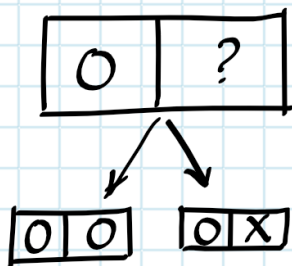
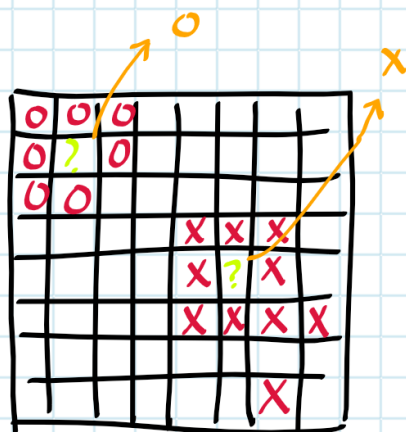
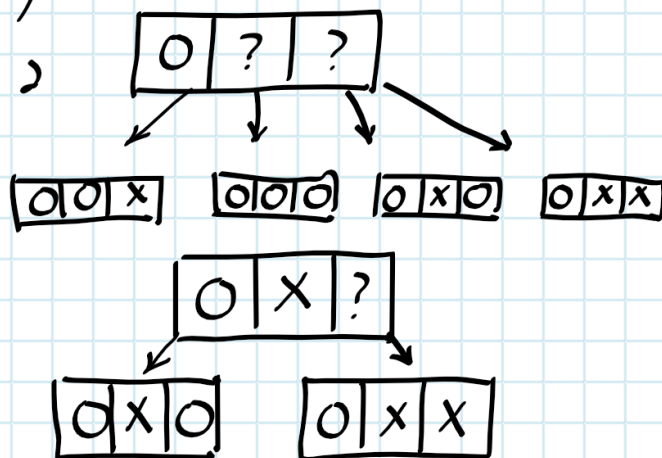


没有免费午餐定理 (No Free Lunch Theorem)

如果我们不对特征空间有先验假设, 则所有算法的平均表现是一样的!



0: Class 1
X: Class 2



0 花瓣
X 蜜蜂

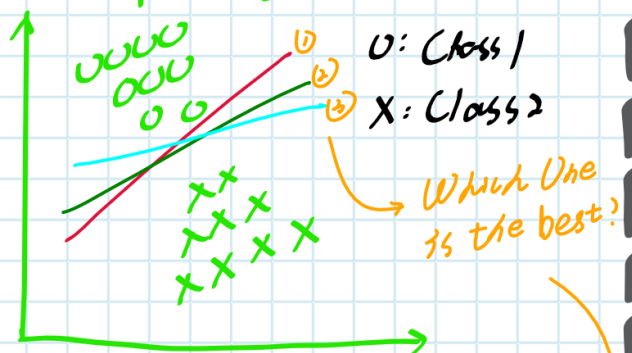
特征差距小的样本更有可能是一类~

支持向量机 (Support Vector Machine)

小样本方法

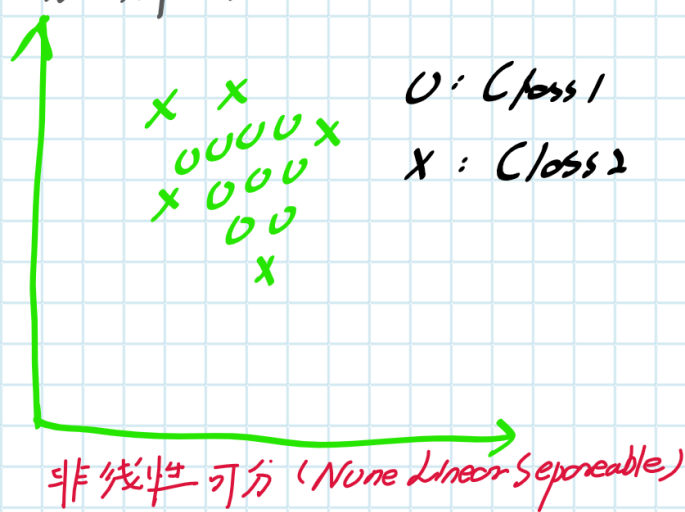
小样本依旧能获得很好的结果!!!

① 线性模型

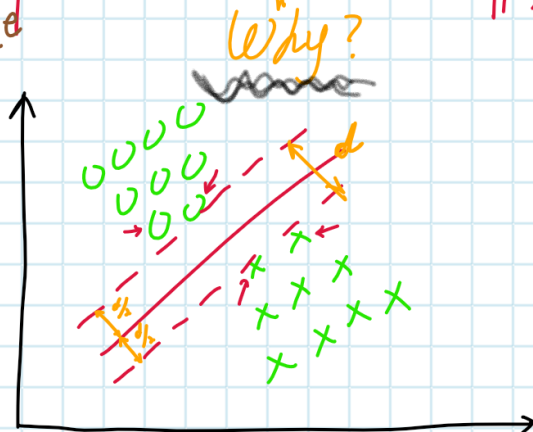


线性可分 (Linear Separable)

performance measure
性能指标



非线性可分 (None Linear Separable)



d: Margin

支持向量 (Support Vectors)
"→" 所指的数据

定义:

- ① 训练数据及标签 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
 \vec{x} 向量 \vec{y} 标签 $(y = +1/-1)$
- ② 线性模型 (W, b) $W^T x + b = 0$ (超平面)
 $(Hyperspace)$
 $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$
 \vec{w} 向量 b 常数

③ 一个训练集线性可分是指:

- $\{ (x_i, y_i) \}_{i=1 \sim N}$
- $\exists (w, b)$, 使: 对 $\forall i=1 \sim N$, 有:
- (a) 若 $y_i = +1$, 则 $w^T x_i + b \geq 0$
- (b) 若 $y_i = -1$, 则 $w^T x_i + b < 0$



$y_i [w^T x_i + b] \geq 0$ ---- 公式1

优化问题: (凸优化问题, 二次规划问题)

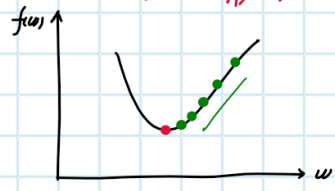
- 最小化 (Minimize): $\|w\|^2 \cdot \frac{1}{2}$ 求导方便
- 限制条件 (Subject to): $y_i (w^T x_i + b) \geq 1, i=1 \sim N$

事实1: $w^T x + b = 0$ 与 $aw^T x + ab = 0$ 是同一个平面
 $a \in \mathbb{R}^+$
 若 (w, b) 满足公式1, 则 (aw, ab) 也满足公式1

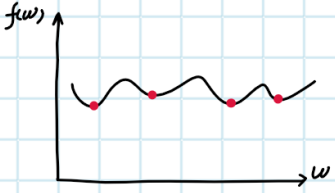
事实2: 点到面的距离公式:
 平面: $w_1 x + w_2 y + b = 0$, 点: (x_0, y_0) 到平面的距离
 $d = \frac{|w_1 x_0 + w_2 y_0 + b|}{\sqrt{w_1^2 + w_2^2}}$
 向量 x_0 到超平面 $w^T x + b = 0$ 的距离
 $d = \frac{|w^T x_0 + b|}{\|w\|} \xrightarrow{\text{求导方便}} \sqrt{w_1^2 + w_2^2 + \dots + w_n^2}$
 我们可以用 a 来缩放 $(w, b) \rightarrow (aw, ab)$
 最终使在支持向量 x_0 上有:
 $|w^T x_0 + b| = 1$
 此时支持向量与平面距离:
 $d = 1/\|w\|$

- 二次规划 (Quadratic Programming):
- ① 目标函数 (Objective Function) 是二次项
- ② 限制条件是一次项

要么无解/要么只有一个取值



\Rightarrow 全局最优解



\Rightarrow Hard to solve

- ① SVM是最大化间隔 (Margin) 的分类算法
- ② 优化问题

训练样本 $\{ (x_i, y_i) \}_{i=1 \sim N}$
 \vec{x} 向量 \vec{y} 标签

- ③ 最小化: $\frac{1}{2} \|w\|^2$
- 限制条件: $y_i [w^T x_i + b] \geq 1 (i=1 \sim N)$

SVM处理非线性:

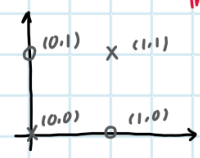
- ① 最小化: $\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$ (Soft Variable) 松弛变量
- ② 限制条件: $y_i [w^T x_i + b] \geq 1 - \xi_i (i=1 \sim N)$
 $\xi_i \geq 0$

$C \sum_{i=1}^N \xi_i$: 正则项 (Regulation Term)
 \rightarrow 事先规定的参数



定义一个高维的映射 $\phi(x)$:

\vec{x} (低维) $\xrightarrow{\phi}$ $\phi(\vec{x})$ (高维)



\Rightarrow 异或问题

$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in C, x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in C, x_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in C, x_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in C$
 $\phi(x): x = \begin{bmatrix} a \\ b \end{bmatrix} \xrightarrow{\phi} \phi(x) = \begin{bmatrix} a^2 \\ b^2 \\ a \\ b \end{bmatrix}$
 $\phi(x_1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \phi(x_2) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \in C$
 $\phi(x_3) = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \phi(x_4) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \in C$

我们可以不知道无限维映射 $\phi(x)$ 的显式表达,
 只需要知道一个核函数 (Kernel function)

$k(x_1, x_2) = \phi(x_1)^T \phi(x_2)$

则①这个优化式依然可解

核函数

① $k(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}$ (高斯核)
 $= \phi(x_1)^T \phi(x_2)$

② $k(x_1, x_2) = (x_1^T x_2 + 1)^d$ (多项式核)
 $= \phi(x_1)^T \phi(x_2)$

$k(x_1, x_2) \Rightarrow \phi(x_1)^T \phi(x_2)$ 的必要条件:

- ① $k(x_1, x_1) = k(x_2, x_2)$ (对称性)
- ② $\forall C_i, x_i (i=1 \sim N)$, 有: (半正定性)
 $\sum_{i=1}^N \sum_{j=1}^N C_i C_j k(x_i, x_j) \geq 0$

优化理论:

- ① « Convex optimization »
- ② « Nonlinear Programming »

原问题 (Primal Problem) → 非常奇怪

最小化: $f(w)$

限制条件: $g_i(w) \leq 0 \quad (i=1 \sim K)$
 $A_i(w) = 0 \quad (i=1 \sim M)$

对偶问题 (Dual Problem):

① 定义: $L(w, \alpha, \beta)$

$$= f(w) + \sum_{i=1}^K \alpha_i g_i(w) + \sum_{i=1}^M \beta_i A_i(w)$$

$$= f(w) + \alpha^T g(w) + \beta^T A(w)$$

$$\begin{bmatrix} g_1(w) \\ g_2(w) \\ \vdots \\ g_K(w) \end{bmatrix} \quad \begin{bmatrix} A_1(w) \\ A_2(w) \\ \vdots \\ A_M(w) \end{bmatrix}$$

② 对偶问题的定义:

最大化: $\theta(\alpha, \beta) = \inf_{w \in \mathbb{R}^n} L(w, \alpha, \beta)$ → 最小值

限制条件: $\alpha_i \geq 0 \quad (i=1 \sim K)$

定理: 如果 w^* 是原问题的解, 而 α^*, β^* 是对偶问题的解, 则有:

$$f(w^*) \geq \theta(\alpha^*, \beta^*)$$

$$\text{证: } \theta(\alpha^*, \beta^*) = \inf_{w \in \mathbb{R}^n} L(w, \alpha^*, \beta^*) \leq L(w^*, \alpha^*, \beta^*) \\ = f(w^*) + \sum_{i=1}^K \alpha_i^* g_i(w^*) + \sum_{i=1}^M \beta_i^* A_i(w^*) \leq f(w^*)$$

$\geq 0 \leq 0 \quad = 0$

$$\text{定义: } G = f(w^*) - \theta(\alpha^*, \beta^*) \geq 0$$

G 叫做原问题与对偶问题的问距 (Duality Gap)

对于某些特殊的优化问题, 可以证明: $G = 0$

强对偶定理:

若 $f(w)$ 为凸函数, 且 $g_i(w) = A_i w + b$, $A_i(w) = C_i w + b$,

则此优化问题的原问题与对偶问题问距为 0, 即 $f(w^*) = \theta(\alpha^*, \beta^*)$

对于 $\forall i=1 \sim K$, 或者 $\alpha_i^* = 0$, 或者 $g_i(w^*) = 0$
(KKT 条件)

将复杂的原问题的原问题转化为对偶问题

最小化 $\frac{1}{2} \|w\|^2 + C \sum \xi_i$

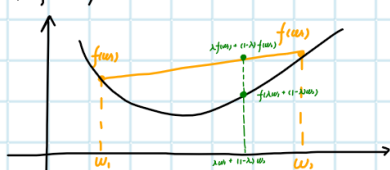
限制条件 $\odot y_i [w^T \phi(x_i) + b] \geq 1 - \xi_i$
 $\odot \xi_i \geq 0$

$$\frac{1}{2} \|w\|^2 - C \sum \xi_i$$

$$\odot y_i [w^T \phi(x_i) + b] \leq 1 + \xi_i$$

$$\odot \xi_i \leq 0$$

关于凸函数:



$\forall w_1, w_2, \forall \lambda \in [0, 1]$,

$$f(\lambda w_1 + (1-\lambda) w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2)$$

数学定义

对偶问题:

最大化 $\theta(\alpha, \beta) = \inf_{w \in \mathbb{R}^n} \left\{ \frac{1}{2} \|w\|^2 - C \sum \xi_i + \sum \alpha_i [1 - \xi_i - y_i w^T \phi(x_i) - y_i b] \right\}$

限制条件 $\alpha_i \geq 0 \quad \beta_i \geq 0 \quad (i=1 \sim N)$

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \Rightarrow w - \sum_{i=1}^N \alpha_i y_i \phi(x_i) = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow -C + \beta_i + \alpha_i = 0 \\ \frac{\partial L}{\partial b} = 0 \Rightarrow -\sum_{i=1}^N \alpha_i y_i = 0 \end{cases}$$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \quad f(w) \rightarrow \text{数}$$

若 $f(w) = \frac{1}{2} \|w\|^2$

则 $\frac{\partial f}{\partial w} = w$

若 $f(w) = w^T x$

则 $\frac{\partial f}{\partial w} = x$

最大化: $\theta(\alpha, \beta) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j)$

限制条件: $\odot 0 \leq \alpha_i \leq C$

$$\odot \sum_{i=1}^N \alpha_i y_i = 0$$

凸优化问题

SMO 算法

测试流程:

测试样本 x ,

$$\begin{cases} \text{若 } w^T \phi(x) + b \geq 0, \text{ 则 } y = +1 \\ \text{若 } w^T \phi(x) + b < 0, \text{ 则 } y = -1 \end{cases}$$

SVM 算法:

① 训练流程

输入训练样本 $\{(x_i, y_i)\}_{i=1 \sim n}$

(解优化问题)

最大化: $\theta(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j)$

限制条件: $\odot 0 \leq \alpha_i \leq C$

$$\odot \sum_{i=1}^n \alpha_i y_i = 0$$

算 b , 找一个 $0 < \alpha_i < C$.

$$b = \frac{1 - y_i \sum_{j=1}^n \alpha_j y_j K(x_i, x_j)}{y_i}$$

② 测试流程

输入测试样本 x

$$\begin{cases} \text{若 } \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b \geq 0, \text{ 则 } y = +1 \\ \text{若 } \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b < 0, \text{ 则 } y = -1 \end{cases}$$

SVM内核函数:

Linear (线性内核): $K(x, y) = x^T y$

Poly (多项式内核): $K(x, y) = (x^T y + 1)^d$

Rbf (高斯径向基函数内核): $K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$

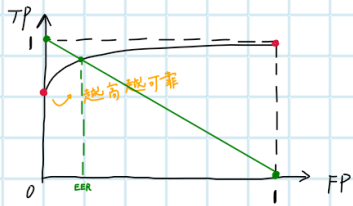
Tanh (Tanh 核): $K(x, y) = \tanh(\beta x^T y + b)$ $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

混淆矩阵

		预测	
实际		正样本	负样本
	正样本	True Positive (TP)	False Negative (FN)
	负样本	False Positive (FP)	True Negative (TN)

$$1. TP + FN = FP + TN = 1$$

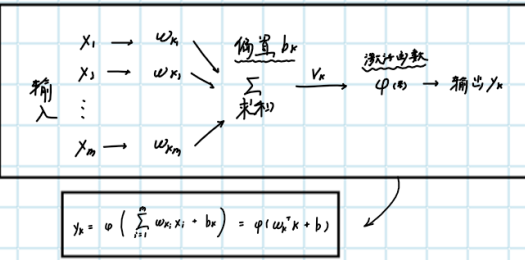
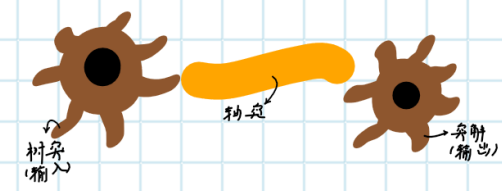
2. 对于同一个系统, 若 TP ↑, 则 FP ↓



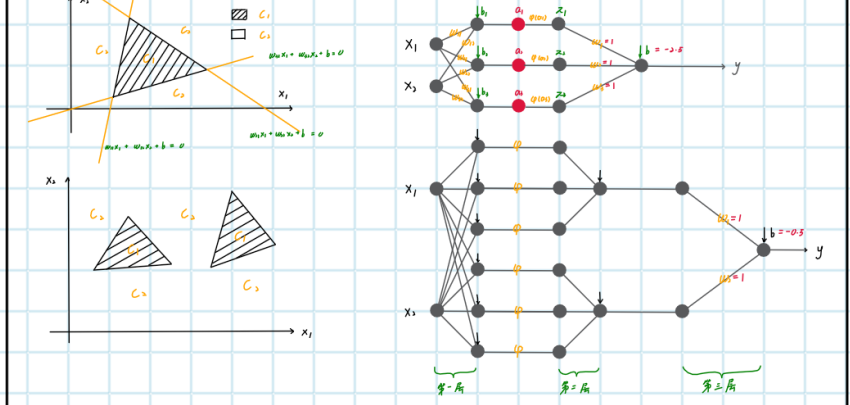
集错误率 (Equal Error Rate):



用于判别系统好坏



定理：三层神经网络可以模拟所有的决策面。



感知器算法 (Perceptron Algorithm):

- ① 取一个初始 w, b
- ② 取一个训练样本 (X, Y)
 - ③ 若 $w^T x + b > 0$ 且 $y = -1$, 则

$$W = w - x \quad b = b - 1$$
 - ④ 若 $w^T x + b < 0$ 且 $y = +1$, 则

$$W = w + x \quad b = b + 1$$
- ⑤ 再取一个 (X, Y) , 回到 ②
- ⑥ 终止条件: 直到所有输入输出对都不满足 ③ 中的 w 和 b 又 -1 , 则退出循环。

数学证明感知器算法的有效性:

定义一个滑移向量 \bar{x} :
 ① 若 $y = +1$, 则 $\bar{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$
 ② 若 $y = -1$, 则 $\bar{x} = \begin{bmatrix} x \\ -1 \end{bmatrix}$

定义一个滑移的 w :

$$w = \begin{bmatrix} w \\ b \end{bmatrix}$$

输入 \bar{x}_i

- ① 随机取 w
- ② 挑一个 \bar{x}_i
- ③ 若 $w^T \bar{x}_i < 0$, 则 $w = w + \bar{x}_i$
- ④ 返回 ②, 直到对所有的 \bar{x}_i 都不成立

感知器收敛定理:

输入 $\{\bar{x}_i\}_{i=1}^N$, 若线性可分, 即 $\exists w$, 使 $w_{opt}^T \bar{x}_i > 0 \quad (i=1 \sim N)$

则利用上述感知器算法, 经过有限步后, 得到 w , 使 $w^T \bar{x}_i > 0 \quad (i=1 \sim N)$

证明: 不失一般性, 设 $\|w_{opt}\| = 1$ (用 w_{opt} 与 αw_{opt} 是同一个平面)

假设第 k 步的 w 是 $w(k)$, 且有一个 \bar{x}_i , 使 $w(k)^T \bar{x}_i < 0$

根据感知器算法:

$$w(k+1) = w(k) + \bar{x}_i \Rightarrow \|w(k+1) - \alpha w_{opt}\|^2 = \|w(k) + \bar{x}_i - \alpha w_{opt}\|^2$$

$$= \|(w(k) - \alpha w_{opt}) + \bar{x}_i\|^2$$

$$= \|w(k) - \alpha w_{opt}\|^2 + \|\bar{x}_i\|^2 + 2w(k)^T \bar{x}_i - 2\alpha w_{opt}^T \bar{x}_i$$

一定可以取一个很大的 α , 使:

$$\|w(k+1) - \alpha w_{opt}\|^2 < \|w(k) - \alpha w_{opt}\|^2$$

定义 $\beta = \max_{i=1 \sim N} \|\bar{x}_i\|$, $\gamma = \min_{i=1 \sim N} (w_{opt}^T \bar{x}_i)$

取 $\alpha = \frac{\beta^2 + 1}{2\gamma}$, 则:

$$\|w(k+1) - \alpha w_{opt}\|^2 < \|w(k) - \alpha w_{opt}\|^2 - 1$$

取 $D = \|w(1) - \alpha w_{opt}\|^2$, 则最多经过 D 步, w 将收敛到 αw_{opt}

