

第十一章 凸优化与支持向量机

Margin

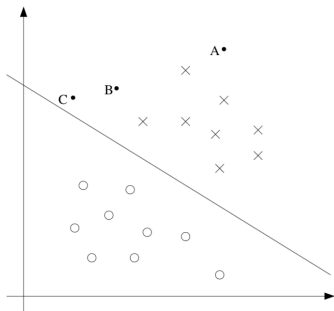
- 本章以支持向量机 (Support Vector Machine, SVM) 为例, 介绍带限制条件的凸优化问题的一般解法

Margin

- 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 $P(Y = 1 | \mathbf{x}, \boldsymbol{\theta}) \geq 0.5$, 预测 $Y = 1$, 反之 $Y = 0 \Leftrightarrow \boldsymbol{\theta}^\top \mathbf{x} \geq 0$, 预测 $Y = 1$, 反之 $Y = 0$
 - ▶ $\boldsymbol{\theta}^\top \mathbf{x}$ 越大, 预测 $Y = 1$ 越有信心; $\boldsymbol{\theta}^\top \mathbf{x}$ 越小, 预测 $Y = 0$ 越有信心

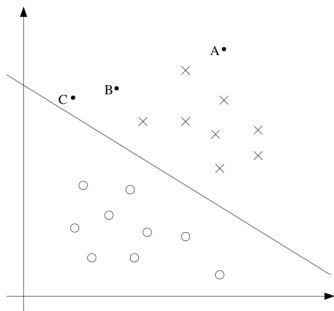
Margin

- 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 $P(Y=1 | \mathbf{x}, \boldsymbol{\theta}) \geq 0.5$, 预测 $Y=1$, 反之 $Y=0 \Leftrightarrow \boldsymbol{\theta}^\top \mathbf{x} \geq 0$, 预测 $Y=1$, 反之 $Y=0$
 - ▶ $\boldsymbol{\theta}^\top \mathbf{x}$ 越大, 预测 $Y=1$ 越有信心; $\boldsymbol{\theta}^\top \mathbf{x}$ 越小, 预测 $Y=0$ 越有信心



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 $P(Y=1 | \mathbf{x}, \boldsymbol{\theta}) \geq 0.5$, 预测 $Y=1$, 反之 $Y=0 \Leftrightarrow \boldsymbol{\theta}^\top \mathbf{x} \geq 0$, 预测 $Y=1$, 反之 $Y=0$
 - ▶ $\boldsymbol{\theta}^\top \mathbf{x}$ 越大, 预测 $Y=1$ 越有信心; $\boldsymbol{\theta}^\top \mathbf{x}$ 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 $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ 中, 每个点 i 由一个特征向量 \mathbf{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:

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

- 决策规则为: $\omega^\top \mathbf{x} + b \geq 0$, 预测 $y = 1$; 反之, 预测 $y = -1$

Margin

在训练集 $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ 中, 每个点 i 由一个特征向量 \mathbf{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:

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

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

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

- $\gamma_i > 0$ 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 $\boldsymbol{\omega}$ 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

Margin

在训练集 $S = \{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ 中, 每个点 i 由一个特征向量 \mathbf{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:

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

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

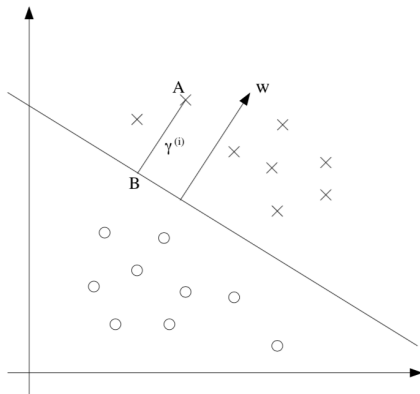
$$\gamma_i = y_i(\boldsymbol{\omega}^\top \mathbf{x}_i + b) \quad (1)$$

- $\gamma_i > 0$ 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 $\boldsymbol{\omega}$ 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

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

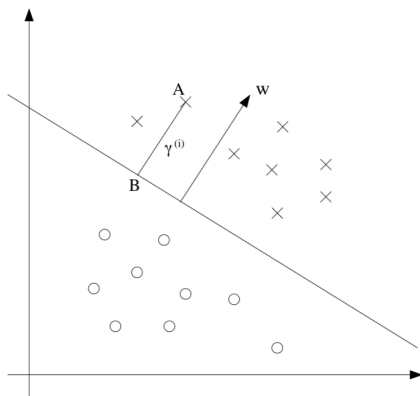
Margin

- Margin 的几何意义



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:

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

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:

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

在(3)中令 $\|\omega\| = 1/\gamma$, 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:

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

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:

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

在(3)中令 $\|\omega\| = 1/\gamma$, 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:

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

- (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:

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

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

凸优化理论

一般的凸优化问题:

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

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

- (5)可以写为以下等价的无限制优化问题:

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

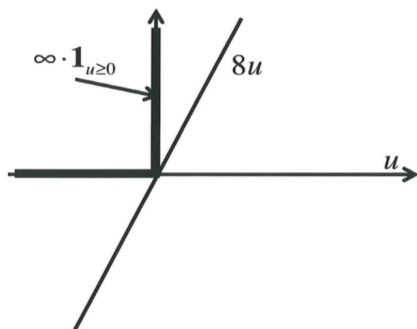
称(6)为**原始优化问题 (primal optimization)**

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

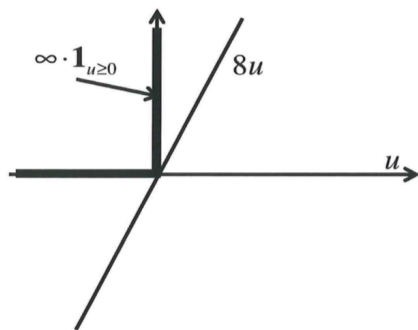
凸优化理论

- (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



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)**:

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

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

凸优化理论

- 定义**拉格朗日函数 (Lagrangian)**:

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

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

- 可以证明

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

凸优化理论

- 定义**拉格朗日函数 (Lagrangian)**:

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

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

- 可以证明

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

- 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:

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

凸优化理论

- 定义**拉格朗日函数 (Lagrangian)**:

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

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

- 可以证明

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

- 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:

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

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

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)**:

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

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

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)**:

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

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

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

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)**:

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

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

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

定理

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

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)**:

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

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

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

定理

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

- 如果原始和对偶问题满足 $d^* = p^*$, 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 \mathbf{x}^* 使所有不等式限制条件都严格成立, $g_i(\mathbf{x}^*) < 0, \forall i$

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)

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 $\alpha_i^* g_i(\mathbf{x}^*) = 0, i = 1, \dots, m$.

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 $\alpha_i^* g_i(\mathbf{x}^*) = 0, i = 1, \dots, m$.

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

KKT 条件

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

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

称等式(9)为**拉格朗日不动性 (Lagrangian stationarity)**

KKT 条件

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

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

称等式(9)为**拉格朗日不动性 (Lagrangian stationarity)**

- (9)表明在最优解 \mathbf{x}^* 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

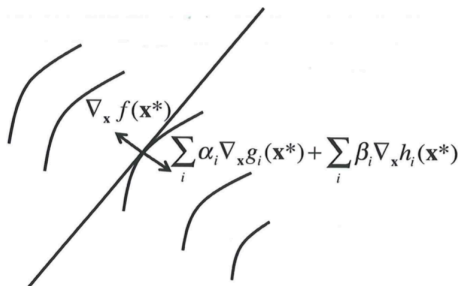
KKT 条件

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

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

称等式(9)为**拉格朗日不动性 (Lagrangian stationarity)**

- (9)表明在最优解 \mathbf{x}^* 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 $\mathbf{x}^* \in \mathbb{R}^d$, $\boldsymbol{\alpha}^* \in \mathbb{R}^m$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ 满足以下条件:

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

则 \mathbf{x}^* 是原始问题最优解, $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 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 \mathbf{x}^* 及任何对偶问题最优解 $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 必须满足以上条件.

KKT 条件

定理 (KKT 条件)

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

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

则 \mathbf{x}^* 是原始问题最优解, (α^*, β^*) 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 \mathbf{x}^* 及任何对偶问题最优解 (α^*, β^*) 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：

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

下面列出(10)的最优解需要满足的 KKT 条件：

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:

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

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

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

计算 \mathcal{L} 关于 ω 和 b 的梯度并令其等于零:

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

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

SVM: 最大化最小 margin

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

SVM: 最大化最小 margin

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

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

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

SVM: 最大化最小 margin

考虑到 α^* 还需满足条件(12)和对偶可行性, α^* 是以下对偶优化问题的解:

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

SVM: 最大化最小 margin

考虑到 α^* 还需满足条件(12)和对偶可行性, α^* 是以下对偶优化问题的解:

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

- 当特征 \mathbf{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

SVM: 最大化最小 margin

考虑到 α^* 还需满足条件(12)和对偶可行性, α^* 是以下对偶优化问题的解:

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

- 当特征 \mathbf{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 α^* , 由(11)可以得到原始问题最优解:

$$\omega^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$$

但是仍然不知道 b^* 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 $\alpha_i^* [-y_i(\omega^{*\top} \mathbf{x}_i + b^*) + 1] = 0, \forall i$, 可得:

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

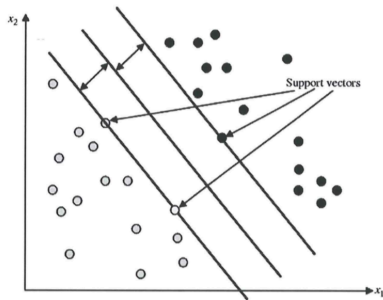
即在 $\alpha_i^* > 0$ 对应的点 (\mathbf{x}_i, y_i) 处, 不等式限制条件以等式成立, 说明该点到决策边界的距离最小 (为 $\gamma = 1 / \|\omega^*\|$), 训练集中这样的点 (\mathbf{x}_i, y_i) 被称为**支持向量 (support vectors)**, 它们是最靠近决策边界的点

支持向量

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

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

即在 $\alpha_i^* > 0$ 对应的点 (\mathbf{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 $\gamma = 1/\|\omega^*\|$), 训练集中这样的点 (\mathbf{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

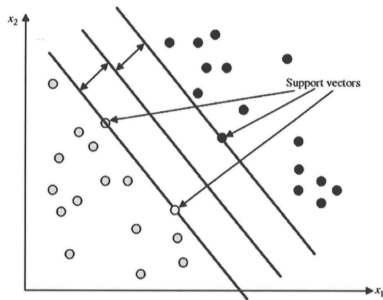


支持向量

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

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

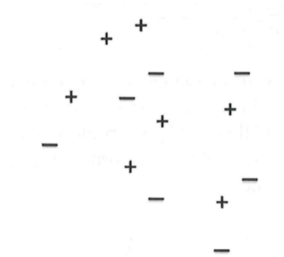
即在 $\alpha_i^* > 0$ 对应的点 (\mathbf{x}_i, y_i) 处, 不等式限制条件以等式成立, 说明该点到决策边界的距离最小 (为 $\gamma = 1/\|\omega^*\|$), 训练集中这样的点 (\mathbf{x}_i, y_i) 被称为**支持向量 (support vectors)**, 它们是最靠近决策边界的点



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

线性不可分情形

很多实际问题不在线性决策边界 (超平面) 可以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：

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

线性不可分情形

修改后的 SVM 求解的优化问题变为：

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

- (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 $y_i(\omega^\top \mathbf{x}_i + b) \geq 1$, 令 $\xi_i = 0$ 可以避免惩罚
 - ▶ 如果观察点 i 出现 $y_i(\omega^\top \mathbf{x}_i + b) = 1 - \xi_i$ 且 $\xi_i > 0$, 则需要付出代价 $C\xi_i$

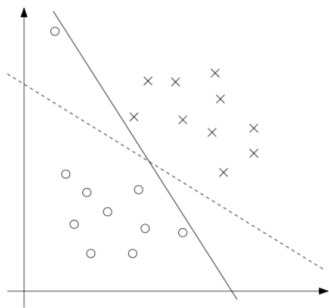
线性不可分情形

修改后的 SVM 求解的优化问题变为：

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

- (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 $y_i(\omega^\top \mathbf{x}_i + b) \geq 1$, 令 $\xi_i = 0$ 可以避免惩罚
 - ▶ 如果观察点 i 出现 $y_i(\omega^\top \mathbf{x}_i + b) = 1 - \xi_i$ 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样本点被正确分类 (ii) 使支持向量的 margin $\gamma = 1/\|\omega\|$ 尽可能大

线性不可分情形



- C 较大: 所有点的 margins 都是正的, 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

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

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

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

- 令 \mathcal{L} 关于 ξ_i 的一阶偏导数等于 0 得

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - r_i = 0 \implies \alpha_i^* = C - r_i^* \quad (16)$$

因此 $0 \leq \alpha_i^* \leq C, i = 1, \dots, n$

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

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

- 令 \mathcal{L} 关于 ξ_i 的一阶偏导数等于 0 得

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - r_i = 0 \implies \alpha_i^* = C - r_i^* \quad (16)$$

因此 $0 \leq \alpha_i^* \leq C, i = 1, \dots, n$

- \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 α^* 是以下对偶问题的解:

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

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 $0 \leq \alpha_i \leq C$

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^* 的计算与之前的方法不同, 由互补松弛性条件可得:

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

$$y_i(\omega^{\star\top} \mathbf{x}_i + b^*) < 1 \Rightarrow \xi_i^* > 0 \Rightarrow r_i^* = 0 \Rightarrow \alpha_i^* = C$$

$$0 < r_i^* < C \Rightarrow 0 < \alpha_i^* < C, \xi_i = 0 \Rightarrow y_i(\omega^{\star\top} \mathbf{x}_i + b^*) = 1$$

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^* 的计算与之前的方法不同, 由互补松弛性条件可得:

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

$$y_i(\omega^{\star\top} \mathbf{x}_i + b^*) < 1 \Rightarrow \xi_i^* > 0 \Rightarrow r_i^* = 0 \Rightarrow \alpha_i^* = C$$

$$0 < r_i^* < C \Rightarrow 0 < \alpha_i^* < C, \xi_i = 0 \Rightarrow y_i(\omega^{\star\top} \mathbf{x}_i + b^*) = 1$$

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

SMO 算法

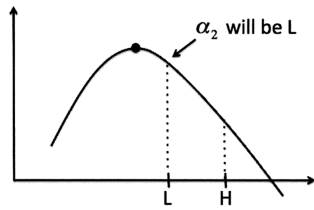
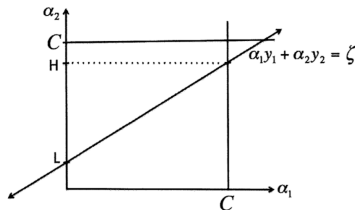
- Sequential Minimal Optimization (SMO) 是为求解 SVM 优化问题(17)设计的一个非常高效的算法，本质上是一种坐标下降算法

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 $\alpha_1, \dots, \alpha_n$ 满足(17)中所有限制条件, 如果固定 $\alpha_2, \dots, \alpha_n$, 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 $\alpha_2, \dots, \alpha_n$ 固定时, α_1 也被固定了

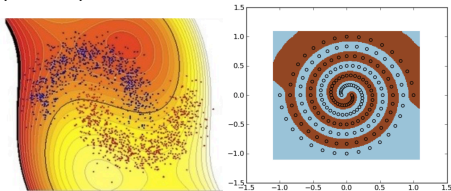
SMO 算法

- 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 $\alpha_1, \dots, \alpha_n$ 满足(17)中所有限制条件, 如果固定 $\alpha_2, \dots, \alpha_n$, 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 $\alpha_2, \dots, \alpha_n$ 固定时, α_1 也被固定了
- 考虑同时更新 α 中的 2 个元素, 比如固定 $\alpha_3, \dots, \alpha_n$, 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



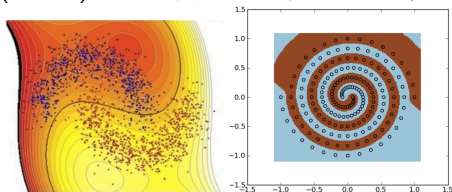
核函数

- 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

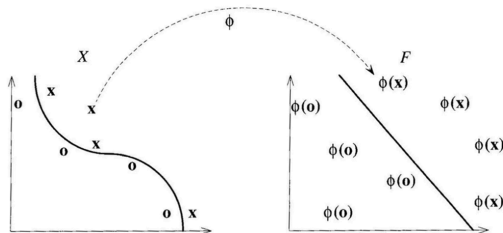


核函数

- 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 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 x 的特征空间升维到 $\phi(x)$ 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:

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

核函数

SVM 的优化问题可以转化为求解以下对偶问题:

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

- 注意到(18)只用到了特征的内积 $\mathbf{x}_i^{\top} \mathbf{x}_k$, 因此只需将(18)中 $\mathbf{x}_i^{\top} \mathbf{x}_k$ 替换为 $\phi(\mathbf{x}_i)^{\top} \phi(\mathbf{x}_k)$, 就得到了 $\phi(\mathbf{x})$ 空间的对偶问题

核函数

SVM 的优化问题可以转化为求解以下对偶问题:

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

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

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

核函数

- 很多时候计算核函数的成本很小，但计算 $\phi(\mathbf{x})$ 的成本却很高
 - ▶ 例如 $\mathbf{x} \in \mathbb{R}^d$ ，令 $\phi(\mathbf{x}) = (x_1^2, x_1 x_2, \dots, x_1 x_d, \dots, x_d x_1, \dots, x_d^2)^\top$ ，它的核函数为 $K_\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2$
 - ▶ $\phi(\mathbf{x})$ 的计算量为 $O(d^2)$ ，而 $K_\phi(\mathbf{x}, \mathbf{z})$ 的计算量只有 $O(d)$

核函数

- 很多时候计算核函数的成本很小，但计算 $\phi(\mathbf{x})$ 的成本却很高
 - ▶ 例如 $\mathbf{x} \in \mathbb{R}^d$ ，令 $\phi(\mathbf{x}) = (x_1^2, x_1x_2, \dots, x_1x_d, \dots, x_dx_1, \dots, x_d^2)^\top$ ，它的核函数为 $K_\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2$
 - ▶ $\phi(\mathbf{x})$ 的计算量为 $O(d^2)$ ，而 $K_\phi(\mathbf{x}, \mathbf{z})$ 的计算量只有 $O(d)$
- 如果核函数的形式为 $K_\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^r$ ，称其为**多项式核函数**，它对应的 $\phi(\mathbf{x})$ 中的每个元素都是一个 r 次多项式 $x_{i_1}x_{i_2}\dots x_{i_r}$ ，此时 $\phi(\mathbf{x})$ 的计算量为 $O(d^r)$ ，而 $K_\phi(\mathbf{x}, \mathbf{z})$ 的计算量仍为 $O(d)$ 。

核函数

- 很多时候计算核函数的成本很小，但计算 $\phi(\mathbf{x})$ 的成本却很高
 - ▶ 例如 $\mathbf{x} \in \mathbb{R}^d$ ，令 $\phi(\mathbf{x}) = (x_1^2, x_1 x_2, \dots, x_1 x_d, \dots, x_d x_1, \dots, x_d^2)^\top$ ，它的核函数为 $K_\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2$
 - ▶ $\phi(\mathbf{x})$ 的计算量为 $O(d^2)$ ，而 $K_\phi(\mathbf{x}, \mathbf{z})$ 的计算量只有 $O(d)$
- 如果核函数的形式为 $K_\phi(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^r$ ，称其为**多项式核函数**，它对应的 $\phi(\mathbf{x})$ 中的每个元素都是一个 r 次多项式 $x_{i_1} x_{i_2} \dots x_{i_r}$ ，此时 $\phi(\mathbf{x})$ 的计算量为 $O(d^r)$ ，而 $K_\phi(\mathbf{x}, \mathbf{z})$ 的计算量仍为 $O(d)$ 。
- 从计算的角度，如果只需要知道 $K_\phi(\mathbf{x}, \mathbf{z})$ 的值，不一定要先计算出 $\phi(\mathbf{x})$ 和 $\phi(\mathbf{z})$

核函数

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

核函数

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

- 若 SVM 在 x 的特征空间的最优线性决策边界为

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

当测试点 z 满足 $\omega^{*\top} z + b^* \geq 0$, 预测其为正, 反之为负

核函数

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

- 若 SVM 在 \mathbf{x} 的特征空间的最优线性决策边界为

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

当测试点 z 满足 $\boldsymbol{\omega}^{\star\top} \mathbf{z} + b^{\star} \geq 0$, 预测其为正, 反之为负

- 由拉格朗日不动性条件 (11) 可得 $\boldsymbol{\omega}^{\star} = \sum_{i=1}^n \alpha_i^{\star} y_i \mathbf{x}_i$, 则最优决策边界可写为:

$$\sum_{i=1}^n \alpha_i^{\star} y_i \mathbf{x}_i^{\top} \mathbf{x} + b^{\star} = 0 \quad (19)$$

核函数

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

- 若 SVM 在 \mathbf{x} 的特征空间的最优线性决策边界为

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

当测试点 z 满足 $\boldsymbol{\omega}^{\star\top} \mathbf{z} + b^{\star} \geq 0$, 预测其为正, 反之为负

- 由拉格朗日不动性条件 (11) 可得 $\boldsymbol{\omega}^{\star} = \sum_{i=1}^n \alpha_i^{\star} y_i \mathbf{x}_i$, 则最优决策边界可写为:

$$\sum_{i=1}^n \alpha_i^{\star} y_i \mathbf{x}_i^{\top} \mathbf{x} + b^{\star} = 0 \quad (19)$$

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

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

核函数

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

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

核函数

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

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

- (20)在 \mathbf{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 \mathbf{x} 的空间就得到一条多项式决策边界

核函数

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

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

- (20)在 \mathbf{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 \mathbf{x} 的空间就得到一条多项式决策边界
- 对于测试点 \mathbf{z} , 如果 $\sum_{i=1}^n \alpha_i^* y_i K_\phi(\mathbf{x}_i, \mathbf{z}) + b^* \geq 0$, 预测其为正, 反之为负

核函数

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

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

- (20)在 \mathbf{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 \mathbf{x} 的空间就得到一条多项式决策边界
- 对于测试点 \mathbf{z} , 如果 $\sum_{i=1}^n \alpha_i^* y_i K_\phi(\mathbf{x}_i, \mathbf{z}) + b^* \geq 0$, 预测其为正, 反之为负
- 以上分析表明, 不需要知道映射 $\phi(\cdot)$ 的具体形式, 只需要定义一个核函数 $K(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$?

核函数

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

- 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 $K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z}, \mathbf{x})$

核函数

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

- 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 $K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z}, \mathbf{x})$
- ▶ 对 \mathbb{R}^d 上的任意 n 个点 $\mathbf{x}_1, \dots, \mathbf{x}_n$, 定义矩阵

$$\mathbf{K} = (\mathbf{K}_{ij})_{n \times n} \quad (21)$$

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

定理 (Mercer)

函数 $K(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $\mathbf{x}_1, \dots, \mathbf{x}_n$, 由(21)定义的矩阵 \mathbf{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)**:

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

- ▶ (22)反映了点 \mathbf{x} 和 \mathbf{z} 的相似度

核函数

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

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

- ▶ (22)反映了点 \mathbf{x} 和 \mathbf{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

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

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

- ▶ (22)反映了点 \mathbf{x} 和 \mathbf{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 $\mathbf{x}^\top \mathbf{z}$ ，就可以将其替换为一个核函数 $K(\mathbf{x}, \mathbf{z})$ ，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)**