

我们准备开始了。同学们先不要刷屏了啊。

首先，模型的学习我建议三个方面相结合。公式、图示、代码。

机器学习基础模型的代码学习推荐两本书《机器学习实战》和《集体智慧编程》

下面按照昨天的思路，从线性回归->感知机->SVM->逻辑斯谛回归模型

## 一、线性回归的梳理

线性回归模型：

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \sum_{i=1}^n w^{(i)} \cdot x^{(i)} + b$$

其中， $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$  是输入， $\mathbf{w} = (w^{(1)}, w^{(2)}, \dots, w^{(n)})^\top \in \mathbb{R}^n$  和  $b \in \mathbb{R}$  是参数， $\mathbf{w}$  称为权值向量， $b$  称为偏置， $\mathbf{w} \cdot \mathbf{x}$  为  $\mathbf{w}$  和  $\mathbf{x}$  的内积。

线性回归的线性关系体现在权值与特征的线性关系上。

$$w^{(i)} \cdot x^{(i)}$$

损失函数：平方损失损失函数

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$

因为是个回归学习，损失函数采用平凡损失

模型参数最优解：

$$\hat{\mathbf{w}}^* = \arg \min_{\hat{\mathbf{w}}} \sum_{i=1}^N (y_i - f(\hat{\mathbf{x}}_i))^2$$

学习策略还是经验风险最小化

求解方法有两种方式：

令  $E_{\hat{\mathbf{w}}} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^\top (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$ ，对  $\hat{\mathbf{w}}$  求偏导，得

$$\frac{\partial E_{\hat{\mathbf{w}}}}{\partial \hat{\mathbf{w}}} = 2\mathbf{X}^\top (\mathbf{X}\hat{\mathbf{w}} - \mathbf{y})$$

当  $\mathbf{X}^\top \mathbf{X}$  为满秩矩阵或正定矩阵时，令上式为零可得最优闭式解

$$\hat{\mathbf{w}}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

求导，令为 0，解除闭式解。

周志华老师的《机器学习方法》中有关于闭式解的标量展开形式。

有兴趣的同学可以学习一下。

或者，采用迭代梯度法，计算梯度，延着梯度方向收敛到极小点。

岭回归 (Ridge Regression) 正则化项:

$$\alpha \|\mathbf{w}\|^2, \alpha \geq 0.$$

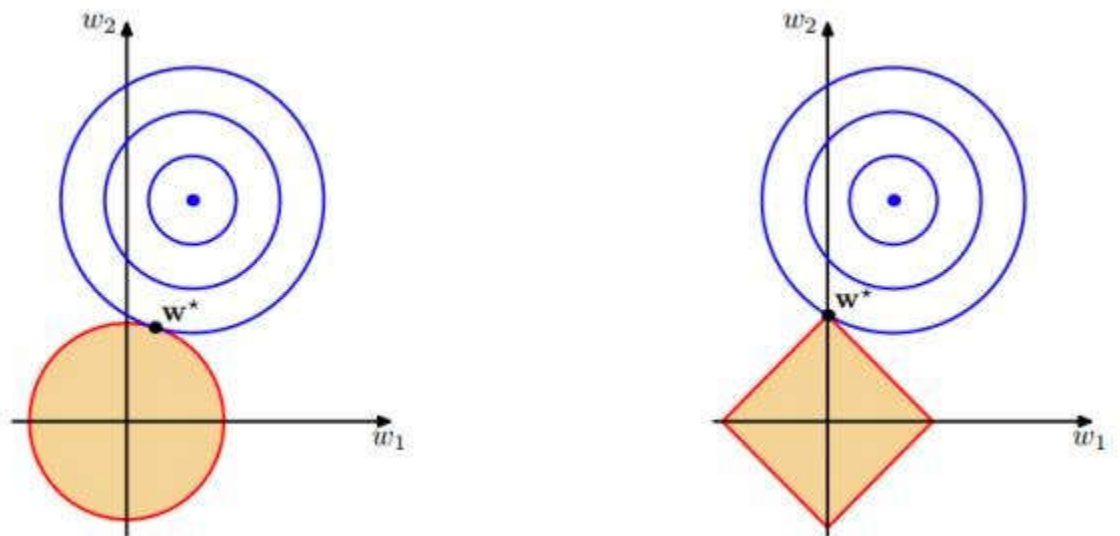
套索回归 (Lasso Regression) 正则化项:

$$\alpha \|\mathbf{w}\|_1, \alpha \geq 0.$$

弹性网络回归 (Elastic Net) 正则化项:

$$\alpha \rho \|\mathbf{w}\|_1 + \frac{\alpha(1-\rho)}{2} \|\mathbf{w}\|^2, \alpha \geq 0, 1 \geq \rho \geq 0.$$

可以在经验风险函数上加正则化项，构成结构风险函数。



说明一下稀疏解的情况。图中坐标  $w_1$   $w_2$  是权值，左图红色曲线是  $L_2$  正则项，右图红色曲线是  $L_1$  正则项。

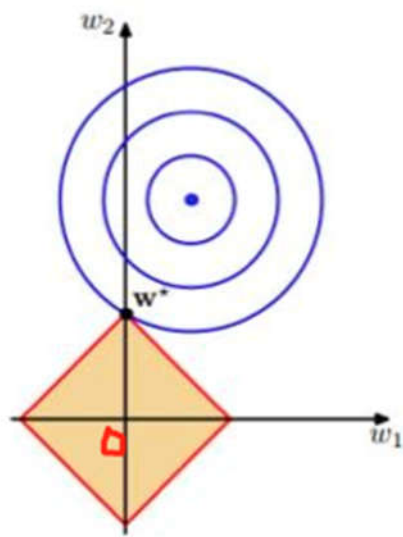
为什么一个是圈儿，一个是菱形，自己思考一下。

介绍一下蓝色圈圈，是风险函数的等高线表示方法。可以想象为一只碗。

圆心为碗底，外圈儿为碗延儿。

等高线和正则项相交的点，就是在结构风险最小化原则下的最优解。

可以看到  $L_1$  的交点在坐标轴上相交的可能性



这样的最优解在某些维度上的值为 0，就是所谓的稀疏解了  
比如上图在  $w_1$  这个维度上是 0  
所以有时候也用  $l_1$  做特征选择的一种方式。  
所谓特征选择就是在输入空间中选择部分有“代表”的部分特征。  
以上，是线性回归的梳理。

## 二、感知机梳理

假设输入空间 $\mathcal{X} \subseteq \mathbb{R}^n$ ，输出空间 $\mathcal{Y} = \{+1, -1\}$ 。输入 $\mathbf{x} \in \mathcal{X}$ 表示实例的特征向量，对应于输入空间的点；输出 $y \in \mathcal{Y}$ 表示实例的类别。由输入空间到输出空间的函数

$$f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

称为感知机。其中， $\mathbf{w}$ 和 $b$ 为感知机模型参数， $\mathbf{w} \in \mathbb{R}^n$ 叫做权值或权值向量， $b \in \mathbb{R}$ 叫偏置， $\mathbf{w} \cdot \mathbf{x}$ 表示 $\mathbf{w}$ 和 $\mathbf{x}$ 的内积。 $\text{sign}$ 是符号函数，即

$$\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

感知机是一种线性分类模型。感知机模型的假设空间是定义在特征空间中的所有线性分类模型或线性分类器，即函数集合 $\{f | f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b\}$ 。

$$(\mathbf{w} \cdot \mathbf{x} + b)$$

感知机模型是以线性回归的输出作为其输入

$$\text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$$

加入  $\text{sign}$  非线性映射函数

将实数值映射到 2 元值

$$\text{sign}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

输入空间 $\mathcal{X}$ 中的任一点 $\mathbf{x}_0$ 到超平面 $S$ 的距离：

$$\frac{1}{\|\mathbf{w}\|} |\mathbf{w} \cdot \mathbf{x}_0 + b|$$

其中 $\|\mathbf{w}\|$ 是权值向量 $\mathbf{w}$ 的 $L_2$ 范数。

对于误分类数据 $(\mathbf{x}_i, y_i)$ ，当 $\mathbf{w} \cdot \mathbf{x} + b > 0$ 时， $y_i = -1$ ， $(\mathbf{x}_i, y_i)$ ，当 $\mathbf{w} \cdot \mathbf{x} + b < 0$ 时， $y_i = +1$ ，则有 $-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) > 0$

误分类点 $\mathbf{x}_i$ 到分离超平面的距离：

$$-\frac{1}{\|\mathbf{w}\|} y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$$

假设超平面 $S$ 的误分类点集合为 $M$ ，则所有误分类点到超平面 $S$ 的总距离：

$$-\frac{1}{\|\mathbf{w}\|} \sum_{\mathbf{x}_i \in M} y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$$

感知机模型的重点内容

损失函数的设计

误分类数据

核心聚焦在对误分类点的处理

对于误分类数据 $(\mathbf{x}_i, y_i)$ , 当 $\mathbf{w} \cdot \mathbf{x} + b > 0$ 时,  $y_i = -1$ ,  $(\mathbf{x}_i, y_i)$ , 当 $\mathbf{w} \cdot \mathbf{x} + b < 0$ 时,  $y_i = +1$ , 则有

$$-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) > 0$$

对于误分类点的两种情况分别的讨论可以归纳出误分类点的判别公式, 即所有误分类点都满足

$$-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) > 0$$

有了以上的准备损失函数可表示

假设超平面 $S$ 的误分类点集合为 $M$ , 则所有误分类点到超平面 $S$ 的总距离:

$$-\frac{1}{\|\mathbf{w}\|} \sum_{\mathbf{x}_i \in M} y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$$

通过权值缩放, 可以使得权值的 2 范数为 1, 所以最后的损失函数描述

给定训练数据集

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$$

其中,  $\mathbf{x}_i \in \mathcal{X} = R^n, y_i \in \mathcal{Y} = \{+1, -1\}, i = 1, 2, \dots, N$ 。感知机 $sign(\mathbf{w} \cdot \mathbf{x} + b)$ 的损失函数定义为

$$L(\mathbf{w}, b) = - \sum_{\mathbf{x}_i \in M} y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$$

$$L(\mathbf{w}, b) = - \sum_{\mathbf{x}_i \in M} y_i (\mathbf{w} \cdot \mathbf{x}_i + b)$$

注意损失的累加是在误分类点集合中进行的。

有了损失函数后, 剩下的就还是老的套路, 求损失的极小化问题

假设误分类点集合 $M$ 是固定的, 则损失函数 $L(\mathbf{w}, b)$ 的梯度

$$\begin{aligned} \nabla_{\mathbf{w}} L(\mathbf{w}, b) &= - \sum_{\mathbf{x}_i \in M} y_i \mathbf{x}_i \\ \nabla_b L(\mathbf{w}, b) &= - \sum_{\mathbf{x}_i \in M} y_i \end{aligned}$$

随机选取一个误分类点 $(\mathbf{x}_i, y_i)$ , 对 $\mathbf{w}, b$ 进行更新:

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i \\ b &\leftarrow b + \eta y_i \end{aligned}$$

其中,  $\eta (0 < \eta \leq 1)$ 是步长, 称为学习率。

分别对  $\mathbf{w}$  和  $b$  求偏导, 得到梯度, 沿着梯度方向迭代求最小值。

感知机算法（原始形式）：

输入：训练数据集  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ ，其中  $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^n$ ,  $y_i \in \mathcal{Y} = \{+1, -1\}$ ,  $i = 1, 2, \dots, N$ ；学习率  $\eta$  ( $0 < \eta \leq 1$ )。

输出： $\mathbf{w}, b$ ；感知机模型  $f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$

1. 选取初值  $\mathbf{w}_0, b_0$
2. 在训练集中选取数据  $(\mathbf{x}_i, y_i)$
3. 如果  $y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \leq 0$

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_i$$

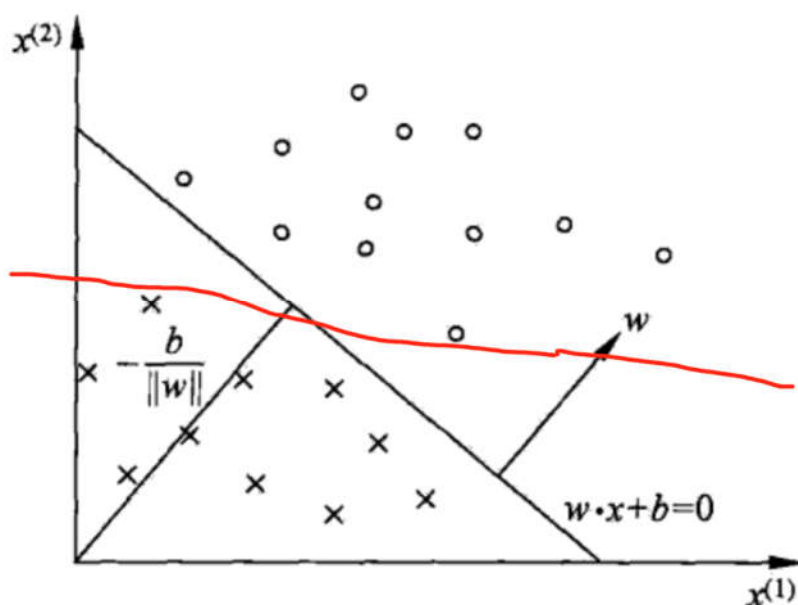
$$b \leftarrow b + \eta y_i$$

4. 转至2，直至训练集中没有误分类点。

在感知机算法中，初始值  $\mathbf{w}_0$  和  $b_0$  是随机设置的。另外，2. 在训练集中选取数据  $(\mathbf{x}_i, y_i)$

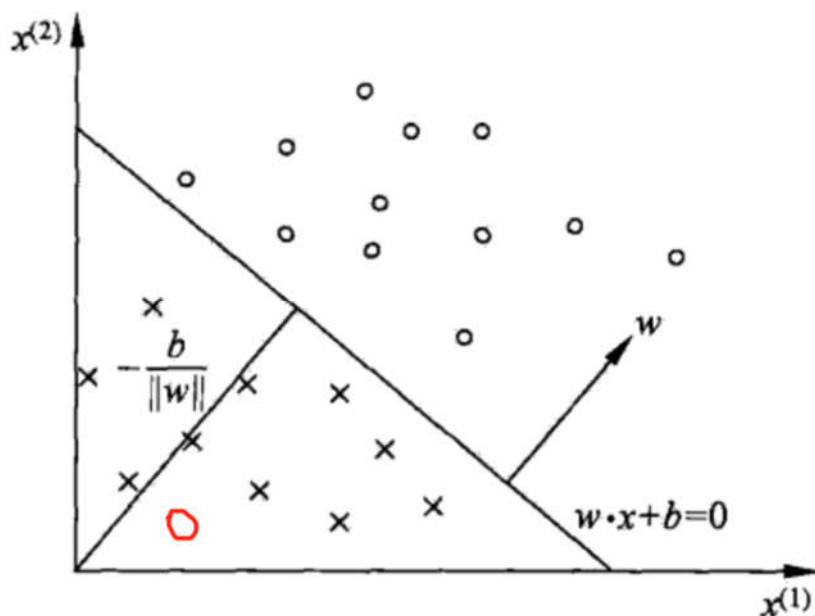
选取顺序也是随机的。

有以上两点，感知机算法在同一数据集上也可能得到不同的分类超平面。



可以看到感知机算存在多个“最优”解的结果。而且更重要的是感知机仅能处理数据本是线性可分的情况

如上图正负样本是可以被直线分开的，感知机算法是可以收敛，即找到分类边界的。



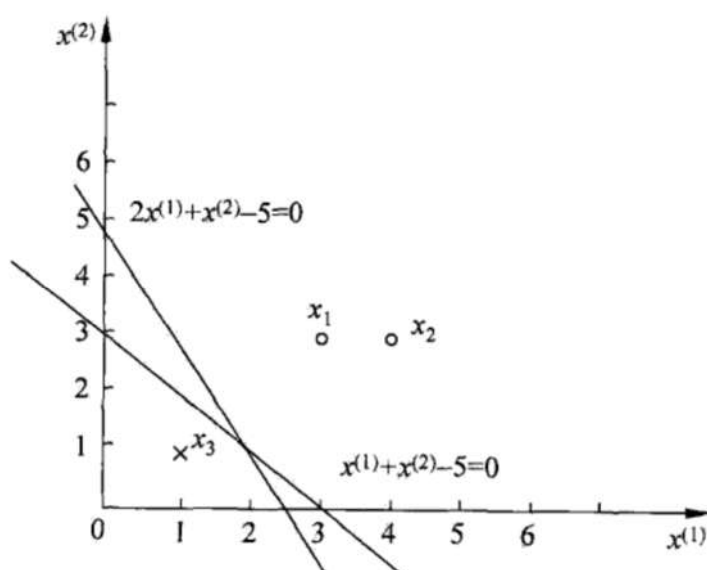
当数据本就不是线性可分的时候，可以根据感知机算法发现，算法会一直执行，长时间不能收敛。

这是感知机算法的最大局限

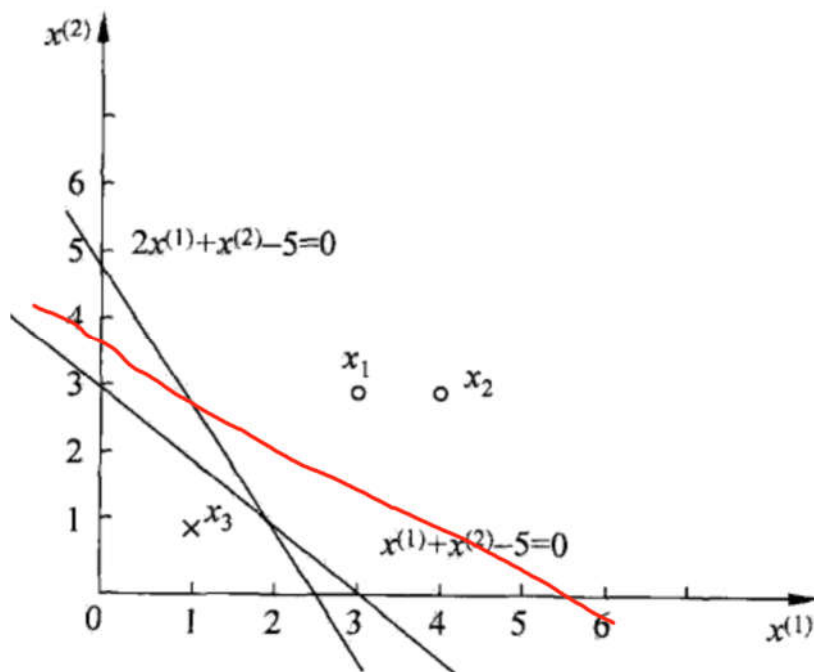
总结一下，感知机以误分类点为处理对象，可以完成线性可分分类。无法完成线性不可分分类。另外，分类边界也不足够“好”。

### 三、SVM 梳理

#### 1. 硬间隔 SVM

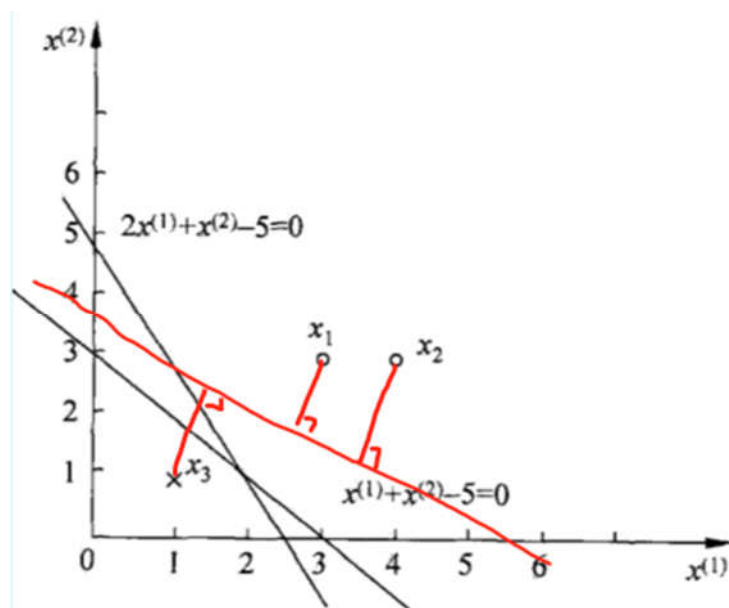


上图中两个分类边界都完成了分类，但是显然有更“好”的分类边界



大约啊，感受一下就好。

那我们可以把“好”进行准确的定义为数据集中最近的点距离分类边界也做够的远。



大家看到这样的分类超平面不仅完成了分类，而且使得样本离超平面还足够的远。当有新的样本时，分类就更可能会是正确的。



训练数据集

$$T = \left\{ \left( \mathbf{x}_1, y_1 \right), \left( \mathbf{x}_2, y_2 \right), \dots, \left( \mathbf{x}_N, y_N \right) \right\}$$

其中,  $\mathbf{x}_i \in X = \mathbb{R}^n, y_i \in Y = \{+1, -1\}, i = 1, 2, \dots, N$ ,  $\mathbf{x}_i$  为第  $i$  个特征向量 (实例),  $y_i$  为第  $\mathbf{x}_i$  的类标记, 当  $y_i = +1$  时, 称  $\mathbf{x}_i$  为正例; 当  $y_i = -1$  时, 称  $\mathbf{x}_i$  为负例,  $(\mathbf{x}_i, y_i)$  称为样本点。

线性可分支持向量机 (硬间隔支持向量机): 给定线性可分训练数据集, 通过间隔最大化或等价地求解相应地凸二次规划问题学习得到分离超平面为

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

以及相应的分类决策函数

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^* \cdot \mathbf{x} + b^*)$$

称为线型可分支持向量机。其中,  $\mathbf{w}^*$  和  $b^*$  为感知机模型参数,  $\mathbf{w}^* \in \mathbb{R}^n$  叫做权值或权值向量,  $b^* \in \mathbb{R}$  叫偏置,  $\mathbf{w}^* \cdot \mathbf{x}$  表示  $\mathbf{w}^*$  和  $\mathbf{x}$  的内积。sign 是符号函数。

所以会发现感知模型和 SVM 模型在模型形式上是一样的

间隔最大化或等价地求解相应地凸二次规划问题学习得

不同的是学习策略。感知机是要求误分类点距离最小, svm 是要求分对的点离超平面尽可能远

超平面  $(\mathbf{w}, b)$  关于样本点  $(\mathbf{x}_i, y_i)$  的几何间隔为

$$\gamma_i = y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right)$$

超平面  $(\mathbf{w}, b)$  关于训练集  $T$  的几何间隔

$$\gamma = \min_{i=1,2,\dots,N} \gamma_i$$

即超平面  $(\mathbf{w}, b)$  关于训练集  $T$  中所有样本点  $(\mathbf{x}_i, y_i)$  的几何间隔的最小值。

需要区分关于数据点的几何间隔和关于数据集的几何间隔的区别和联系

样本点的几何间隔就是点到分类平面距离。数据集的几何间隔是数据集中最近的样本点的几何间隔

有了上述定义后, SVM 的学习问题可表述为带约束条件的最优化问题

.. ..

最大间隔分离超平面等价求解

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N \end{aligned}$$

使得最近的样本距离分类平面尽可能的远

$$y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right)$$

这个距离公式也使用和感知机相同的分析过程，不同的是感知机是从误分类点角度，SVM 是从分类正确的点的角度进行归纳

同样可以引入函数间隔的概念，主要是为了进行最优化问题的转化。

最大间隔分离超平面等价求解

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{\hat{\gamma}}{\|\mathbf{w}\|} \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \hat{\gamma}, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

最大间隔分离超平面等价求解

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{\hat{\gamma}}{\|\mathbf{w}\|} \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \hat{\gamma}, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

数值上有相等的关系，就是个替换。

最大间隔分离超平面等价于求解

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\max \frac{1}{\|\mathbf{w}\|}$$

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{\hat{\gamma}}{\|\mathbf{w}\|} \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \hat{\gamma}, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

同样使用权值缩放，将  $\{\gamma\}$  缩放为 1

最大间隔分离超平面等价于求解

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \gamma \\ \text{s.t.} \quad & y_i \left( \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}_i + \frac{b}{\|\mathbf{w}\|} \right) \geq \gamma, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\max \frac{1}{\|\mathbf{w}\|}$$

$$\begin{aligned} \max_{\mathbf{w}, b} \quad & \frac{\hat{\gamma}}{\|\mathbf{w}\|} \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq \hat{\gamma}, \quad i = 1, 2, \dots, N \end{aligned}$$

等价的

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

分母的极大问题等价于分子的极小问题

平方不影响求极值，乘以 1/2 常数不影响求极值。

### 1. 构建并求解约束最优化问题

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

求得最优解  $\mathbf{w}^*, b^*$ 。

问题转化为以上形式。

通过拉格朗日乘子法求解。

拉个朗日乘子法今天就不做介绍了。有问题个别提问吧。

(硬间隔) 支持向量: 训练数据集的样本点中与分离超平面距离最近的样本点的实例, 即使约束条件等号成立的样本点

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 = 0$$

对  $y_i = +1$  的正例点, 支持向量在超平面

$$H_1: \mathbf{w} \cdot \mathbf{x} + b = 1$$

对  $y_i = -1$  的正例点, 支持向量在超平面

$$H_1: \mathbf{w} \cdot \mathbf{x} + b = -1$$

$H_1$  和  $H_2$  称为间隔边界。

$H_1$  和  $H_2$  之间的距离称为间隔, 且  $|H_1 H_2| = \frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$ 。

强调: 关于支持向量和间隔边界的定义是重点

(硬间隔) 支持向量: 训练数据集的样本点中与分离超平面距离最近的样本点的实例, 即使约束条件等号成立的样本点

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 = 0$$

对  $y_i = +1$  的正例点, 支持向量在超平面

$$H_1: \mathbf{w} \cdot \mathbf{x} + b = 1$$

对  $y_i = -1$  的正例点, 支持向量在超平面

$$H_1: \mathbf{w} \cdot \mathbf{x} + b = -1$$

$H_1$  和  $H_2$  称为间隔边界。

$H_1$  和  $H_2$  之间的距离称为间隔, 且  $|H_1 H_2| = \frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$ 。

支持向量是某些特殊的样本点。

间隔边界是两条直线

$$|H_1 H_2| = \frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

当做结论

最优化问题的求解:

1. 引入拉格朗日乘子  $\alpha_i \geq 0, i = 1, 2, \dots, N$  构建拉格朗日函数

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + 1] \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_{i=1}^N \alpha_i \end{aligned}$$

其中,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$  为拉格朗日乘子向量。

## 构建拉格朗日函数

最优化问题的求解：

1. 引入拉格朗日乘子  $\alpha_i \geq 0, i = 1, 2, \dots, N$  构建拉格朗日函数

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^N \alpha_i [-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + 1] \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + \sum_{i=1}^N \alpha_i \end{aligned}$$

其中,  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$  为拉格朗日乘子向量。

标记为求解目标

2. 求  $\min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha)$ :

$$\begin{aligned} \nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) &= \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0 \\ \nabla_b L(\mathbf{w}, b, \alpha) &= - \sum_{i=1}^N \alpha_i y_i = 0 \end{aligned}$$

得

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^N \alpha_i y_i &= 0 \end{aligned}$$

代入拉格朗日函数, 得

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i y_i \left[ \left( \sum_{j=1}^N \alpha_j y_j \mathbf{x}_j \right) \cdot \mathbf{x}_i + b \right] + \sum_{i=1}^N \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i y_i b + \sum_{i=1}^N \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i \end{aligned}$$

即

$$\min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i$$

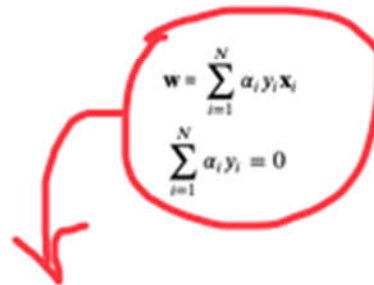
先求极小

2. 求  $\min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha)$ :

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0$$

$$\nabla_b L(\mathbf{w}, b, \alpha) = - \sum_{i=1}^N \alpha_i y_i = 0$$

得



$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

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

代入拉格朗日函数，得

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i y_i \left[ \left( \sum_{j=1}^N \alpha_j y_j \mathbf{x}_j \right) \cdot \mathbf{x}_i + b \right] + \sum_{i=1}^N \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i y_i b + \sum_{i=1}^N \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i \end{aligned}$$

即

$$\min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i$$

求偏导得到的两个公式带入到拉格朗日函数，  
化简整理

3. 求  $\max_{\alpha} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha)$ :

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

等价的

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i \\ s.t. & \sum_{i=1}^N \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, N \end{aligned}$$

再求极大

3. 求  $\max_{\alpha} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \alpha)$ :

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

等价的

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

负号提前，求极大等价求极小

求得最优解  $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)$ 。

2. 计算

$$\mathbf{w}^* = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i$$

并选择  $\alpha^*$  的一个正分量  $\alpha_j^* > 0$ ，计算

$$b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$$

3. 得到分离超平面

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

以及分类决策函数

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^* \cdot \mathbf{x} + b^*)$$

求出拉格朗日乘子，求出  $\mathbf{w}$  和  $b$ ，求出分类边界。  
以上是硬间隔 SVM

## 2. 软间隔 SVM

线性支持向量机（软间隔支持向量机）：给定线性不可分训练数据集，通过求解凸二次规划问题

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

软间隔在硬间隔基础上引入松弛变量，允许数据点距离分类边界的距离小于 1

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

同样适用拉格朗日乘子法求解

1. 引入拉格朗日乘子  $\alpha_i \geq 0, \mu_i \geq 0, i = 1, 2, \dots, N$  构建拉格朗日函数

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \mu) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i [-y_i (\mathbf{w} \cdot \mathbf{x}_i + b) + 1 - \xi_i] + \sum_{i=1}^N \mu_i (-\xi_i) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^N \mu_i \xi_i \end{aligned}$$

其中， $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^T$  以及  $\mu = (\mu_1, \mu_2, \dots, \mu_N)^T$  为拉格朗日乘子向量。

求极小



2. 求 $\min_{\mathbf{w}, b} L(\mathbf{w}, b, \xi, \alpha, \mu)$ :

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \xi, \alpha, \mu) = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0$$

$$\nabla_b L(\mathbf{w}, b, \xi, \alpha, \mu) = - \sum_{i=1}^N \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L(\mathbf{w}, b, \xi, \alpha, \mu) = C - \alpha_i - \mu_i = 0$$

得

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^N \alpha_i y_i &= 0 \\ C - \alpha_i - \mu_i &= 0 \end{aligned}$$

代入拉格朗日函数, 得

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \mu) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i y_i \left[ \left( \sum_{j=1}^N \alpha_j y_j \mathbf{x}_j \right) \cdot \mathbf{x}_i + b \right] + \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \alpha_i \xi_i - \sum_{i=1}^N \mu_i \xi_i \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i y_i b + \sum_{i=1}^N \alpha_i + \sum_{i=1}^N \xi_i (C - \alpha_i - \mu_i) \\ &= -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i \end{aligned}$$

即

$$\min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \mu) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i$$

仔细化简

求极大

3. 求 $\max_{\alpha} \min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \mu)$ :

$$\max_{\alpha} -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i$$

$$s. t. \sum_{i=1}^N \alpha_i y_i = 0$$

$$C - \alpha_i - \mu_i = 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0, \quad i = 1, 2, \dots, N$$

等价的

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i$$

$$s. t. \sum_{i=1}^N \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N$$

3. 求  $\max_{\alpha} \min_{w, b, \xi} L(w, b, \xi, \alpha, \mu)$ :

$$\max_{\alpha} -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^N \alpha_i$$

$$s. t. \sum_{i=1}^N \alpha_i y_i = 0$$

$$C - \alpha_i - \mu_i = 0$$

$$\alpha_i \geq 0$$

$$\mu_i \geq 0, \quad i = 1, 2, \dots, N$$

等价的

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^N \alpha_i$$

$$s. t. \sum_{i=1}^N \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N$$

注意这个条件是怎么得到的。仔细一些。

实例  $x_i$  的几何间隔

$$\gamma_i = \frac{y_i (w \cdot x_i + b)}{\|w\|} = \frac{|1 - \xi_i|}{\|w\|}$$

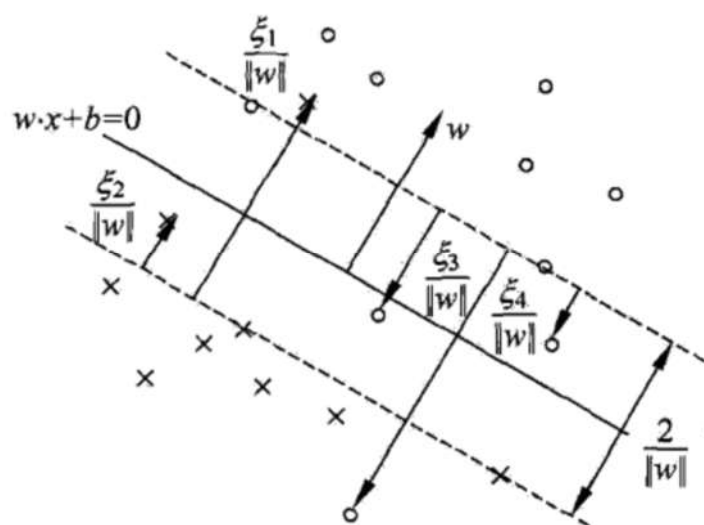
$$\text{且 } \frac{1}{2} |H_1 H_2| = \frac{1}{\|w\|}$$

则实例  $x_i$  到间隔边界的距离

$$\frac{\xi_i}{\|w\|}$$

$$\xi_i \geq 0 \Leftrightarrow \begin{cases} \xi_i = 0, x_i \text{ 在间隔边界上;} \\ 0 < \xi_i < 1, x_i \text{ 在间隔边界与分离超平面之间;} \\ \xi_i = 1, x_i \text{ 在分离超平面上;} \\ \xi_i > 1, x_i \text{ 在分离超平面误分类一侧;} \end{cases}$$

强调：对  $x_i$  的不同取值对应的数据的位置分析



此处忽略了对  $C$  值得讨论。《统计学习方法》中有详细讨论。可参考。  
以上是软间隔 SVM

### 3.非线性 SVM

核函数

设 $\mathcal{X}$ 是输入空间（欧氏空间 $\mathbb{R}^n$ 的子集或离散集合）， $\mathcal{H}$ 是特征空间（希尔伯特空间），如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

$$\phi(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{H}$$

使得对所有 $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ ，函数 $K(\mathbf{x}, \mathbf{z})$ 满足条件

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

则称 $K(\mathbf{x}, \mathbf{z})$ 为核函数， $\phi(\mathbf{x})$ 为映射函数，式中 $\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ 为 $\phi(\mathbf{x})$ 和 $\phi(\mathbf{z})$ 的内积。

核函数

设 $\mathcal{X}$ 是输入空间（欧氏空间 $\mathbb{R}^n$ 的子集或离散集合）， $\mathcal{H}$ 是特征空间（希尔伯特空间），如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

$$\phi(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{H}$$

使得对所有 $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ ，函数 $K(\mathbf{x}, \mathbf{z})$ 满足条件

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

则称 $K(\mathbf{x}, \mathbf{z})$ 为核函数， $\phi(\mathbf{x})$ 为映射函数，式中 $\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ 为 $\phi(\mathbf{x})$ 和 $\phi(\mathbf{z})$ 的内积。

\phi 函数是不同空间上的映射函数

将低维线性不可分数据映射到高维空间

$$\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

在高维空间做内积计算

核方法最大的优势在：

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

右面在高维空间做的内积计算与等号左面的在低维

空间的核函数的结果相等！

也就是说虽然要映射到高维空间再进行分类，但是！但是！但是！仅在低维空间做核函数计算就可以同样实现！！！！

常用核函数：

1. 多项式核函数

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z} + 1)^p$$

2. 高斯核函数

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

常用核函数

面试常问问题，为什么高斯核函数可以映射到无穷维空间？（没啥味道的面试题）请自行百度。

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N \end{aligned}$$

引入核函数后，所有内积计算都可以替换为核函数计算。

以上是非线性 SVM

SMO 自己学习

## 四、逻辑回归与 sigmoid 函数梳理

sigmoid函数：

$$\text{sigmoid}(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

其中,  $z \in \mathbb{R}$ ,  $\text{sigmoid}(z) \in (0, 1)$ 。

sigmoid函数的导数：

$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$

二项逻辑斯谛回归模型是如下的条件概率分布：

$$\begin{aligned} P(y = 1|\mathbf{x}) &= \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))} \\ &= \frac{\exp(\mathbf{w} \cdot \mathbf{x} + b)}{1 + \exp(\mathbf{w} \cdot \mathbf{x} + b)} \\ P(y = 0|\mathbf{x}) &= 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \frac{1}{1 + \exp(\mathbf{w} \cdot \mathbf{x} + b)} \end{aligned}$$

其中,  $\mathbf{x} \in \mathbb{R}^n$ ,  $y \in \{0, 1\}$ ,  $\mathbf{w} \in \mathbb{R}^n$ 是权值向量,  $b \in \mathbb{R}$ 是偏置,  $\mathbf{w} \cdot \mathbf{x}$ 为向量内积。

逻辑回归的重点就是 sigmoid 函数

将  $\mathbb{R}$  映射到  $(0, 1)$

完成分类任务

二项逻辑斯谛回归模型是如下的条件概率分布：

$$\begin{aligned} P(y = 1|\mathbf{x}) &= \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} + b))} \\ &= \frac{\exp(\mathbf{w} \cdot \mathbf{x} + b)}{1 + \exp(\mathbf{w} \cdot \mathbf{x} + b)} \\ P(y = 0|\mathbf{x}) &= 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \frac{1}{1 + \exp(\mathbf{w} \cdot \mathbf{x} + b)} \end{aligned}$$

其中,  $\mathbf{x} \in \mathbb{R}^n$ ,  $y \in \{0, 1\}$ ,  $\mathbf{w} \in \mathbb{R}^n$ 是权值向量,  $b \in \mathbb{R}$ 是偏置,  $\mathbf{w} \cdot \mathbf{x}$ 为向量内积。

同样以线性回归输出为  $s$  函数输入

将  $\mathbb{R}^n$  映射到  $(0, 1)$



0.5 作为阈值尽心分类标签的确定是正例还是负例

给定训练数据集

$$D = \{(\hat{\mathbf{x}}_1, y_1), (\hat{\mathbf{x}}_2, y_2), \dots, (\hat{\mathbf{x}}_N, y_N)\}$$

其中,  $\hat{\mathbf{x}}_i \in \mathbb{R}^{n+1}, y_i \in \{0, 1\}, i = 1, 2, \dots, N$ 。

设

$$P(y = 1|\hat{\mathbf{x}}) = \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}), \quad P(y = 0|\hat{\mathbf{x}}) = 1 - \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}})$$

似然函数

$$\begin{aligned} L(\hat{\mathbf{w}}) &= \prod_{i=1}^N P(y_i|\hat{\mathbf{x}}_i) \\ &= \prod_{i=1}^N [\sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}_i)]^{y_i} [1 - \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}_i)]^{1-y_i} \end{aligned}$$

对数似然函数

$$\begin{aligned} l(\hat{\mathbf{w}}) &= \log L(\hat{\mathbf{w}}) \\ &= \sum_{i=1}^N [y_i \log \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}_i) + (1 - y_i) \log(1 - \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}_i))] \end{aligned}$$

最大似然估计

$$\hat{\mathbf{w}}^* = \arg \max_{\hat{\mathbf{w}}} l(\hat{\mathbf{w}})$$

用-对数似然作为损失函数

$$l(\hat{\mathbf{w}}) = \log L(\hat{\mathbf{w}})$$

求-对数似然的极小

令  $\hat{y}_i = \sigma(\hat{\mathbf{w}} \cdot \hat{\mathbf{x}}_i)$ ，则对数似然函数  $l(\hat{\mathbf{w}})$  关于  $\hat{\mathbf{w}}$  的偏导数

$$\begin{aligned}\frac{\partial l(\hat{\mathbf{w}})}{\partial \hat{\mathbf{w}}} &= \sum_{i=1}^N \left( y_i \frac{\hat{y}_i (1 - \hat{y}_i)}{\hat{y}_i} \hat{\mathbf{x}}_i + (1 - y_i) \frac{\hat{y}_i (1 - \hat{y}_i)}{1 - \hat{y}_i} \hat{\mathbf{x}}_i \right) \\ &= \sum_{i=1}^N (y_i (1 - \hat{y}_i) \hat{\mathbf{x}}_i + (1 - y_i) \hat{y}_i \hat{\mathbf{x}}_i) \\ &= \sum_{i=1}^N \hat{\mathbf{x}}_i (y_i - \hat{y}_i)\end{aligned}$$

采用梯度上升法，初始化  $\hat{\mathbf{w}} = \mathbf{0}$ ，进行迭代

$$\hat{\mathbf{w}}_{t+1} \leftarrow \hat{\mathbf{w}}_t + \alpha \sum_{i=1}^N \hat{\mathbf{x}}_i (y_i - \hat{y}_i^{\hat{\mathbf{w}}_t})$$

其中， $\alpha$  是学习率， $\hat{y}_i^{\hat{\mathbf{w}}_t}$  是当参数  $\hat{\mathbf{w}}_t$  时模型的输出。

解出最优值。

以上是逻辑斯谛回归模型