支持向量机（SVM）

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

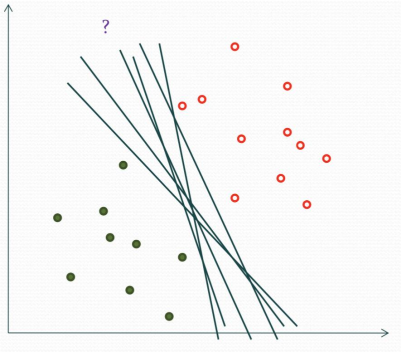
2023年春

课程目标

- 理解硬间隔最大化思想和线性可分SVM算法。
- 掌握拉格朗日对偶性的主要概念和结论，并灵活运用。
- 学会推导支持向量机的对偶优化问题。
- 理解支持向量的作用。

问题的提出

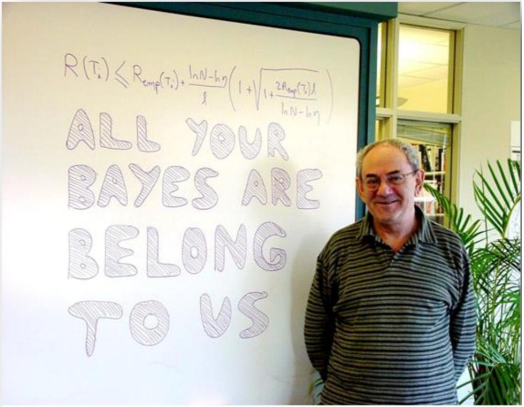
- 给定如图所示的一个线性可分的训练数据集，能将该数据集中的两类样本完全正确分开的直线有很多。
- 利用感知机算法可以找到一条直线，且找到的直线不是唯一的。图中每条直线都可能是感知机算法求得的直线。



- ✓ 哪一条直线最好?
- ✓ 感知机无法回答这个问题。
- ✓ SVM告诉我们，在所有能将训练样本正确分类的直线中**具有最大的间隔**的那条直线是最优的。

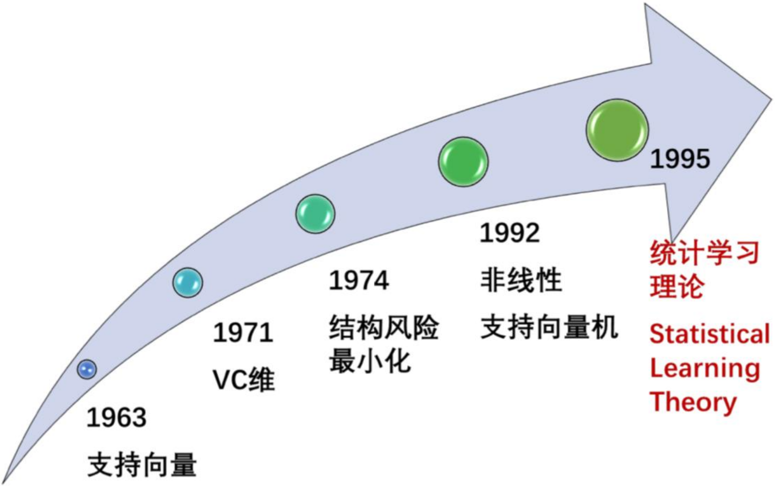
SVM简介

- 支持向量机算法是由俄罗斯著名的统计学家和数学家弗拉基米尔·瓦普尼克 (Vladimir Naumovich Vapnik)于1963 年提出的。



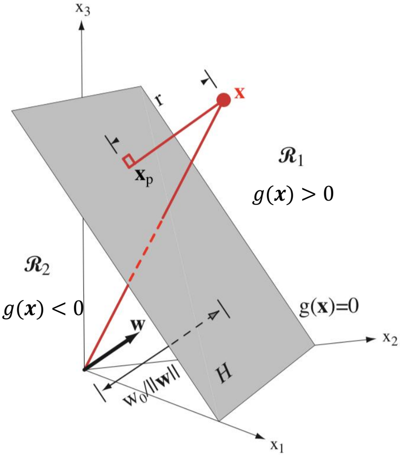
Vladimir N. Vapnik

SVM的发展历程


1963
支持向量
1971
VC维
1974
结构风险
最小化
1992
非线性
支持向量机
1995
统计学习
理论
Statistical
Learning
Theory

点到超平面的垂直距离

- 给定一个分割超平面$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b=0$，其对应的判别函数为$g(\boldsymbol{x})=\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b$，分割超平面将整个特征空间分为两个决策域：$g(\boldsymbol{x})>0$和$g(\boldsymbol{x})<0$。



- ✓ 点$\boldsymbol{x}$到超平面$\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b=0$的**垂直距离**为：

$$r=\frac{|g(\boldsymbol{x})|}{||\boldsymbol{w}||}=\frac{|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}+b|}{||\boldsymbol{w}||}$$

- ✓ r值的大小可以反应分类的确信程度。r值越小，点离决策超平面越近，对其分类的不确信程度就越高。极端情况r=0，此时无法确定分类。

间隔

- 给定一个训练数据集$\{(\boldsymbol{x}_1, y_1), \cdots (\boldsymbol{x}_N, y_N)\}$，其中$\boldsymbol{x}_i \in \mathbb{R}^d$，$y_i \in \{+1, -1\}$。$y_i = +1$对应的$\boldsymbol{x}_i$称为正类样本，$y_i = -1$对应的$\boldsymbol{x}_i$称为负类样本。
- **(间隔)**：超平面$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$关于样本点$(\boldsymbol{x}_i, y_i)$的间隔定义为：

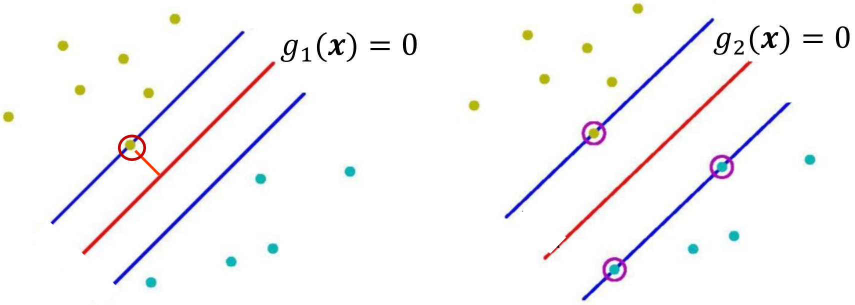
$$r_i = \frac{|\boldsymbol{w}^\top \boldsymbol{x}_i + b|}{||\boldsymbol{w}||}$$

超平面$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$关于训练数据集的间隔定义为：

超平面关于训练数据集中**所有样本点的间隔中的最小间隔**：

$$r_{min} = \min_{i=1,\cdots N} r_i = \min_{i=1,\cdots N} \frac{|\boldsymbol{w}^\top \boldsymbol{x}_i + b|}{||\boldsymbol{w}||}$$

Support Vector Machine (SVM)的思想

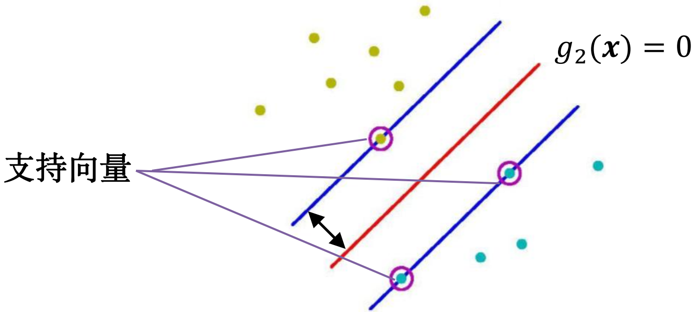


- SVM的思想是：寻找能够正确划分训练数据集并且**在训练数据集上具有最大间隔**的超平面。

✓ 左图中的间隔小于右图中的间隔，所以右图的超平面更优。

Support Vector Machine (SVM)的思想

- **间隔最大化**的直观解释：不仅能将训练样本正确分开，而且对于最难分的样本点(离超平面最近)，也能够以足够大的置信度对其分类。满足这些条件的超平面一定具有很好的泛化性能。



✓ 训练集中具有**最小间隔的样本点**称为"**支持向量**"。

SVM思想的形式化描述

- **假设训练数据是线性可分的。**
- 目标：找到分割超平面$\boldsymbol{w}^\top\boldsymbol{x}+b=0$，满足两个条件：

① 正确划分训练样本：$\forall i=1\cdots N$，都有$y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)>0$

② 最大化在训练集上的间隔：

$$\max_{\boldsymbol{w},b}\left(\min_{i=1\cdots N}\frac{|\boldsymbol{w}^\top\boldsymbol{x}_i+b|}{||\boldsymbol{w}||}\right)$$

- 满足条件①的超平面有无穷多个，同时满足①②的超平面只有一个。
- 公式简化：因为条件①，$|\boldsymbol{w}^\top\boldsymbol{x}_i+b|=y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)$

$$\max_{\boldsymbol{w},b}\left(\min_{i=1\cdots N}\frac{y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)}{||\boldsymbol{w}||}\right)=\max_{\boldsymbol{w},b}\left(\frac{1}{||\boldsymbol{w}||}\min_{i=1\cdots N}y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\right)$$

SVM思想的形式化描述

- 令 $\min_{i=1\cdots N} y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$，则有

$$\forall i, y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1$$

- 原优化问题等价于：

$$\max_{\boldsymbol{w},b} \frac{1}{||\boldsymbol{w}||}$$

$$s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \forall i = 1\cdots N$$

- 因为：

$$\underset{\boldsymbol{w}}{\operatorname{argmax}} \frac{1}{||\boldsymbol{w}||} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{2}||\boldsymbol{w}||^2$$

- 原优化问题等价于：

$$\boxed{\begin{gathered}\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \\ s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ \forall i = 1\cdots N\end{gathered}}$$

硬间隔最大化

凸二次规划

线性可分SVM的学习算法

- 输入：线性可分的训练数据集$\{(\boldsymbol{x}_1, y_1)\cdots(\boldsymbol{x}_N, y_N)\}$，其中$\boldsymbol{x}_i \in \mathbb{R}^d$，$y_i \in \{+1, -1\}$。
- 输出：最大间隔的分离超平面和分类决策函数。

1. 求解如下优化问题

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \quad s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1, \forall i = 1\cdots N$$

得到最优解$\boldsymbol{w}^*$，b^*;

2. 得到分割超平面：$\boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 0$

分类决策函数$f(\boldsymbol{x}) = sign(\boldsymbol{w}^{*\top}\boldsymbol{x} + b^*)$

硬间隔最大化的SVM学习算法

支持向量

- 训练集中具有**最小间隔的样本点**称为“**支持向量**”。
- 支持向量$\boldsymbol{x}_i$一定满足：$y_i(\boldsymbol{w}^{*\top}\boldsymbol{x}_i + b^*) = 1$
- 支持向量到决策超平面$\boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 0$的距离为：$\dfrac{1}{||\boldsymbol{w}^*||}$
- 对于正类($y_i = +1$)的支持向量在超平面

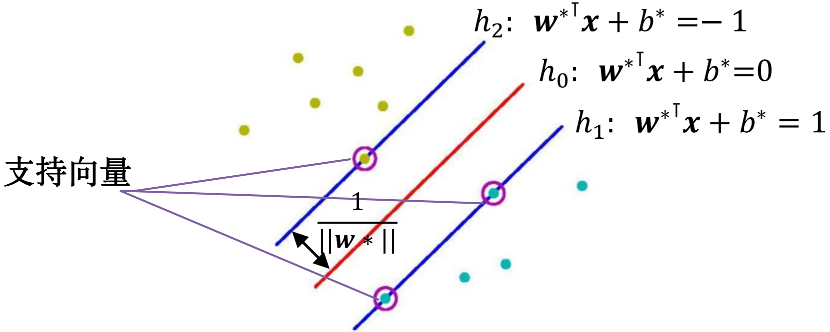
$$h_1:\ \boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 1$$

对于负类($y_i = -1$)的支持向量在超平面

$$h_2:\ \boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = -1$$

超平面h_1和h_2称为间隔边界。

支持向量



- 在决定分类决策超平面的位置时，只有支持向量起作用，其它样本点不起作用，即使删除这些样本点，也不影响分类决策边界。所以这种分类模型称为“**支持向量机**”。
- 在线性可分的数据集上，支持向量的数目一般很少。

SVM的对偶优化问题

- SVM的原始优化问题：

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

$$s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ \forall i = 1\cdots N$$

- 如何求解这个约束优化问题?
- 在约束最优化问题中，常常利用拉格朗日对偶性(Lagrange duality)将原始问题转换为对偶问题，通过解对偶问题得到原始问题的解。
- 这么做的好处是：

① 对偶问题更容易求解；

② 自然而言地引入核函数，进而将算法推广到求解非线性分类问题。

拉格朗日对偶性的主要概念和结论

- 假设$f_0(\boldsymbol{x})$，$f_i(\boldsymbol{x})$，$h_i(\boldsymbol{x})$是定义在$\mathbb{R}^d$上的连续可微函数，一般的约束优化问题都可以表示为：

$$
\begin{aligned}
& \min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \\
\text{subject to} \quad & f_i(\boldsymbol{x}) \leq 0,\ i = 1,2\cdots k \\
& h_i(\boldsymbol{x}) = 0,\ i = 1,2\cdots l \qquad (1)
\end{aligned}
$$

称该问题为原始优化问题。

- (<u>广义拉格朗日函数</u>) 上述问题对应的拉格朗日函数为：

$$L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{k} \alpha_i f_i(\boldsymbol{x}) + \sum_{i=1}^{l} \beta_i h_i(\boldsymbol{x})$$

其中$\alpha_i \geq 0$是第i个不等式约束的拉格朗日乘子，β_i是等式约束的乘子，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots \alpha_k) \geq \mathbf{0}$，$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots \beta_l)$。

原始问题

- 考虑如下的函数：

$$\theta_p(\boldsymbol{x}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq\boldsymbol{0}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

- 思考$\theta_p(\boldsymbol{x})$的取值是什么？

  ① 当$\boldsymbol{x}$满足原始问题(1)的约束条件时，

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq\boldsymbol{0}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) = f_0(\boldsymbol{x})$$

  ② 当$\boldsymbol{x}$违反原始问题(1)中至少一个约束条件时，至少存在一个i使得$f_i(\boldsymbol{x}) > 0$，或者至少存在一个i使得$h_i(\boldsymbol{x}) \neq 0$，此时总是找到一个$\boldsymbol{\alpha} \geq \boldsymbol{0}$或$\boldsymbol{\beta}$，使得

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq\boldsymbol{0}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) =+\infty$$

- 综上所述：

$$\theta_p(\boldsymbol{x}) = \begin{cases} f_0(\boldsymbol{x}), & \text{如果}\boldsymbol{x}\text{满足原始问题(1)的约束条件} \\ +\infty, & \text{如果}\boldsymbol{x}\text{违反原始问题(1)的约束条件} \end{cases}$$

原始问题

- 因此，$\min_{\boldsymbol{x}} \theta_p(\boldsymbol{x})$ 这个优化问题**等价于原始优化问题(1)**，即两个优化问题有相同的解，又因为

$$\min_{\boldsymbol{x}} \theta_p(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq 0} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

- 观察到，原始优化问题(1)等价于拉格朗日函数的极小极大问题。
- 定义如下的函数：$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$
- 原始优化问题(1)的对偶优化问题定义为：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq 0} \theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\,\boldsymbol{\alpha}\geq 0} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

可以看到，原始优化问题(1)的**对偶问题**是拉格朗日函数的极大极小问题。

原始问题的对偶问题

- 换一种表述：

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})$$
$$\text{s.t.} \quad \alpha_i \geq 0,\ i = 1 \cdots k \qquad (2)$$

- 对任意的一个优化问题，无论原始问题是否是凸优化问题，其对偶问题一定是凸优化问题。

- 记原始优化问题(1)的最优值为p^*：

$$p^* = \min_{\boldsymbol{x}} \theta_p(\boldsymbol{x})$$

- 记对偶问题(2)的最优值为d^*：

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\ \boldsymbol{\alpha} \geq \boldsymbol{0}} \theta_D(\boldsymbol{\alpha},\boldsymbol{\beta})$$

原始问题和对偶问题的关系(弱对偶性)

- **定理1:** 对任意的一个优化问题，若原始问题和对偶问题都有最优值，则弱对偶性成立，即$d^* \leq p^*$。

证明：对于任意的$\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}$，有

$$\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}:\, \boldsymbol{\alpha} \geq \boldsymbol{0}} L(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \theta_p(\boldsymbol{x})$$

也即是：$\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \theta_p(\boldsymbol{x})$。

因此，$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}:\, \boldsymbol{\alpha} \geq \boldsymbol{0}} \theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leq \min_{\boldsymbol{x}} \theta_p(\boldsymbol{x})$，即$d^* \leq p^*$。

- 将$p^* - d^*$称为对偶差距(duality gap)。
- 如果$d^* = p^*$成立，我们可以说，强对偶性成立。在什么情况下，优化问题满足$d^* = p^*$?

Slater条件 (强对偶性成立的条件)

- **定义1：** 凸优化问题是形如：

$$\boxed{\begin{aligned} &\min_{\boldsymbol{x}} f_0(\boldsymbol{x}) \\ \text{s.t.}\quad & f_i(\boldsymbol{x}) \leq 0,\quad i = 1,2\cdots k \\ & \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} = b_i,\quad i = 1,2\cdots l \end{aligned}}$$

  的问题，其中，$f_0(\boldsymbol{x}), f_1(\boldsymbol{x}), \cdots f_k(\boldsymbol{x})$都是凸函数。

- **定理2:** 如果一个凸优化问题满足**Slater条件**：存在$\boldsymbol{x}$满足：$f_i(\boldsymbol{x}) < 0,\ i = 1,2\cdots k,\quad \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} = b_i,\ i = 1,2\cdots l$，
  那么，强对偶性成立，即$d^* = p^*$成立。

- 也就是说，如果强对偶性成立，则对原始问题的求解可以通过求解相应的对偶问题得到。

Slater条件 (强对偶性成立的条件)

- 定理2的语言描述:

对原始优化问题，如果其目标函数$f_0(\boldsymbol{x})$和所有的不等式约束函数$f_i(\boldsymbol{x})$是凸函数，所有的等式约束函数$h_i(\boldsymbol{x})$是仿射函数，并且所有的不等式约束是严格可行的，也即存在$\boldsymbol{x}$满足对$i = 1,2\cdots k$都有$f_i(\boldsymbol{x}) < 0$，对所有的$i = 1,2\cdots l$，有$h_i(\boldsymbol{x}) = 0$，则存在$\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$，使得$\boldsymbol{x}^*$是原始问题的最优解，$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$是对偶问题的最优解，并且

$$d^* = p^* = L(\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

改进的Slater条件

- 推论1: 如果一个优化问题具有如下形式:

$$\min_{\boldsymbol{x}} f_0(\boldsymbol{x})$$
$$\text{s.t.} \quad f_i(\boldsymbol{x}) \le 0, \quad i = 1,2\cdots k$$
$$\boldsymbol{a}_i^\intercal \boldsymbol{x} = b_i, \quad i = 1,2\cdots l$$

其中，$f_0(\boldsymbol{x})$和前r个不等式约束函数$f_1(\boldsymbol{x}),\cdots f_r(\boldsymbol{x})$是**凸函数**，$f_{r+1}(\boldsymbol{x})\cdots f_k(\boldsymbol{x})$是**仿射函数**，且存在$\boldsymbol{x}$满足:

$$f_i(\boldsymbol{x}) < 0, i = 1,2\cdots r,$$

$$f_i(\boldsymbol{x}) \le 0, i = r+1,\cdots k, \quad \boldsymbol{a}_i^\intercal \boldsymbol{x} = b_i, \ i = 1,2\cdots l$$

那么，强对偶性成立，即$d^* = p^*$成立。

改进的Slater条件

- **推论1的语言描述:** 对原始问题和对偶问题，如果目标函数$f_0(\boldsymbol{x})$是凸函数，不等式约束函数$f_1(\boldsymbol{x})\cdots f_r(\boldsymbol{x})$是凸函数，$f_{r+1}(\boldsymbol{x})\cdots f_k(\boldsymbol{x})$是仿射函数，所有的等式约束函数$h_i(\boldsymbol{x})$是仿射函数，并且所有的**凸不等式约束**是严格可行的，则存在$\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$，使得$\boldsymbol{x}^*$是原始问题的最优解，$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$是对偶问题的最优解，并且

$$d^* = p^* = L(\boldsymbol{x}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$$

- 划线的部分称为“改进的Slater条件”。
- 上述定理说明：如果原始问题的目标函数$f_0(\boldsymbol{x})$是凸函数，所有的等式和不等式约束都是仿射函数，只要该问题的可行域非空，则强对偶性成立。

KKT条件

- **定理3:** 如果一个**凸优化**问题具有**可微的**目标函数和约束函数，且满足Slater条件，则$\boldsymbol{x}^*$和$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$分别是原始问题和对偶问题的最优解的**充分必要条件**是$\boldsymbol{x}^*$和$\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*$满足Karush-Kuhn-Tucker (KKT)条件:

拉格朗日函数关于原问题的变量$\boldsymbol{x}$的梯度在最优解处为0

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*,\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = 0$$

$$\boxed{\alpha_i{}^* f_i(\boldsymbol{x}^*) = 0,\ i = 1,2\cdots k}$$

对偶互补条件

原始问题和对偶问题的约束条件

$$\begin{cases} f_i(\boldsymbol{x}^*) \le 0,\ i = 1,2\cdots k \\ h_i(\boldsymbol{x}^*) = 0,\ i = 1,2\cdots l \\ \alpha_i{}^* \ge 0,\ i = 1\cdots k \end{cases}$$

原始问题到对偶问题的转化

- 线性可分SVM的原始问题:

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

$$s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ \forall i = 1 \cdots N$$

- 原始问题满足slater条件吗?

① $\frac{1}{2}||\boldsymbol{w}||^2$是$\boldsymbol{w}$的凸函数, $1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)$是$\boldsymbol{w}$的仿射函数

② 原始问题的可行域非空。

综上, 原始问题满足改进的slater条件, 强对偶成立。

原问题是凸优化问题 + 满足slater条件

$\Longrightarrow$ KKT条件是最优性的充分必要条件。

线性可分SVM的对偶优化问题

- 线性可分SVM的原始优化问题：

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

$$s.t.\ \ y_i(\boldsymbol{w}^{\intercal}\boldsymbol{x}_i + b) \geq 1,\ \forall i = 1 \cdots N$$

- 推导出其对偶优化问题：

① 定义拉格朗日函数为：

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

② 对偶优化目标函数为：

$$\theta_D(\boldsymbol{\alpha}) = \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$$

如何求$\min\limits_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$？

令$L(\boldsymbol{w}, b, \boldsymbol{\alpha})$关于$\boldsymbol{w}$和$b$的偏导数分别为零：

SVM的对偶优化问题

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

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

- 带入拉格朗日函数得到：

$$\theta_D(\boldsymbol{\alpha}) = \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right)^{\top}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right) + \sum_{i=1}^{N} \alpha_i - \left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right)^{\top}\left(\sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i\right) - \sum_{i=1}^{N} \alpha_i y_i b$$

SVM的对偶优化问题

- 整理得到：

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

- 因为$\max\limits_{\boldsymbol{\alpha}} \theta_D(\boldsymbol{\alpha})$等价于$\min\limits_{\boldsymbol{\alpha}} -\theta_D(\boldsymbol{\alpha})$
- SVM的对偶优化问题为：

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\top}\boldsymbol{x}_j - \sum_{i=1}^{N}\alpha_i$$

$$\text{subject to} \quad \alpha_i \geq 0,\ i = 1\cdots N$$

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

由对偶解构造原始解

- 假设已经求出线性可分SVM的对偶问题的最优解为 $\boldsymbol{\alpha}^* = (\alpha_1{}^* \cdots \alpha_N{}^*)$，如何构造出原始问题的最优解$\boldsymbol{w}^*, b^*$？
- 思路：KKT条件
- 定理3告诉我们，满足KKT条件的$\boldsymbol{w}^*, b^*$和$\boldsymbol{\alpha}^*$一定分别是原始问题和对偶问题的最优解。
- 由KKT条件可得到：

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}^*) = 0 \Rightarrow \boxed{\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i{}^* y_i \boldsymbol{x}_i}$$

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

对偶互补条件： $\alpha_i{}^*(1 - y_i(\boldsymbol{w}^{*\top} \boldsymbol{x}_i + b^*)) = 0, i = 1 \cdots N$

由对偶解构造原始解

原问题约束： $y_i({\boldsymbol{w}^*}^\top \boldsymbol{x}_i + b^*) \geq 1, i = 1 \cdots N$

对偶约束： ${\alpha_i}^* \geq 0,\ i = 1 \cdots N$

- 根据KKT条件可得：至少有一个拉格朗日乘子${\alpha_j}^* > 0$。因为如果$\forall i = 1 \cdots N$，${\alpha_i}^* = 0$，则$\boldsymbol{w}^* = \boldsymbol{0}$，而$(\boldsymbol{0}, b)$不是问题的可行解，产生矛盾。

由${\alpha_j}^* \neq 0$可得 $1 - y_j({\boldsymbol{w}^*}^\top \boldsymbol{x}_j + b^*) = 0$，结合${y_j}^2 = 1$得：

$$b^* = y_j - {\boldsymbol{w}^*}^\top \boldsymbol{x}_j = y_j - \sum_{i=1}^N {\alpha_i}^* y_i\, {\boldsymbol{x}_i}^\top \boldsymbol{x}_j$$

- ✓ 定理：若SVM对偶问题的最优解为$\boldsymbol{\alpha}^* = ({\alpha_1}^* \cdots {\alpha_N}^*)$，则存在下标$j$，使得${\alpha_j}^* > 0$，可按如下公式得到原问题的最优解$(\boldsymbol{w}^*, b^*)$：

$$\boldsymbol{w}^* = \sum_{i=1}^N {\alpha_i}^* y_i \boldsymbol{x}_i,\quad b^* = y_j - \sum_{i=1}^N {\alpha_i}^* y_i\, {\boldsymbol{x}_i}^\top \boldsymbol{x}_j$$

线性可分SVM的对偶学习算法

- 输入：线性可分的训练数据集$\{(\boldsymbol{x}_1, y_1)\cdots(\boldsymbol{x}_N, y_N)\}$，其中$\boldsymbol{x}_i \in \mathbb{R}^d$，$y_i \in \{+1, -1\}$。
- 输出：最大间隔的分离超平面和分类决策函数。

1. 求解如下优化问题：

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i{}^{\top}\boldsymbol{x}_j - \sum_{i=1}^{N}\alpha_i$$

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

得到最优解$\boldsymbol{\alpha}^* = (\alpha_1{}^*\cdots\alpha_N{}^*)$;

2. 计算原问题最优解：$\boldsymbol{w}^* = \sum_{i=1}^{N}\alpha_i{}^* y_i \boldsymbol{x}_i$，选择$\boldsymbol{\alpha}^*$的一个正分量$\alpha_j{}^* > 0$，计算$b^* = y_j - \sum_{i=1}^{N}\alpha_i{}^* y_i\, \boldsymbol{x}_i{}^{\top}\boldsymbol{x}_j$

3. 得到分割超平面：$\boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 0$

   分类决策函数$f(\boldsymbol{x}) = sign(\boldsymbol{w}^{*\top}\boldsymbol{x} + b^*)$

支持向量

- 由对偶解和原始解之间的关系可得到：$\boldsymbol{w}^*$和b^*仅依赖于${\alpha_i}^* > 0$的样本$(\boldsymbol{x}_i, y_i)$，其他样本对它们没有影响。
- 事实上，训练数据集中对应于${\alpha_i}^* > 0$的样本$(\boldsymbol{x}_i, y_i)$称为支持向量。
- 根据KKT条件中的对偶互补条件：

$$ {\alpha_i}^*(1 - y_i({\boldsymbol{w}^*}^\top \boldsymbol{x}_i + b^*)) = 0 $$

若${\alpha_i}^* > 0$，则一定有$y_i({\boldsymbol{w}^*}^\top \boldsymbol{x}_i + b^*) = 1$。

所以支持向量在如下超平面上：

$$ h_1:\ {\boldsymbol{w}^*}^\top \boldsymbol{x} + b^* = 1 $$

$$ h_2:\ {\boldsymbol{w}^*}^\top \boldsymbol{x} + b^* = -1 $$

- 这与前面得到的结论是一致的。

小结

- 最简单的支持向量机是线性可分的支持向量机，也称为硬间隔的支持向量机，构建它的条件是训练数据集线性可分，其学习的策略是最大化间隔法，可以形式化为求解如下的凸二次规划问题：

重点

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

$$s.t.\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ \forall i = 1 \cdots N$$

求得最优解记为$\boldsymbol{w}^*, b^*$，得到的分类决策边界为 $\boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 0$，分类决策函数为$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^{*\top}\boldsymbol{x} + b^*)$。

- 线性可分的支持向量机的最优解存在且唯一。
- 距离分类决策超平面最近的样本点称为支持向量。

小结

- 线性可分支持向量机的对偶优化问题为：

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\top}\boldsymbol{x}_j - \sum_{i=1}^{N}\alpha_i$$

$$\text{subject to} \quad \alpha_i \geq 0,\ i = 1\cdots N$$

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

重难点

通过求解对偶问题的最优解$\boldsymbol{\alpha}^*$，可以求出原始问题的最优解$\boldsymbol{w}^*, b^*$。

- 支持向量对应的$\alpha_i^* > 0$，且支持向量位于超平面

$$h_1:\ \boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = 1$$

$$h_2:\ \boldsymbol{w}^{*\top}\boldsymbol{x} + b^* = -1$$