第五章：SVM
——优化论五部曲

Jason 博士

网易微专业 x 稀牛学院

人工智能数学基础微专业

机器学习中两种问题

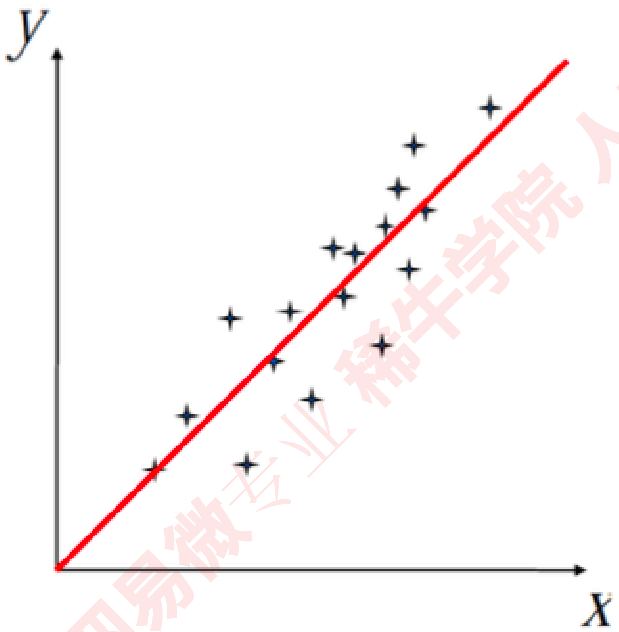
SVM 建模

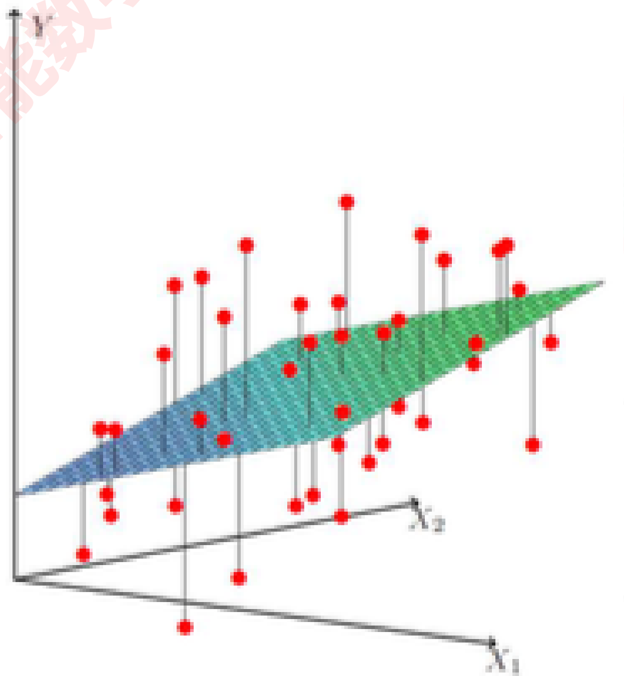
SVM 求解

SVM 扩展

回归

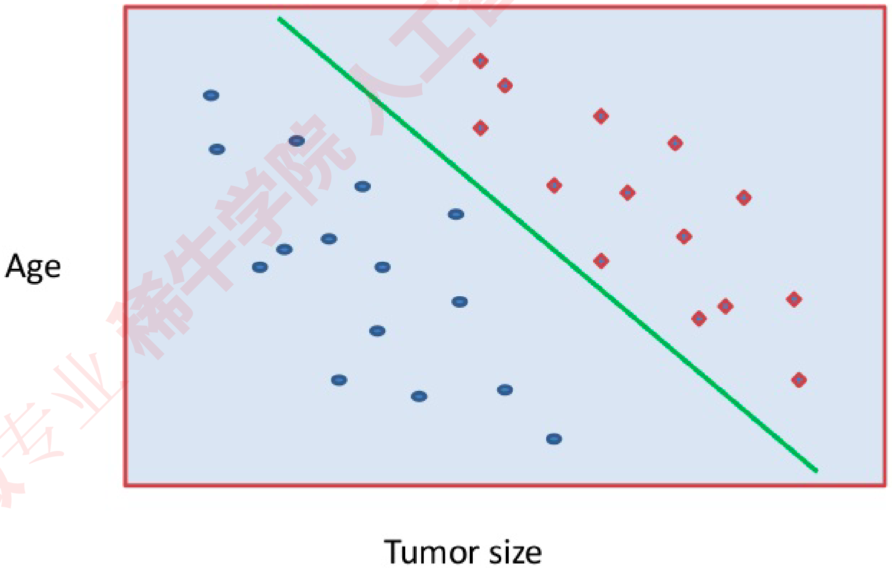
- 给定训练样本集 $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}, \mathbf{x}_i \in \mathbb{R}^d$
- 回归：预测的是连续值





分类

- 给定训练样本集 $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}, \mathbf{x}_i \in \mathbb{R}^d$
- 分类：预测的是离散值
- 二分类：$y_i \in \{-1, +1\}$



机器学习中两种问题

SVM 建模

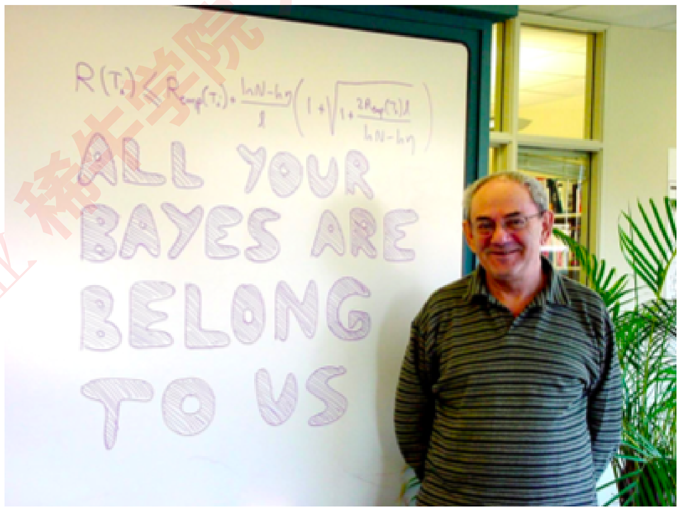
SVM 求解

SVM 扩展

SVM 的发明

Vladimir Vapnik

- Ph.D. in Statistics 1964
- Ins. Control Sci. Moscow 1964-1990 AT&T, USA 1990 2002 (developed Support Vector Machines)
- NEC Laboratories 2002-now
- U.S. National Academy of Engineering 2006



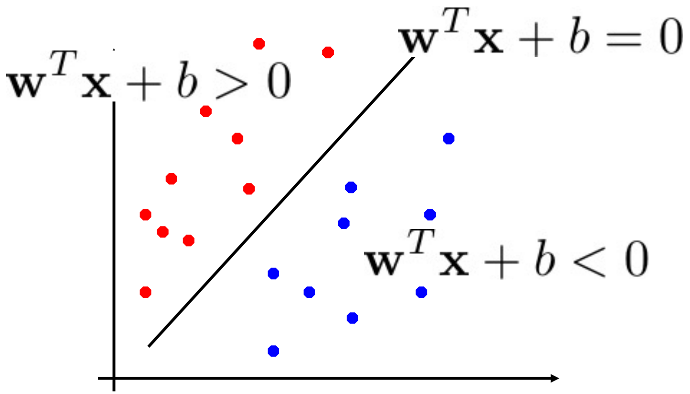
线性分类器和最优线性分类器

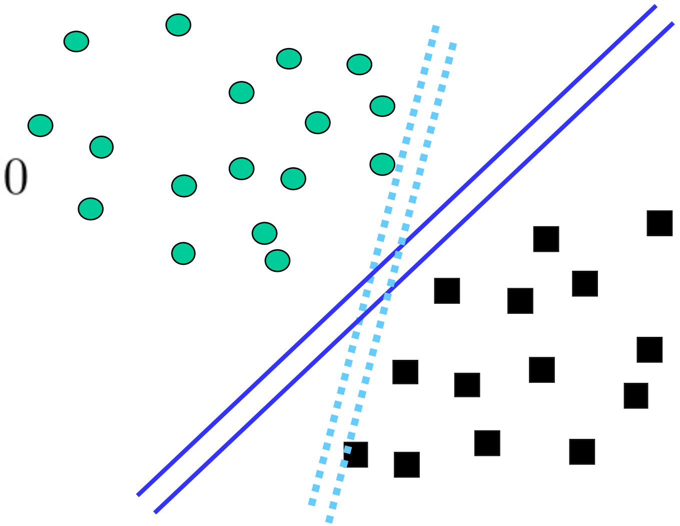
- 样本空间中，切分超平面

$$\mathbf{w}^T\mathbf{x} + b = 0$$

- $\mathbf{w}$ 法向量，b 偏移

$$f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^T\mathbf{x} + b)$$





点到超平面的距离

- 超平面 $H = \{\mathbf{w}^T\mathbf{x} + b = 0\}$
- 空间中任意一点 $\mathbf{p}$ 到平面的距离为

$$r = \frac{|\mathbf{w}^T\mathbf{p} + b|}{||\mathbf{w}||_2}$$

- 课堂推导证明

SVM 目标函数

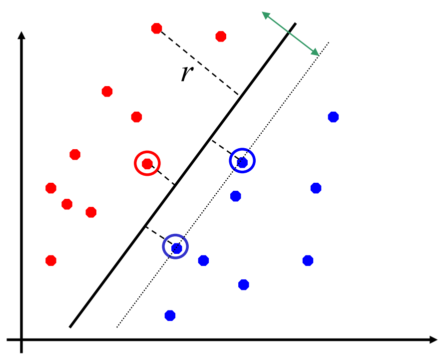
- 几何间隔 (Geometric margin)：

$$M = \min_i \ r_i = \min_i \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{||\mathbf{w}||_2}$$

- SVM 目标函数

$$\max_{\mathbf{w},b} M$$

$$\max_{\mathbf{w},b} \left\{ \min_i \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{||\mathbf{w}||_2} \right\}$$



简化目标函数

1 分子部分 $|\mathbf{w}^T\mathbf{x}_i + b|$ 改写为

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)$$

2 目标函数由 $\max_{\mathbf{w},b}\left\{\min_i \frac{|\mathbf{w}^T\mathbf{x}_i+b|}{||\mathbf{w}||_2}\right\}$ 变为

$$\max_{\mathbf{w},b}\left\{\min_i \frac{y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)}{||\mathbf{w}||_2}\right\}$$

$$\max_{\mathbf{w},b}\left\{\frac{1}{||\mathbf{w}||_2}\min_i\ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right)\right\}$$

简化目标函数

Rescaling

- 目标函数

$$\max_{\mathbf{w},b}\left\{\frac{1}{||\mathbf{w}||_2}\min_i\ y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)\right\}$$

- Rescaling $\mathbf{w}\rightarrow k\mathbf{w}$ 和 $b\rightarrow kb$，点到超平面距离不变.
- 因此可以设计距离最近的一个点，使得

$$y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)=1$$

- 同时有

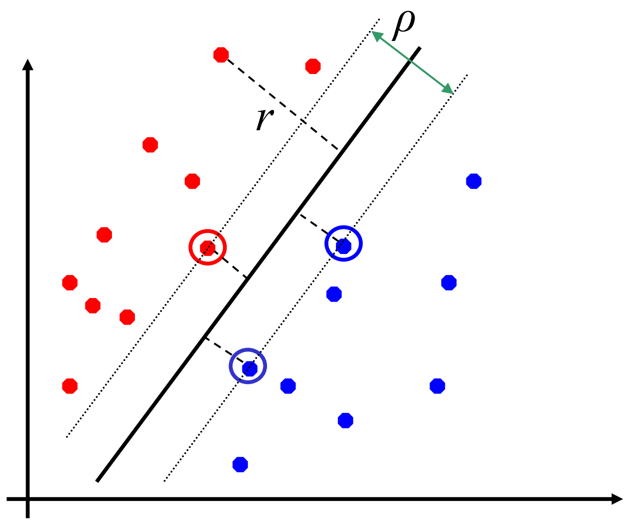
$$y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)\geq 1,\ i=1,\cdots,N$$

简化 SVM 目标函数

- SVM 基本型

$$
\begin{aligned}
\min \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\
\text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1,\ i = 1, \cdots, N
\end{aligned}
$$

- 几何解释



机器学习中两种问题

SVM 建模

SVM 求解

SVM 扩展

QP 问题

- QP 标准型

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{c}^T\mathbf{x} + d \\
\text{subject to} \quad & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\
& \mathbf{A}\mathbf{x} = \mathbf{b}
\end{aligned} \tag{1}
$$

- QP 问题

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i \\
\text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i, \; i = 1, \cdots, N \\
& \xi_i \geq 0
\end{aligned} \tag{2}
$$

转成标准型

- 定义

$$\mathbf{x} \in \mathbb{R}^{k} = \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\xi} \\ b \end{bmatrix}$$

- 定义

$$\mathbf{P} \in \mathbb{R}^{k \times k} = \begin{bmatrix} \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{c} \in \mathbb{R}^{k} = \begin{bmatrix} 0 \\ C \cdot \mathbf{1} \\ 0 \end{bmatrix}$$

$$\mathbf{G} \in \mathbb{R}^{2N \times k} = \begin{bmatrix} -\operatorname{diag}(\mathbf{y})\,\mathbf{X} & -\mathbf{I} & -\mathbf{y} \\ 0 & -\mathbf{I} & 0 \end{bmatrix}, \mathbf{h} \in \mathbb{R}^{2N} = \begin{bmatrix} -\mathbf{1} \\ 0 \end{bmatrix}$$

SVM 对偶问题

- Lagrangian 函数

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}||\mathbf{w}||_2^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)\right)$$

- 回忆对偶问题

$$d^* = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \left(\min_{\mathbf{w}, b} L(\mathbf{w}, b, \boldsymbol{\lambda}) \right)$$

- 令 $L(\mathbf{w}, b, \boldsymbol{\lambda})$ 对 $\mathbf{w}$ 和 b 的偏导为 0，则有 $\mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i$ 和 $0 = \sum_{i=1}^{N} \lambda_i y_i$
- 代入则有

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to} \quad & \sum_{i=1}^{N} \lambda_i y_i = 0, \ \lambda_i \geq 0, i = 1, \cdots, N \end{aligned}$$

KKT

- 假定能够解出 $\boldsymbol{\lambda}^*$，强对偶，根据 KKT 有
    - $y_i\left({\mathbf{w}^*}^T\mathbf{x}_i + b^*\right) \geq 1$
    - $\lambda_i^* \geq 0$
    - $\lambda_i^*\left(y_i\left({\mathbf{w}^*}^T\mathbf{x}_i + b^*\right) - 1\right) = 0$
    - $\mathbf{w}^* = \sum_{i=1}^N \lambda_i^* y_i \mathbf{x}_i, \sum_{i=1}^N \lambda_i^* y_i = 0$

解释

- 若 $\lambda_j^* = 0$，对 $\mathbf{w}^*$ 毫无贡献
- 若 $\lambda_j^* > 0$，则 $y_j\left({\mathbf{w}^*}^T\mathbf{x}_j + b^*\right) = 1$，为支撑向量.
- $\mathbf{w}^* = \sum_{i=1}^N \lambda_i^* y_i \mathbf{x}_i$ 代入 $y_j\left({\mathbf{w}^*}^T\mathbf{x}_j + b^*\right) = 1$，则 $b^* = y_j - \sum_{i=1}^N \lambda_i^* y_i\left(\mathbf{x}_i^T\mathbf{x}_j\right)$

SVM 最终步骤

- 最优分离超平面

$$\mathbf{w}^{*T}\mathbf{x} + b^* = 0$$

$$\sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i \mathbf{x} + b^* = 0$$

- 决策函数

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} \lambda_i^* y_i \mathbf{x}_i \mathbf{x} + b^*\right)$$

SMO 求解 $\boldsymbol{\lambda}^*$[Platt,1998]

思想

- 求解下列无约束优化问题

$$\max_{\boldsymbol{\lambda}} W(\lambda_1, \lambda_2, \cdots, \lambda_m)$$

- 除了之前谈到的算法，还可以考虑这样解：例如固定 $\lambda_2, \cdots, \lambda_m$，求解 λ_1，反复迭代.
- SVM 中不能只留一个变量，因为 $\sum_{i=1}^N \lambda_i y_i = 0$.
  - 选取一对需要更新的变量 λ_i 和 λ_j
  - 固定其它参数，求解 SVM 对偶式，得 λ_i 和 λ_j
  - 每次求解是关于 λ_i 的单变量 QP 问题，仅有的约束是 $\lambda_i \geq 0$，一维搜索问题[a].

[a]Platt,J.(1998)."Sequential minimal optimization: A fast algorithim for training support vector machines."Technical Report MSR-TR-98-14, Microsfot Research

机器学习中两种问题

SVM 建模

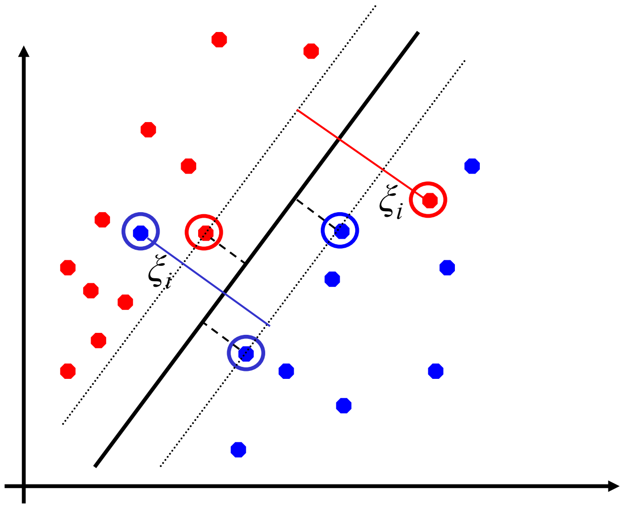
SVM 求解

SVM 扩展

软间隔

- SVM 软间隔

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i \\
\text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i,\ i = 1, \cdots, N \\
& \xi_i \geq 0
\end{aligned}
$$

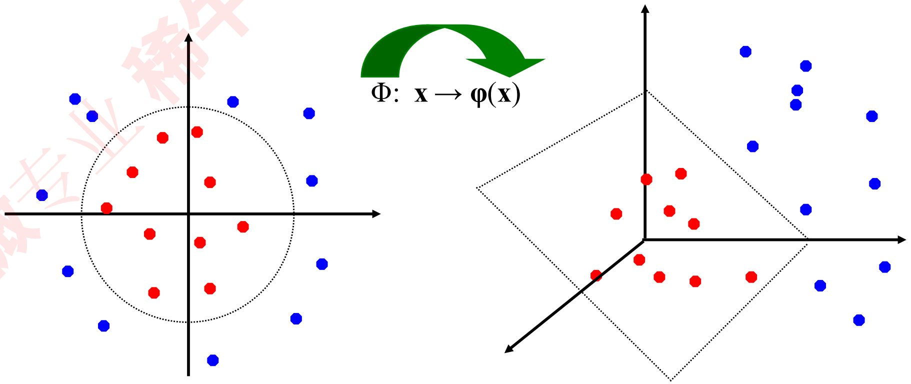


核函数 (1/3)

思想

- 线性不可分时?
- 特征映射 $\mathbf{x} \mapsto \phi(\mathbf{x})$
- SVM 目标函数

$$
\begin{aligned}
\min \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\
\text{subject to} \quad & y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i) + b\right) \geq 1, \; i = 1, \cdots, N
\end{aligned}
$$



核函数 (2/3)

- 特征映射 $\mathbf{x} \mapsto \phi(\mathbf{x})$，对偶问题

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{N} \lambda_i y_i = 0, \ \ \lambda_i \geq 0, i = 1, \cdots, N \end{aligned}$$

- 直接计算 $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ 很困难，假设有如下函数 (核技巧)

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

- 决策函数

$$f(\mathbf{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} \lambda_i^* y_i \kappa(\mathbf{x}_i, \mathbf{x}) + b^* \right)$$

核函数 (3/3)

常用核函数

- 线性核

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{x}_i^T \mathbf{x}_j$$

- 多项式核

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i^T \mathbf{x}_j\right)^d, d \geq 1$$

- 高斯核

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

- 拉普拉斯核

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\sigma}\right)$$

- Sigmoid 核

$$\kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \tanh\left(\beta \mathbf{x}_i^T \mathbf{x}_j + \theta\right)$$

SVM 案例演示

本章总结

本章参考资料

- Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.