsymbol{x}_k$, 因此只需将(18)中 $\boldsymbol{x}_i^\top \boldsymbol{x}_k$ 替换为 $\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_k)$, 就得到了 $\phi(\boldsymbol{x})$ 空间的对偶问题
- 对每一个映射 ϕ, 定义它对应的核函数为:

$$
K_\phi(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{z})
$$

核函数

- 很多时候计算核函数的成本很小，但计算 $\phi(\boldsymbol{x})$ 的成本却很高
  - 例如 $\boldsymbol{x} \in \mathbb{R}^d$，令 $\phi(\boldsymbol{x}) = \left(x_1^2, x_1x_2, \ldots, x_1x_d, \ldots, x_dx_1, \ldots, x_d^2\right)^\top$，它的核函数为 $K_\phi(\boldsymbol{x}, \boldsymbol{z}) = \left(\boldsymbol{x}^\top \boldsymbol{z}\right)^2$
  - $\phi(\boldsymbol{x})$ 的计算量为 $O(d^2)$, 而 $K_\phi(\boldsymbol{x}, \boldsymbol{z})$ 的计算量只有 $O(d)$
- 如果核函数的形式为 $K_\phi(\boldsymbol{x}, \boldsymbol{z}) = \left(\boldsymbol{x}^\top \boldsymbol{z}\right)^r$，称其为**多项式核函数**, 它对应的 $\phi(\boldsymbol{x})$ 中的每个元素都是一个 r 次多项式 $x_{i_1}x_{i_2}\ldots x_{i_r}$，此时 $\phi(\boldsymbol{x})$ 的计算量为 $O(d^r)$，而 $K_\phi(\boldsymbol{x}, \boldsymbol{z})$ 的计算量仍为 $O(d)$.
- 从计算的角度，如果只需要知道 $K_\phi(\boldsymbol{x}, \boldsymbol{z})$ 的值，不一定要先计算出 $\phi(\boldsymbol{x})$ 和 $\phi(\boldsymbol{z})$

核函数

如果不计算 $\phi(\cdot)$ 在任意一点的值，对于一个测试点 $\boldsymbol{z}$, 如何用 $\phi(\boldsymbol{x}_i)$ 所在空间的决策超平面预测其正负?

- 若 SVM 在 $\boldsymbol{x}$ 的特征空间的最优线性决策边界为

$$\boldsymbol{\omega}^{\star\top}\boldsymbol{x} + b^{\star} = 0$$

  当测试点 $\boldsymbol{z}$ 满足 $\boldsymbol{\omega}^{\star\top}\boldsymbol{z} + b^{\star} \geq 0$, 预测其为正，反之为负
- 由拉格朗日不动性条件 (11)可得 $\boldsymbol{\omega}^{\star} = \sum_{i=1}^{n} \alpha_i^{\star} y_i \boldsymbol{x}_i$, 则最优决策边界可写为:

$$\sum_{i=1}^{n} \alpha_i^{\star} y_i \boldsymbol{x}_i^{\top}\boldsymbol{x} + b^{\star} = 0 \tag{19}$$

- 注意到(19)只用到点的内积 $\boldsymbol{x}_i^{\top}\boldsymbol{x}$，因此在 $\phi(\boldsymbol{x})$ 的空间中，最优决策超平面应具有以下形式:

$$\sum_{i=1}^{n} \alpha_i^{\star} y_i K_{\phi}(\boldsymbol{x}_i, \boldsymbol{x}) + b^{\star} = 0 \tag{20}$$

核函数

- (20)中 $b^\star$ 可以从某个 $0 < \alpha_j^\star < C$ 对应的观察点 $(\boldsymbol{\phi}(\boldsymbol{x}_j), y_j)$ 处计算得到:

$$b^\star = y_j - \boldsymbol{\omega}^{\star\top}\boldsymbol{\phi}(\boldsymbol{x}_j) = y_j - \sum_{i=1}^{n} \alpha_i^\star y_i \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j) = y_j - \sum_{i=1}^{n} \alpha_i^\star y_i K_\phi(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- (20)在 $\boldsymbol{x}$ 所在的空间一般对应一条曲线或曲面
  - 如果在(20)中使用多项式核函数，在 $\boldsymbol{x}$ 的空间就得到一条多项式决策边界
- 对于测试点 $\boldsymbol{z}$，如果 $\sum_{i=1}^{n} \alpha_i^\star y_i K_\phi(\boldsymbol{x}_i, \boldsymbol{z}) + b^\star \geq 0$, 预测其为正，反之为负
- 以上分析表明，不需要知道映射 $\boldsymbol{\phi}(\cdot)$ 的具体形式，只需要定义一个核函数 $K(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 就可以得到 SVM 的决策边界

核函数

如何证明确实存在一个映射 $\phi(\cdot)$ 使得 $K(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{z})$?

- 首先考察核函数需要具备的必要条件
  - $K(\cdot, \cdot)$ 需要满足对称关系 $K(\boldsymbol{x}, \boldsymbol{z}) = K(\boldsymbol{z}, \boldsymbol{x})$
  - 对 $\mathbb{R}^d$ 上的任意 n 个点 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, 定义矩阵

$$\boldsymbol{K} = (\boldsymbol{K}_{ij})_{n\times n} \tag{21}$$

其中 $\boldsymbol{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. 此时对 $\forall \boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{z}^\top \boldsymbol{K} \boldsymbol{z} \geq 0$, 因此矩阵 $\boldsymbol{K}$ 是一个半正定矩阵

定理 (Mercer)

函数 $K(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ 是一个有效核函数的充分必要条件是：对 $\mathbb{R}^d$ 上的任意有限个点 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, 由(21)定义的矩阵 $\boldsymbol{K}$ 是一个对称半正定矩阵.

核函数

- SVM 中一个常用的核函数是**高斯核函数 (Gaussian kernel)**:

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}\right) \tag{22}$$

  - (22)反映了点 $\boldsymbol{x}$ 和 $\boldsymbol{z}$ 的相似度
  - (22)对应的映射 $\phi(\cdot)$ 将原特征映射到一个无穷维空间

- 核函数的应用不仅限于 SVM，只要一个算法仅用到特征的内积 $\boldsymbol{x}^\top \boldsymbol{z}$，就可以将其替换为一个核函数 $K(\boldsymbol{x}, \boldsymbol{z})$，从而能在更高维的空间继续使用该算法，这个方法被称为**核函数技巧 (kernel trick)**