支持向量机从入门到入门

Apple Zhang

Shenzhen University

2022年5月6日

目录

目录

深圳大学 SHENZHEN UNIVERSITY

符号定义

- 第i个训练样本: $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(d)}]^T \in \mathbb{R}^d$.
- 第i个训练样本的标签: $y_i \in \{-1, 1\}$.
- 训练样本矩阵: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$.
- 训练样本标签向量: $\mathbf{y} = [y_1, y_2, \cdots, y_n] \in \{-1, 1\}^n$.
- 超平面的法向量: $\mathbf{w} \in \mathbb{R}^d$.
- 超平面的偏移: $b \in \mathbb{R}$
- L_2范数:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}.$$

- 损失函数: $\mathcal{L}(\cdot)$.
- 拉格朗日函数: $L(\cdot)$.
- 求梯度/求导: ∇f.

深圳大学 SHENZHEN UNIVERSITY

什么是超平面?

- 在$\mathbb{R}^2$空间中，一条直线可以表示为

$$w_1 x^{(1)} + w_2 x^{(2)} + b = 0.$$

- 在$\mathbb{R}^3$空间中，一个平面可以表示为

$$w_1 x^{(1)} + w_2 x^{(2)} + w_3 x^{(3)} + b = 0.$$

- 在$\mathbb{R}^d$空间中，一个超平面可以表示为

$$w_1 x^{(1)} + w_2 x^{(2)} + \cdots + w_d x^{(d)} + b = 0.$$

即：

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

数据点到超平面的距离

- 在 $\mathbb{R}^2$ 空间中，数据点 $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}]^T$ 到一条直线的距离：

$$\frac{|w_1 x_i^{(1)} + w_2 x_i^{(2)} + b|}{\sqrt{w_1^2 + w_2^2}}.$$

- 在 $\mathbb{R}^3$ 空间中，数据点 $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, x_i^{(3)}]^T$ 到一个平面的距离：

$$\frac{|w_1 x_i^{(1)} + w_2 x_i^{(3)} + w_2 x_i^{(3)} + b|}{\sqrt{w_1^2 + w_2^2 + w_3^2}}.$$

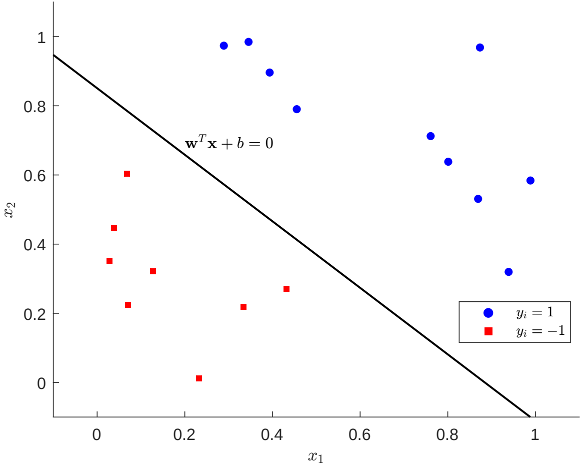
- 在 $\mathbb{R}^d$ 空间中，数据点 $\mathbf{x}_i = [x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(d)}]^T$ 到超平面的距离：

$$\frac{|w_1 x_i^{(1)} + w_2 x_i^{(2)} + \cdots + w_d x_i^{(d)} + b|}{\sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}}$$

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

分类问题——线性可分

问题:
对于一个二分类问题，假设数据是线性可分的 (linear separatable)，如何找到一个超平面把两类数据分开?



关键词:

- 超平面: $\mathbf{w}^T\mathbf{x} + b = 0$.
- 预测函数: $h(\mathbf{x}) = \text{sgn}(\mathbf{w}^T\mathbf{x} + b)$.
- 线性可分: 存在至少一个超平面，使得 $\forall i, y_i = h(\mathbf{x}_i)$.

目录

以减少错误为目的: 感知机

感知机 (Percetron) 减少每个样本分类的错误.

- 损失函数:

$$\ell(\mathbf{x}, y; \mathbf{w}, b) = \max(0, -y(\mathbf{w}^T\mathbf{x} + b)) \tag{1}$$

$$= \begin{cases} -y(\mathbf{w}^T\mathbf{x} + b), & \text{if } y(\mathbf{w}^T\mathbf{x} + b) < 0, \text{ misclassification}, \\ 0, & \text{if } y(\mathbf{w}^T\mathbf{x} + b) \geq 0, \text{ correct}, \end{cases} \tag{2}$$

- 优化目标:

$$(\hat{\mathbf{w}}, \hat{b}) = \arg\min_{\mathbf{w}, b} J(\mathbf{w}, b) \tag{3}$$

其中 J 为目标函数:

$$J(\mathbf{w}, b) = \sum_{i=1}^{n} \ell(\mathbf{x}_i, y_i; \mathbf{w}, b) \tag{4}$$

- 优化方法: 梯度下降

以减少错误为目的: 感知机

注意到只有在分类错误时损失不为 0, 此时梯度才非 0:

$$\nabla_{\mathbf{w}}\ell = -y_i\mathbf{x}_i, \quad \nabla_b\ell = -y_i, \quad \text{only if :} y_i(\mathbf{w}^T\mathbf{x}_i + b) < 0. \tag{5}$$

感知机 (Percetron) 减少每个样本分类的错误. 目标函数的梯度:

$$\nabla_{\mathbf{w}}J = \sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} -y_i\mathbf{x}_i, \tag{6}$$

$$\nabla_b J = \sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} -y_i. \tag{7}$$

设学习率为 η，则梯度下降描述为:

$$\mathbf{w} := \mathbf{w} - \eta\nabla_{\mathbf{w}}J, \quad b := b - \eta\nabla_b J \tag{8}$$

感知机有什么缺点?

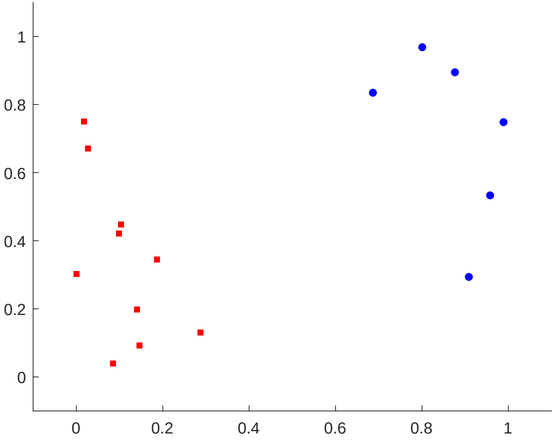


Figure: 这是一个二维线性可分数据集

感知机有什么缺点?

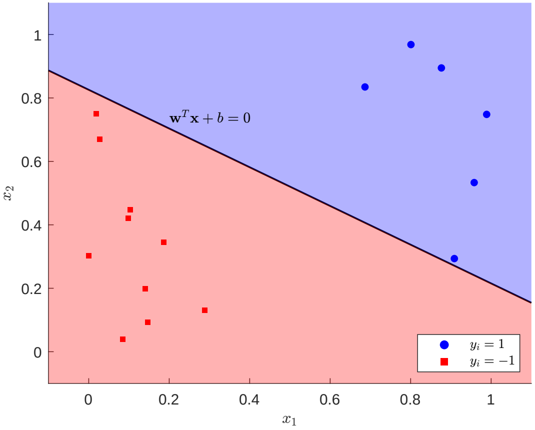
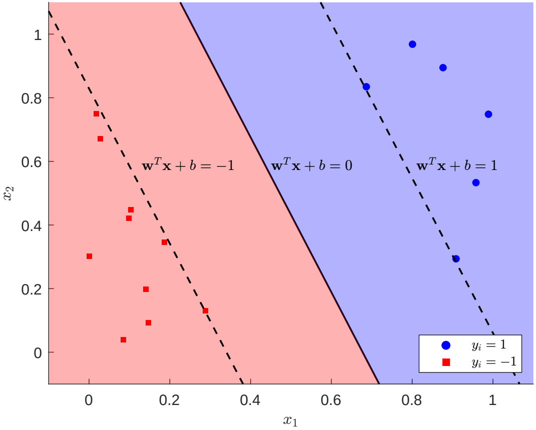


Figure: 感知机的这个分类结果好吗

目录

最大间隔分类器: 支持向量机

SVM 寻找最大化两个类间的间隔的超平面[1].



SVM 的优化目标

定义 (间隔)

样本集合到超平面 $\mathbf{w}^T\mathbf{x} + b = 0$ 的间隔，定义为其到超平面的最小距离:

$$\rho = \min_{i=1,2,\cdots,n} \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{\|\mathbf{w}\|_2} = \min_{i=1,2,\cdots,n} \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|_2} \tag{9}$$

线性可分时，考虑超平面 $\mathbf{w}^T\mathbf{x} + b = 1$ 和 $\mathbf{w}^T\mathbf{x} + b = -1$, SVM 希望所有样本都恰好被分到这两个超平面以外:

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq 1, & \text{if } y_i = 1, \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1, & \text{if } y_i = -1. \end{cases} \Leftrightarrow y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1. \tag{10}$$

“恰好” 意味着两类数据至少有一个样本满足 $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$, 则说明两类到超平面 $\mathbf{w}^T\mathbf{x} + b = 0$ 的间隔均为:

$$\rho = \frac{1}{\|\mathbf{w}\|_2}.$$

深圳大学 SHENZHEN UNIVERSITY

SVM 的优化问题

要最大化两个类之间的间隔, 则有

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|_2}, \quad \text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1. \tag{11}$$

由于

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|_2} \Leftrightarrow \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2.$$

所以最终得到如下的优化问题

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ s.t. } \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1.$$

SVM 的优化问题

该问题带有**不等式约束**，如何求解？

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \text{ s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, n.$$

带不等式约束的优化问题

我们在有等式约束的时候是怎么做的？

$$\min_{\mathbf{a}} \mathbf{a}^T\mathbf{S}\mathbf{a}, \quad \text{s.t. } \mathbf{a}^T\mathbf{a} = 1. \tag{12}$$

定义其**拉格朗日函数**为:

$$L(\mathbf{a}, \lambda) = \mathbf{a}^T\mathbf{S}\mathbf{a} + \lambda(1 - \mathbf{a}^T\mathbf{a}). \tag{13}$$

求导令 $\nabla_{\mathbf{a}} L = 0$ 得到

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}. \tag{14}$$

深圳大学 SHENZHEN UNIVERSITY

带不等式约束的优化问题

考虑一个带不等式约束的优化问题:

$$\begin{aligned} &\min_{\mathbf{u}} f_0(\mathbf{u}), \\ &\text{s.t. } f_i(\mathbf{u}) \le 0,\ i = 1, 2, \cdots, n. \end{aligned} \tag{15}$$

定义 (15) 对应的拉格朗日函数为

$$L(\mathbf{u}, \boldsymbol{\alpha}) = f_0(\mathbf{u}) + \sum_{i=1}^{n} \alpha_i f_i(\mathbf{u}). \tag{16}$$

分析

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \quad \text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, n.$$

SVM 对应的拉格朗日函数:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)]. \tag{17}$$

定理

求解 *SVM* 优化问题等价于求解以下的**对偶问题***:*

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}), \quad \text{s.t. } \boldsymbol{\alpha} \geq \mathbf{0} \tag{18}$$

分析

对于原问题:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \quad \text{s.t.} \;\; 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \le 0, \quad i = 1, 2, \cdots, n.$$

我们首先构造一个辅助函数$\theta(\mathbf{w}, b)$，且满足

$$\theta(\mathbf{w}, b) = \begin{cases} \frac{1}{2}\|\mathbf{w}\|_2^2 & \forall i,\; 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \le 0, \\ \infty & \exists i_0,\; 1 - y_{i_0}(\mathbf{w}^T\mathbf{x}_{i_0} + b) > 0, \end{cases}$$

这样构造的原因是:

$$\begin{aligned} & \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \; \text{s.t.} \;\; y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge 1 \\ \Leftrightarrow & \min_{\mathbf{w},b} \theta(\mathbf{w}, b). \end{aligned}$$

由此消去了优化问题中的不等式约束.

分析

$$\theta(\mathbf{w}, b)=\begin{cases}\frac{1}{2}\|\mathbf{w}\|_2^2 & \forall i,\ 1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\leq 0,\\ \infty & \exists i_0,\ 1-y_{i_0}(\mathbf{w}^T\mathbf{x}_{i_0}+b)>0.\end{cases}$$

接下来将说明:

$$\theta(\mathbf{w}, b)=\max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\alpha}),\ \text{s.t. } \boldsymbol{\alpha}\geq\mathbf{0}.$$

事实上, 设

$$h(\mathbf{w}, b)=\max_{\boldsymbol{\alpha}}\ \frac{1}{2}\|\mathbf{w}\|_2^2+\sum_{i=1}^{n}\alpha_i\left[1-y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)\right],\ \text{s.t. } \alpha\geq\mathbf{0}.$$

- $\forall i,\ 1-y_i(\mathbf{w}^T\mathbf{x}_i+b)\leq 0\Rightarrow h(\mathbf{w}, b)=\frac{1}{2}\|\mathbf{w}\|_2^2.$
- $\exists i_0,\ 1-y_{i_0}(\mathbf{w}^T\mathbf{x}_{i_0}+b)>0\Rightarrow h(\mathbf{w}, b)=\infty.$

因此, $\theta(\mathbf{w}, b)=h(\mathbf{w}, b)$.

分析

记录一下，我们现在得到了：

$$
\begin{aligned}
&\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \ \text{s.t.} \ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \\
\Leftrightarrow &\min_{\mathbf{w},b} \theta(\mathbf{w}, b) \\
\Leftrightarrow &\min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}\geq\mathbf{0}} L(\mathbf{w}, b, \boldsymbol{\alpha}).
\end{aligned}
$$

回顾一下我们证明的目标是：

$$
\max_{\alpha\geq\mathbf{0}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2, \ \text{s.t.} \ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1.
$$

事实上，我们有：

$$
\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}\geq\mathbf{0}} L(\mathbf{w}, b, \boldsymbol{\alpha}). \tag{19}
$$

该性质称为强对偶性.

Slater 定理

Slater 条件是原问题满足强对偶性的充分条件:

定理 (Slater 条件)

设定义在 $\mathcal{D}$ 上的函数 $f_i(\cdot), i = 0, 1, \cdots, n$ 均为凸函数，对于优化问题

$$\min_{\mathbf{u}} f_0(\mathbf{u}), \quad \text{s.t. } f_i(\mathbf{u}) \leq 0.$$

如果存在点 $\tilde{\mathbf{u}} \in \mathbf{relint}\mathcal{D}$ （即定义域 $\mathcal{D}$ 的相对内点）, 使得 $f_i(\tilde{\mathbf{u}}) < 0$, 则强对偶性成立.

分析

求解原问题的对偶问题:

$$\max_{\alpha \geq \mathbf{0}} \min_{\mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \alpha_i \left[1 - y_i \left(\mathbf{w}^T \mathbf{x}_i + b\right)\right].$$

首先对于 $\mathbf{w}, b$ 最小化 $L(\mathbf{w}, b, \boldsymbol{\alpha})$:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \quad \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0.$$

得到了如下的二次规划问题:

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

分析

化简为矩阵形式:

$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}, \quad \text{s.t. } \alpha_i \geq 0,\ \mathbf{y}^T\boldsymbol{\alpha} = 0.$$

这里 $\mathbf{1}$ 是全1的列向量, $\mathbf{Q} \in \mathbb{R}^{n\times n}$的每一个元素是

$$\mathbf{Q}_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j.$$

实际实现中可以调用 Matlab 的函数: `quadprog`[1], 对于Python, 可以使用`cvxopt` 中的 `solvers.qp`[2] 函数.
最后，我们可以把预测函数写为

$$h(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^T\mathbf{x} + b) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b\right).$$

[1] 需要 Matlab Optimization toolbox
[2] 需要 numpy-mkl

小贴士 (支持向量)

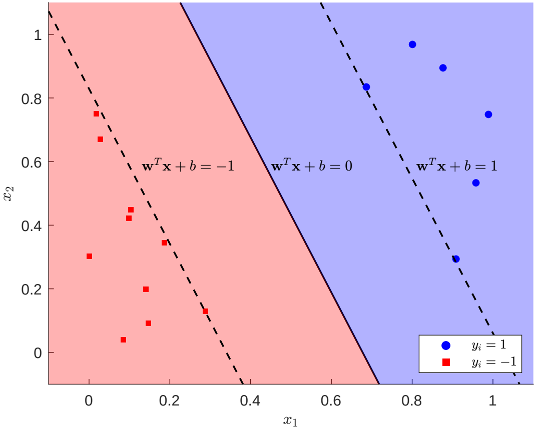
$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}, \quad \text{s.t. } \alpha_i \geq 0,\ \mathbf{y}^T\boldsymbol{\alpha} = 0.$$

- 上式解出的α是稀疏的. 如果 $\alpha_i > 0$, 则其对应的训练样本 $\mathbf{x}_i$ **称为支持向量(Support Vector**, SV).
- b可以用下面的公式算出:

$$b = \frac{1}{\|\alpha\|_0} \sum_{\alpha_i > 0} \left(\frac{1}{y_i} - \sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j^T \mathbf{x}_i \right).$$

其中$\|\cdot\|_0$表示L_0范数算子, 返回向量非零元素个数.

小贴士 (支持向量)



$\mathbf{w}^T\mathbf{x} + b = -1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = 1$
$y_i = 1$
$y_i = -1$
x_1
x_2
0
0.2
0.4
0.6
0.8
1

小贴士 (二次规划问题)

$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}, \quad \text{s.t. } \boldsymbol{\alpha} \geq \mathbf{0},\ \mathbf{y}^T\boldsymbol{\alpha} = 0. \tag{20}$$

Matlab函数: `x = quadprog(H, f, A, b, Aeq, Beq, lb, ub, x0);`

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{f}^T\mathbf{x}, \text{ s.t. } \begin{cases} \mathbf{A}\cdot\mathbf{x} \leq \mathbf{b} \\ \mathbf{Aeq}\cdot\mathbf{x} = \mathbf{beq} \\ lb \leq \mathbf{x} \leq ub \end{cases}$$

因此对偶问题可以按以下方式求解:

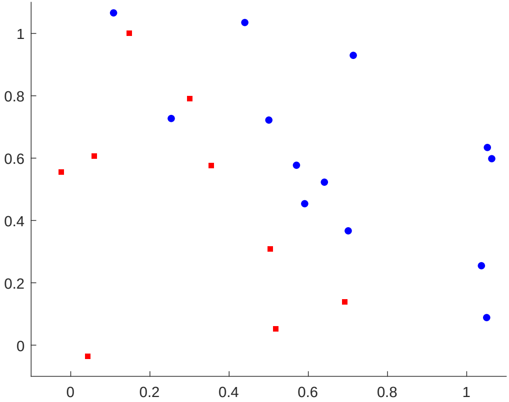
$$\texttt{alpha = quadprog}(\mathbf{Q},\ -\mathbf{1},\ \texttt{[]},\ \texttt{[]},\ \mathbf{y}^T,\ 0,\ \mathbf{0},\ \texttt{[]},\ \mathbf{x}_0)$$

注意: 需要安装Matlab Optimization toolbox.

目录

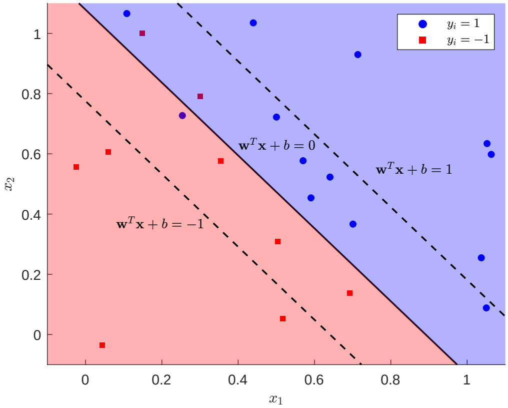
不是完美的线性可分?

前面讲的 **硬间隔 SVM** 无法用于这类数据.



不是完美的线性可分?

因此我们需要**软间隔 SVM**.
事实上，软间隔 SVM 才是我们平常用的 SVM. 不作特殊说明，一般讲到 SVM 模型就是指的软间隔 SVM.



软间隔SVM

软间隔 SVM 的优化问题:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|_2^2 + c\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$$

其中 c 是惩罚参数. 上述问题的拉格朗日函数:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + c\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\left[\alpha_i\left(1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right)\right] - \sum_{i=1}^{n}\beta_i\xi_i. \tag{21}$$

对偶问题:

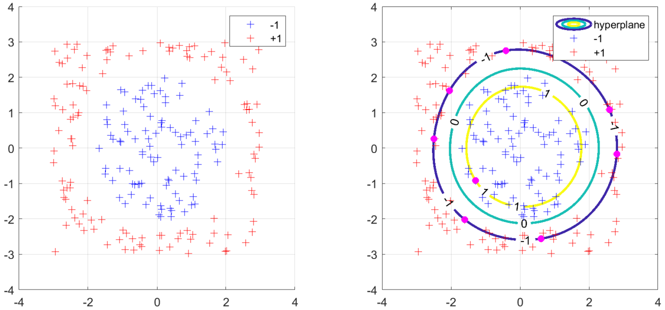
$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha}, \quad \text{s.t. } 0 \leq \alpha_i \leq c,\ \mathbf{y}^T\boldsymbol{\alpha} = 0.$$

目录

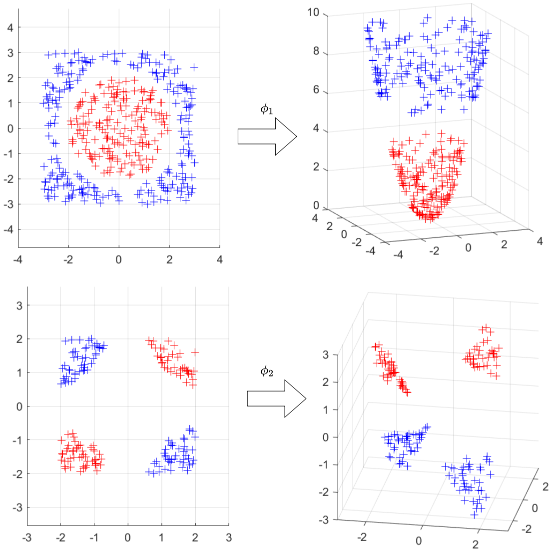
深圳大学 SHENZHEN UNIVERSITY

非线性分类: 核 SVM

前面使用的都是线性模型. 如果数据是非线性分布的, 则效果会较差.
核 SVM: 通过核方法完成非线性分类.



非线性分类: 核 SVM



核技巧

定义 (正定对称核)

定义函数 $\mathcal{K}:\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}$. 若对任意在 $\mathbb{R}^d$ 上的数据集: $\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$, 矩阵 $\mathcal{K}\in\mathbb{R}^{n\times n}$ 满足 $K_{ij}=\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)$, 且 $\mathcal{K}$ 是对称正定的, 则称 $\mathcal{K}$ 是正定对称核.

对正定对称核, 存在唯一的特征映射 $\phi:\mathbb{R}^d\mapsto\mathcal{H}$, 使得

$$\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)=\langle\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\rangle=\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j). \tag{22}$$

典型的例子是 RBF 核函数:

$$\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)=\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{t}\right),\quad t>0. \tag{23}$$

核函数可以隐式确定一个非线性映射 $\phi(\cdot)$, 于是不需要知道 $\phi(\cdot)$ 的具体表达式就可以计算非线性映射后的内积.

核 SVM

对于经过非线性映射后的数据 $\phi(\mathbf{x})$, 运用软间隔 SVM, 求以下超平面:

$$\mathbf{w}^T\phi(\mathbf{x}) + b = 0. \tag{24}$$

对应的 SVM 优化问题是:

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 + c\sum_{i=1}^{n}\xi_i, \\ &\text{s.t. } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0. \end{aligned} \tag{25}$$

实际上, 只是将软间隔 SVM 中的 $\mathbf{x}_i$ 都替换成了 $\phi(\mathbf{x}_i)$.

核技巧

拉格朗日函数:

$$L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|_{\mathcal{H}}^2 + c\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\left[\alpha_i\left(1 - \xi_i - y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b)\right)\right] - \sum_{i=1}^{n}\beta_i\xi_i. \tag{26}$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \phi(\mathbf{x}_i),$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0,$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow c - \alpha_i - \beta_i = 0.$$

核方法

对偶问题:

$$\max_{\boldsymbol{\alpha}} \mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}^{\Phi} \boldsymbol{\alpha},$$
$$\text{s.t. } 0 \leq \alpha_i \leq c,\ \mathbf{y}^T \alpha = 0.$$

其中 $\mathbf{Q}^{\Phi} \in \mathbb{R}^{n \times n}$ 定义为:

$$\mathbf{Q}_{ij}^{\Phi} = y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) = y_i y_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j).$$

注意: 在核 SVM 中, 测试样本的标签需要通过以下方式计算:

$$h(\mathbf{x}) = \text{sgn}(\mathbf{w}^T \phi(\mathbf{x}) + b) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b\right).$$

目录

更多有关SVM...

高效的SVM解决方案(C++实现): **LIBSVM, LIBLINEAR**[2, 3].

- 已有C++, Java, MATLAB, Python接口.
- 贼快.

其他SVM的变种:

- (1999 NPL) Least Square Support Vector Machine (LSSVM)[4].
- (2001 JMLR) Crammar-Singer Support Vector Machine (CSSVM)[5].
- (2006 JMLR) Laplacian Support Vector Machine (Lap-SVM)[6].
- (2007 PAMI) Twin Support Vector Machine (TWSVM)[7].
- (2019 NN) Capped-norm Twin Support Vector Machine (CTWSVM)[8].
- (2022 TCYB) Robust Manifold Twin Bounded SVM.

结束

谢谢!

https://github.com/Apple-Zhang/SVM-Intro

参考文献

[1] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, “Liblinear: A library for large linear classification,” *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[4] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[5] K. Crammer and Y. Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” *Journal of machine learning research*, vol. 2, no. Dec, pp. 265–292, 2001.

[6] M. Belkin, P. Niyogi, and V. Sindhwani, “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.,” *Journal of machine learning research*, vol. 7, no. 11, 2006.

[7] R. Khemchandani, S. Chandra, *et al.*, “Twin support vector machines for pattern classification,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[8] C. Wang, Q. Ye, P. Luo, N. Ye, and L. Fu, “Robust capped l1-norm twin support vector machine,” *Neural Networks*, vol. 114, pp. 47–59, 2019.

答疑