支持向量机（Support Vector Machine, SVM）

研究生课程深入讲解

赵凯

上海大学大学通信与信息工程学院

https://kaizhao.net/teaching

本次课程结构

1. 引言与直观理解
2. Soft-margin SVM：加入松弛变量
3. 对偶问题与 KKT 条件
4. 核方法与非线性 SVM
5. SVM 的优缺点与应用
6. 小结与思考题

为什么需要支持向量机?

- 线性分类的基本形式:

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b), x \in \mathbb{R}^d, W \in \mathbb{R}^d.$$

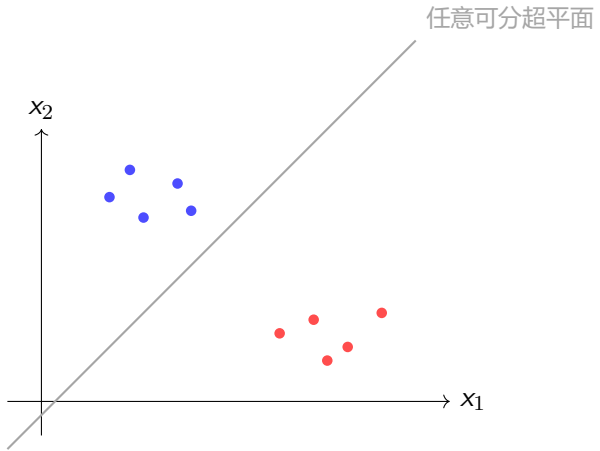
- 感知机（Perceptron）可以找到**某一个**可分超平面，但:
  - 不唯一;
  - 对噪声和边界样本敏感;
  - 没有显式最大化泛化能力。
- SVM 的核心思想:

最大间隔原则

在所有能正确分类训练数据的超平面中，选取**几何间隔**最大的那个。

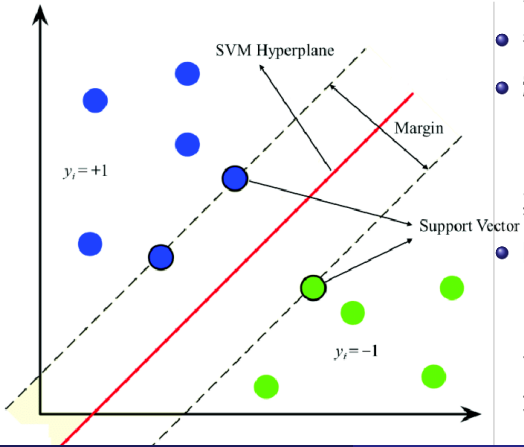
- 几何直观：更大的间隔 $\Rightarrow$ 更好的泛化能力。

线性可分情况下的几何示意（一）



- 只要数据线性可分，很多条直线都可以把正负类分开。
- 问题：**哪一条直线**具有更好的泛化能力？

线性可分情况下的几何示意（二）：最大间隔



- 输入特征向量：$\boldsymbol{x} \in \mathbb{R}^d$（本课默认 $\boldsymbol{x}$ 为 $d \times 1$ 的列向量）。
- 类别标签：$y \in \{+1, -1\}$。
- 参数：$\boldsymbol{w} \in \mathbb{R}^d$（同样视为 $d \times 1$ 列向量），$b \in \mathbb{R}$。
- **超平面（决策边界）**：

$$H = \{\boldsymbol{x} \mid \boldsymbol{w}^\top \boldsymbol{x} + b = 0\},$$

其中 $\boldsymbol{w}$ 是该超平面的**法向量**（图中紫色箭头）。

- **间隔边界（margins）**：

$$\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1,$$

两条虚线之间的距离为 $2/\|\boldsymbol{w}\|$；从决策边界到任一边界的距离为 $\frac{1}{\|\boldsymbol{w}\|}$。

什么是函数间隔？什么是几何间隔？（直观理解）

函数间隔（functional margin）对样本 (x_i, y_i) 定义为：

$$\hat{\gamma}_i = y_i(w^\top x_i + b).y_i \in \{+1, -1\}$$

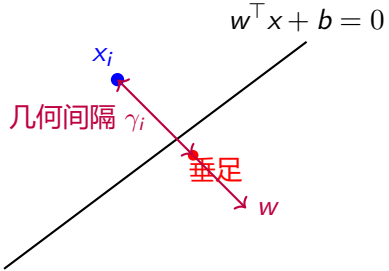
解释：

- 若符号正确，则样本被正确分类；
- 数值越大，分类“越自信”；
- **但它依赖 w 的缩放：**把 w 放大 10 倍，函数间隔也放大 10 倍。

因此函数间隔不能代表真实“距离”。

几何间隔（geometric margin）定义为：

$$\gamma_i = \frac{\hat{\gamma}_i}{\|w\|}.y_i \in \{+1, -1\}$$



Hard-margin SVM 的优化问题

- 对线性可分数据，SVM 通过约束形式将间隔归一化：

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \quad i = 1, \ldots, N.$$

- 在此约束下，几何间隔为：

$$\gamma = \frac{1}{\|\boldsymbol{w}\|}.$$

- 因此最大化间隔等价于：

$$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|} \quad \Longleftrightarrow \quad \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

- 得到 Hard-margin SVM 的原始问题：

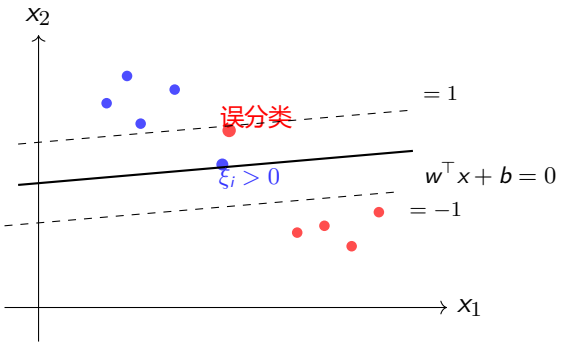
$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \quad i = 1, \ldots, N. \end{aligned}$$

- 这是一个带线性约束的凸二次规划（QP）。

为什么需要 Soft-margin?

- 实际数据通常：
  - 有噪声；
  - 存在离群点；
  - 甚至并非严格线性可分。
- 如果仍然强制所有样本 $y_i(\boldsymbol{w}^\top x_i + b) \geq 1$，会产生：
  - 过拟合；
  - 决策边界被少数异常点严重影响。
- Soft-margin SVM 的思路：
  - 允许部分样本**违背间隔约束**；
  - 但对违背程度进行惩罚。

Soft-margin 的几何示意



- 蓝色点在 margin 内部：$0 < \xi_i < 1$。
- 红色点被误分类：$\xi_i > 1$。
- 所有这些“违规程度”通过 ξ_i 和惩罚参数 C 体现。

Soft-margin SVM 的优化问题

- 引入松弛变量 $\xi_i \geq 0$，放宽约束：

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N.$$

- 同时惩罚 ξ_i 之和：

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

- 参数 C 的含义：
  - C 越大：对误分类惩罚越重，更趋向**硬间隔**，但可能过拟合。
  - C 越小：允许更多错误，更**鲁棒**，但可能欠拟合。

为什么需要对偶问题？（动机）

回顾原始优化问题：

$$\min_{\mathbf{w},b} \ \frac{1}{2}\|\mathbf{w}\|^2, \qquad \text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \ \forall i.$$

原始问题（Primal）存在两个局限：

- **(1) 直接在 $\mathbf{w}$ 上优化，不利于处理高维/特征映射**如果把样本映射到高维空间 $\phi(x)$，优化变量维度会急剧上升。
- **(2) 约束很多（N 条），结构不够简洁**不便于理解支持向量的作用，也不便于扩展到核方法。

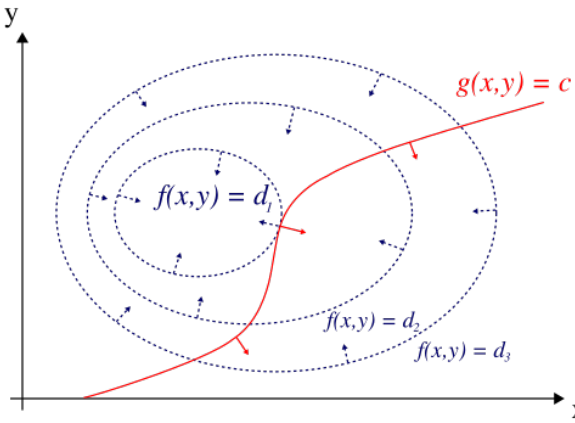
理想状态：如果优化目标可以**只依赖样本之间的内积**

$$\mathbf{x}_i^\top \mathbf{x}_j,$$

那么：

- 可以直接将内积替换为核函数 $K(x_i, x_j)$；
- 自然得到“稀疏解”，即**只有支持向量参与决策**；
- 不需要在高维空间显式计算 $\phi(x)$。

拉格朗日乘子法：优化目标和约束函数在最优点相切



回顾原始优化问题：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2, \qquad \text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \ \forall i.$$

- 红线 $g(x, y)$：约束函数，等式
- 蓝色虚线：目标函数 $f(x, y)$ 的等高线

在最优点，红线与蓝色虚线**相切**，其梯度方向平行，因此可引入乘子 λ。

拉格朗日乘子法：用于带约束优化

目标：求解带等式约束的优化问题：

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } g(\mathbf{x}) = 0.$$

核心思想：在最优解处，目标函数 f 的梯度方向必须与约束曲线 g 的梯度方向一致（或成比例），因此存在某个标量 λ，使得：

$$\nabla f(\mathbf{x}^*) = \lambda \nabla g(\mathbf{x}^*).$$

构造拉格朗日函数：

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x}).$$

优化条件为同时满足：

$$\nabla_{\mathbf{x}} L = 0, \qquad \nabla_{\lambda} L = -g(\mathbf{x}) = 0.$$

几何解释：在最优点处，等高线 $f(x) = \text{const}$ 与约束曲线 $g(x) = 0$ **相切**，其梯度方向平行，因此可引入乘子 λ。

拉格朗日乘子法：用于带约束优化

目标：求解带等式约束的优化问题：

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } g(\mathbf{x}) = 0.$$

在 SVM 中的作用：

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

可通过引入拉格朗日乘子转化为无约束问题，为求解对偶问题奠定基础。

Hard-margin SVM 的对偶形式

- **说明：** 在 Hard-margin SVM 中，每个约束 $y_i(w^\top x_i + b) \geq 1$ 都对应一个拉格朗日乘子。本课件采用 SVM 文献的标准记号 **α_i** (等价于一般优化中的 λ_i)。
- 对 $\boldsymbol{w}$: $\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$.
- 对 b: $\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0$.
- 代回得到对偶问题：

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, N, \\ & \sum_{i=1}^{N} \alpha_i y_i = 0. \end{aligned}$$

- 求得 α_i 后：$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$.
  其中只有 $\alpha_i > 0$ 的样本对应的点是支持向量。

KKT 条件与支持向量（Karush–Kuhn–Tucker Conditions）

KKT 条件 = Karush–Kuhn–Tucker 条件，是求解带不等式约束优化问题的必要条件。它包含四部分：

- **(K) 可行性（Primal feasibility）** 原问题的约束必须满足：$y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \geq 0.$
- **(K) 对偶可行性（Dual feasibility）** 拉格朗日乘子必须非负：$\alpha_i \geq 0.$
- **(T) 互补松弛（Complementary slackness）** $\alpha_i \left(y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1\right) = 0.$ 表示：要么点满足严格间隔条件，要么乘子为正，但不可能两个都“活跃”。

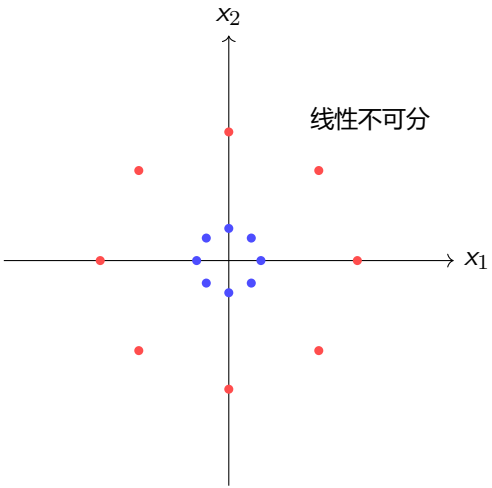
几何意义（非常关键）：

- 若 $\alpha_i = 0$，则点满足 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1$, 在 margin 外，对超平面没有“贡献”。
- 若 $\alpha_i > 0$，则 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$, 点恰好落在 margin 上，称为**支持向量**。它们决定了 $\mathbf{w}$：$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i.$

Soft-margin 情况： $0 \leq \alpha_i \leq C$, 支持向量分为：

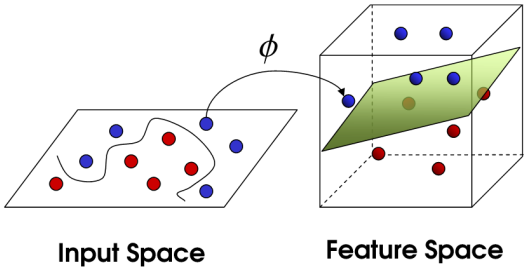
- $0 < \alpha_i < C$：点在 margin 上；
- $\alpha_i = C$：通常是误分类点或严重违反间隔约束的点。

非线性可分数据：为什么需要核?



- 在原始输入空间中，很难用一个线性超平面将两类分开。
- 直观想法：将数据映射到一个更高维的特征空间，使其线性可分。
- 但显式构造高维映射 $\phi(x)$ 的代价非常高、甚至不可行。
- 这就引出了**核函数（Kernel trick）**的思想。

核技巧（Kernel Trick）的直观示意



- SVM 的对偶形式中只出现 $\boldsymbol{x}_i^\top \boldsymbol{x}_j$。
- 将其替换为核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$，即可隐式在高维特征空间中做线性分类。

核 SVM 的决策函数

- 线性 SVM 的决策函数：

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i^\top \boldsymbol{x} + b\right).$$

- 将内积替换为核函数：

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right).$$

- 常见核函数：
  - 线性核：$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^\top \boldsymbol{z}$。
  - 多项式核：$K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^\top \boldsymbol{z} + c)^p$。
  - RBF/Gaussian 核：

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

- 超参数：C（软间隔）和核参数（如 σ、多项式阶数等）需要通过交叉验证选择。

SVM 的优势

- 几何意义清晰，以最大间隔为目标，具有较好的泛化能力。
- 通过核技巧，可以处理复杂的非线性分类问题。
- 在中小规模数据集上表现稳定，常作为强力 baseline。
- 对高维特征（如文本分类、基因表达数据）表现良好。

与 LDA 的对比：

- LDA：基于高斯分布假设的**生成式**模型，强调类间/类内散度。
- SVM：**判别式**模型，不需要分布假设，直接最大化 margin。
- 在数据近似高斯、类协方差相近时，两者决策边界可能相近。

SVM 的局限性

- 对大规模数据（特别是样本数 N 很大）时，二次规划的计算和存储成本较高。
- 参数选择（C、核参数）对性能影响大，需要调参。
- 对极端噪声和强 label noise 情况可能不够鲁棒（需要改进如 robust SVM）。
- 在很多现代深度学习任务中，SVM 更多作为：
  - 特征层之上的分类器；
  - 或用于小数据集下的基准模型。

本课小结

- 支持向量机的核心：**最大间隔线性分类。**
- Hard-margin SVM：

$$\min \frac{1}{2}\|w\|^2, \quad y_i(w^\top x_i + b) \geq 1.$$

- Soft-margin SVM：

$$\min \frac{1}{2}\|w\|^2 + C\sum \xi_i, \quad y_i(w^\top x_i + b) \geq 1 - \xi_i.$$

- 对偶形式只依赖内积 $\Rightarrow$ 核技巧实现非线性分类。
- 支持向量：KKT 条件中 $\alpha_i > 0$ 的样本，对决策边界起关键作用。

课后思考题

1. 推导 Soft-margin SVM 的对偶形式，并写出其约束与目标函数。
2. 对于给定的二维 toy 数据，分别用线性 SVM 和 RBF 核 SVM 训练，比较决策边界形状。
3. 思考：在高度不平衡数据集（正负样本数相差悬殊）中，如何改进 SVM 的损失或权重？
4. 将 SVM 与 Logistic Regression 的决策边界和损失函数进行对比：SVM 的 hinge loss 与 logistic loss 有何异同？

谢谢大家！

Questions?