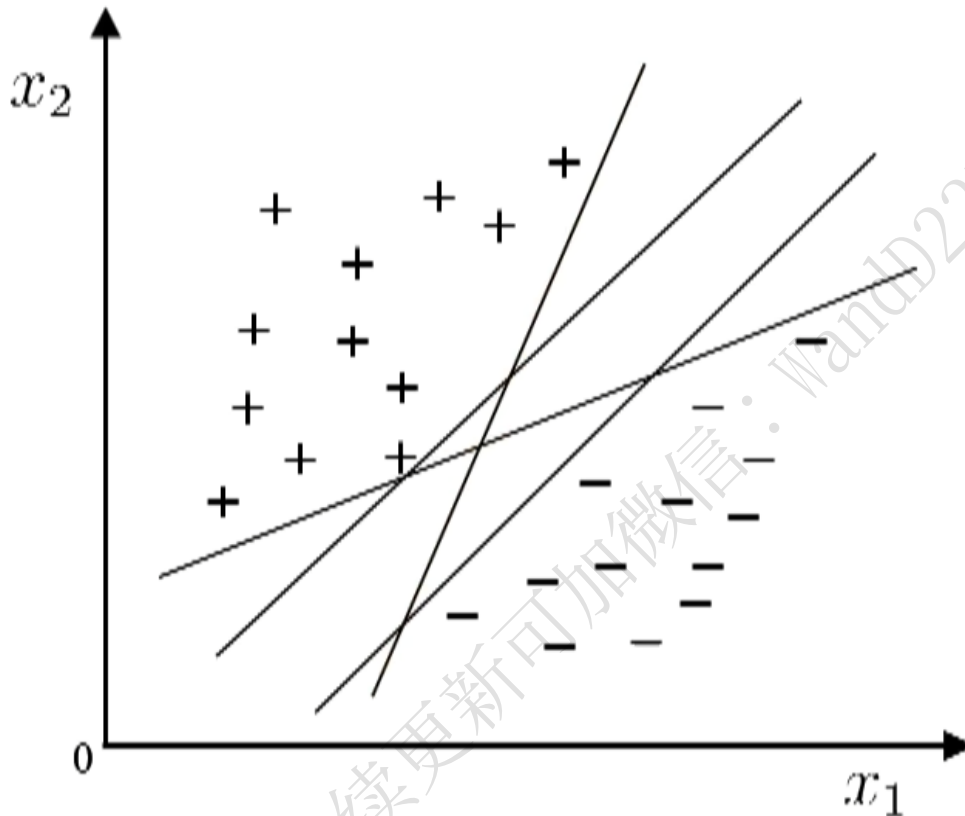


Lecture 6: 支持向量机

支持向量机基础概念

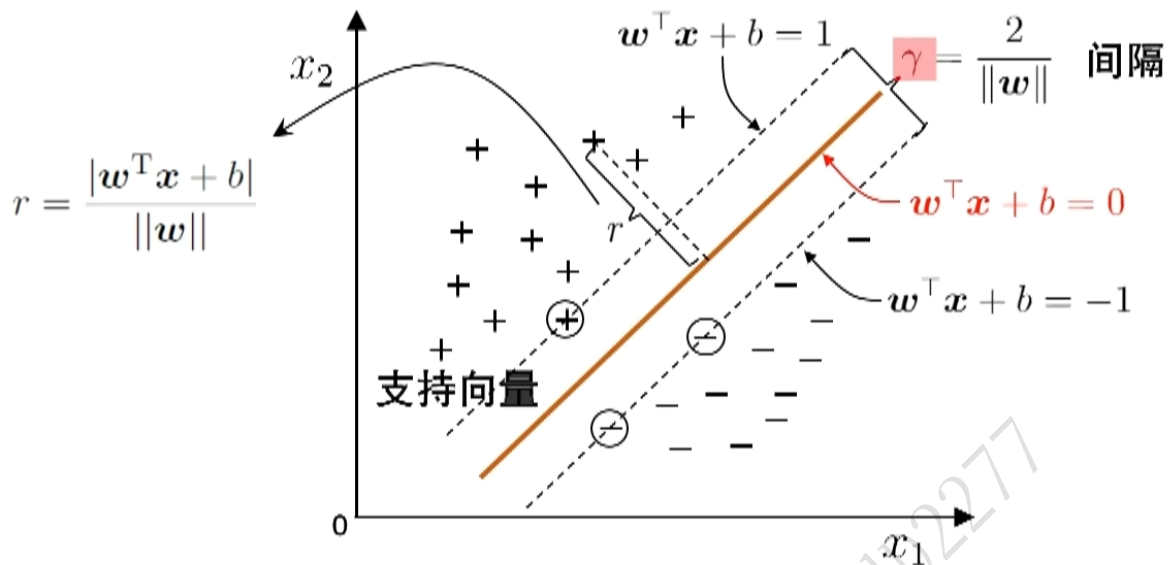
概念：支持向量机基础概念

- 引入：
 - 将训练样本分开的超平面可能有很多，哪一个更好呢？



- 分析：
 - “正中间”的超平面方程：鲁棒性最好，泛化能力最强；
- 效果：
 - 可以尽可能保证分类的正确性
 - 可以尽可能加大两个类别间的距离，使他们更容易区分；
- 目的：
 - 找到一个**间隔最大的参数**，使得超平面方程的泛化能力最好；
 - 假设定义支持向量到直线的举例为 1，则总的间隔就是 $2/||w||$ ，这个数就是超平面相对于整个数据集的间隔；
 - 离得越远越好，这样划分的把握越大、越准确；

超平面方程: $w^\top x + b = 0$



概念: 支持向量机基本型

- 概念:
 - 支持向量机是监督学习的线性分类器;

最大间隔: 寻找参数 w 和 b , 使得 γ 最大

$$\begin{aligned} \arg \max_{w, b} \quad & \frac{2}{\|w\|} \\ \text{s.t.} \quad & y_i(w^\top x_i + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned}$$



$$\arg \min_{w, b} \quad \frac{1}{2} \|w\|^2$$

$$\text{s.t.} \quad y_i(w^\top x_i + b) \geq 1, \quad i = 1, 2, \dots, m.$$

- 目的是最大化两个类别间的距离, 其等价于求最小化间距倒数的平方; , 因此寻找 $2/\|w\|$ 最大值等价于寻找 $1/2 \|w\|^2$ 的最小值;
- 这是一个凸二次规划问题, 能用优化计算包求解, 但可以有更高效的办法;
- 计算:
 - 利用拉格朗日乘数法, 计算 $\frac{1}{2} \|w\|^2$ 的最大值, 其需要满足约束条件:

拉格朗日乘子法

□ 第一步：引入拉格朗日乘子 $\alpha_i \geq 0$ 得到拉格朗日函数

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))$$

□ 第二步：令 $L(\mathbf{w}, b, \alpha)$ 对 $\mathbf{w}$ 和 b 的偏导为零可得

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i, \quad 0 = \sum_{i=1}^m \alpha_i y_i$$

□ 第三步：回代可得

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

- 要求最小值最大，因此回代后需要求 MAX；
- α 是对每一个样本都有一个约束，如果有 m 个样本，就需要解出 m 个 α 的约束；
- 解出 m 个 α 的约束后，得到下面的最终模型；

$$\text{最终模型: } f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

KKT条件：

$$\begin{cases} \alpha_i \geq 0; \\ 1 - y_i f(\mathbf{x}_i) \leq 0; \\ \alpha_i (1 - y_i f(\mathbf{x}_i)) = 0. \end{cases} \quad \longrightarrow \quad \begin{aligned} & \text{必有 } \alpha_i = 0 \text{ 或} \\ & y_i f(\mathbf{x}_i) = 1 \end{aligned}$$

解的**稀疏性**: 训练完成后，最终模型仅与支持向量有关

支持向量机(Support Vector Machine, SVM) 因此而得名

- 概念：解的稀疏性
 - 由于拉格朗日乘数法，自然引入了 KKT 条件，而由 KKT 条件，导致必定有 $\alpha=0$ 或者 $yf(x)=1$ ，因此只有少数点满足，这个是间隔上的点；

求解方法

概念：求解方法 - SMO 求解

基本思路：不断执行如下两个步骤直至收敛

- 第一步：选取一对需更新的变量 α_i 和 α_j
- 第二步：固定 α_i 和 α_j 以外的参数，求解对偶问题更新 α_i 和 α_j

仅考虑 α_i 和 α_j 时，对偶问题的约束 $0 = \sum_{i=1}^m \alpha_i y_i$ 变为

$$\alpha_i y_i + \alpha_j y_j = c, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0$$

用 α_i 表示 α_j ，代入对偶问题

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i^T x_j \quad \text{有闭式解!}$$

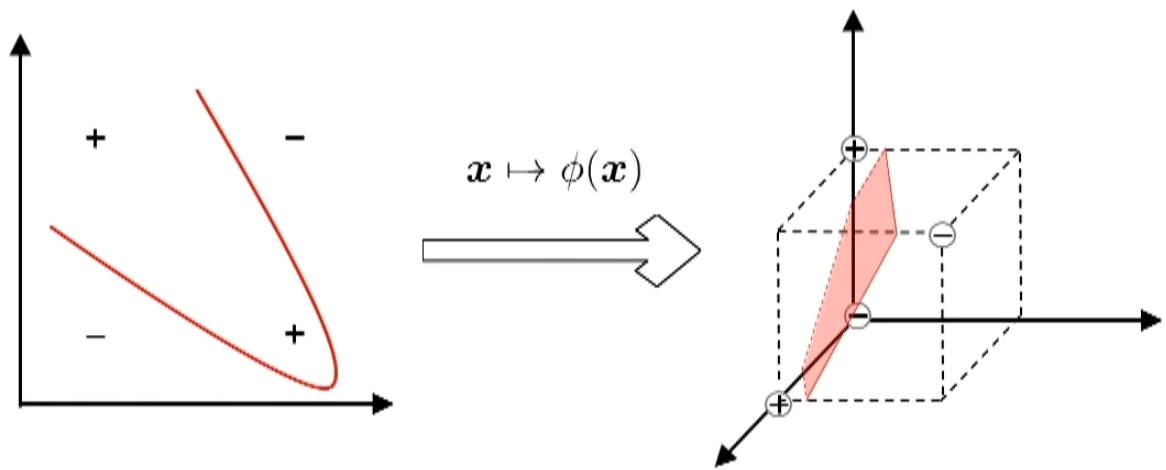
对任意支持向量 (x_s, y_s) 有 $y_s f(x_s) = 1$ 由此可解出 b

- 为提高鲁棒性，通常使用所有支持向量求解的平均值；

特征空间映射

概念：特征空间映射

- 引入：
 - 若不存在一个能正确划分两类样本的超平面，则需要将这些样本，从原始空间映射到一个更高维的特征空间，使样本在这个特征空间内线性可分；
 - 即：特征空间映射；
- 图示：



• 总结:

- 如果原始空间是有限维 (属性数有限), 那么 **一定存在一个高维特征空间使样本线性可分**;
- 这个高维可能是一个无限维;

概念: 映射至高维空间

• 概念:

- 原本是一个有限维上、低纬度上不可划分的向量, 将其提升至高维后可以线性可分, 此时就可以利用支持向量机的一般解法求间隔、划分;

设样本 x 映射后的向量为 $\phi(x)$, 划分超平面为 $f(x) = w^T \phi(x) + b$

原始问题

$$\begin{aligned} \min_{w, b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y_i (w^T \phi(x_i) + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned}$$

对偶问题

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned}$$

- 但也有引入的新问题: 映射为高维后、其维度可能十分高, 此时求其 **MAX** 时、可能会涉及非常高维向量的内积, 这个内积十分难以计算;
- 每一次预测都需要计算非常高维的内积;
- 假设: 每一次都可以利用一个计算, 代替这个高纬度的内积, 就可以减少大量的计算;
- 这个东西就是核函数;

核函数

概念: 核函数

• 概念:

- 基本的思路，是设计一个核函数，在低维的函数空间中进行计算，但其计算的结果满足高维的内积，从而绕过显式考虑特征映射、以及计算高维内积的困难；

基本思路：设计核函数

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

绕过显式考虑特征映射、以及计算高维内积的困难

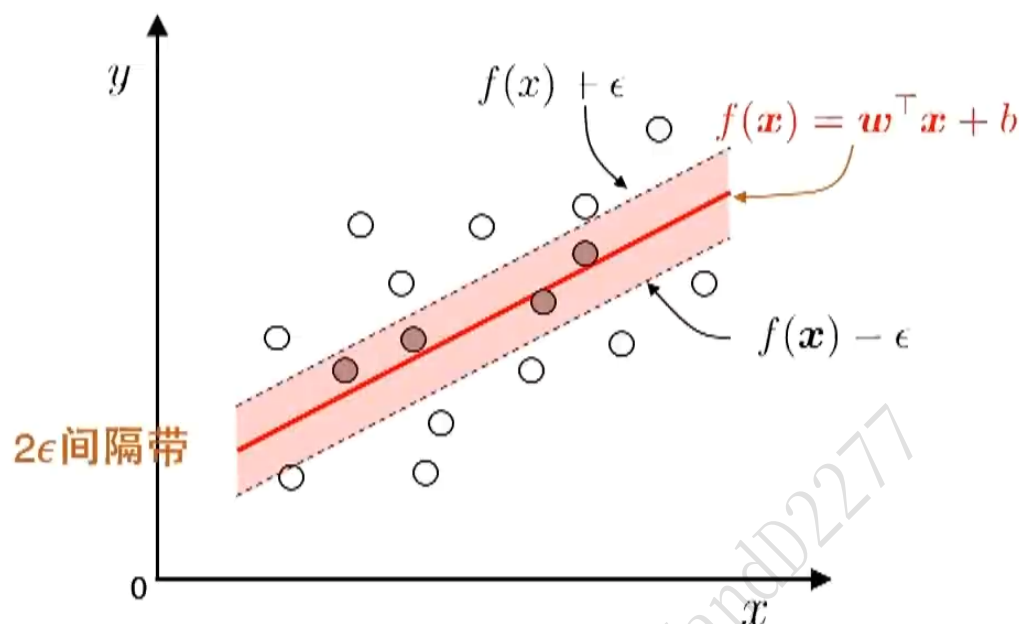
- 概念：Mercer 定理
 - 函数现在起到一个空间的变换的作用；
 - 通过一个空间变换函数，可以将若干点从低维空间中的点、映射到高维空间中的点；
 - 并且：如果两个低维空间中的点，它们之间的线性距离固定了下来，也就可以固定下来其高维线性空间中的距离；
 - 可以通过一个矩阵，来计算空间中两点的距离；
 - 此时可以求核函数、在甚高维空间中、两个点的距离的矩阵；
 - 这个矩阵的主对角线为 0，副对角线对称，一旦满足了这样一个条件，其也就满足了甚高维的空间；
 - 而核矩阵也就指定了任意两点间的距离；
 - 任何一个核函数，都隐式地定义了一个 RKHS (Reproducing Kernel Hilbert Space，再生核希尔伯特空间)；
- 分析：
 - 这样就把甚高维空间中、求点积的问题，转换为了求核函数；
 - 但这里求得的核函数不一定就等于甚高维，因此可能存在很多个核函数，然后映射到甚高维的函数存在于多个核函数中间；
 - 因此：就需要在很多个核函数中，寻找一个最合适的，此时就需要模型选择技术、选择一个最好的核函数；
 - “核函数选择”成为决定支持向量机性能的关键；

如何使用

概念：支持向量机 (SVM) 的使用

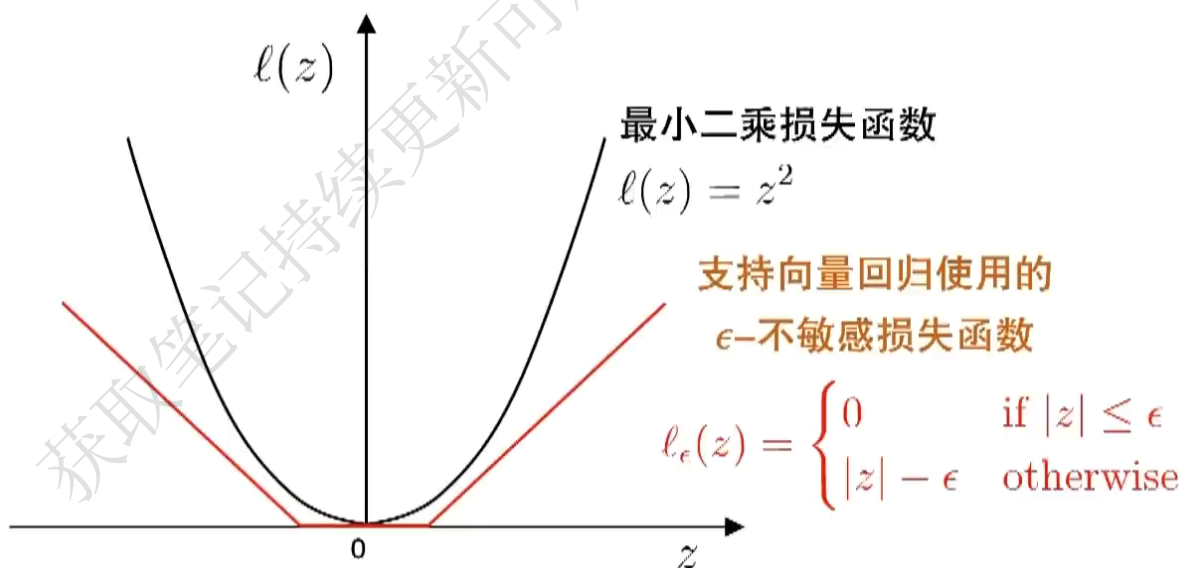
- 概念：

基本思路: 允许模型输出与实际输出间存在 2ϵ 的差别



-
- 让中间的点不计算损失，外面的点计算损失，即计算回归；
- 实现：

落入 2ϵ 间隔带的样本不计算损失



原始问题

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_i, \hat{\xi}_i} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \hat{\xi}_i) \\ \text{s.t.} \quad & \underline{f(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i}, \\ & y_i - f(\mathbf{x}_i) \leq \epsilon + \hat{\xi}_i, \\ & \xi_i \geq 0, \hat{\xi}_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned}$$

对偶问题

$$\begin{aligned} \max_{\alpha, \hat{\alpha}} \quad & \sum_{i=1}^m y_i (\hat{\alpha}_i - \alpha_i) - c(\hat{\alpha}_i + \alpha_i) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) = 0, \quad 0 \leq \alpha_i, \hat{\alpha}_i \leq C \end{aligned}$$

预测

$$f(\mathbf{x}) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \mathbf{x}_i^T \mathbf{x} + b$$

- 概念：如何实现并使用 SVM

如何使用SVM?

- 入门级——实现并使用各种版本SVM
- 专业级——尝试、组合核函数
- 专家级——根据问题而设计目标函数、替代损失、进而……