本文介绍传统机器学习算法中的“支撑向量机（SVM）”，在深度学习变得流行之前，SVM是许多论文中经常会用到的算法。它有非常完备的数学推导，我在刚接触它的时候也被搞得云里雾里。现在打算系统的介绍一下它。

本文一共分成以下几个部分

1. 线性可分支撑向量机及其对偶问题
2. 线性不可分支撑向量机及其对偶问题
3. 非线性支撑向量机（核技巧）
4. SMO算法

文章重点参考了《统计学习方法》的支撑向量机一章。

声明：文章中用到的上下标，上标的标示是括号加数字，例如$\mathbf{x}^{(i)}$表示样本集中第i个样本点，加粗的$\mathbf{x}$表示这是一个向量。下标x_i表示向量$\mathbf{x}$的第i个维度。这种标记方式是follow ng的课程的。

2 线性不可分支撑向量机及其对偶问题

2.1 问题描述

有关数据集T以及$\mathbf{x}$和y的定义和1.1节一致。区别在于，此时的数据集T无法做到线性可分。此时数据集T中存在样本点无法满足$y^{(i)}\left(\mathbf{w}\cdot\mathbf{x}^{(i)}+b\right)-1\geq 0$。此时为每个样本点引入松弛变量$\xi^{(i)}$。将约束条件改写成

$$y^{(i)}\left(\mathbf{w}\cdot\mathbf{x}^{(i)}+b\right)\geq 1-\xi^{(i)}$$

同时修改目标函数，使得目标函数变为

$$\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}\xi^{(i)}$$

其中$C>0$作为超参数(hyperparameter)输入，后面一项为惩罚项。C越大表明对惩罚越严重，反之对惩罚越小。

此时的凸二次规划问题为

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

这是一个目标函数为二次函数，约束条件为一次的凸二次规划问题。通过求解该问题得到$(\mathbf{w}^\star,b^\star)$，则最优超平面为$\mathbf{w}^\star\cdot\mathbf{x}+b^\star=0$

2.2 凸二次问题求解

在2.1节结尾的最小化问题是原始问题，我们依旧可以用在第1.3节中介绍的拉格朗日对偶性质将其转换成对偶问题来求解。定义拉格朗日函数

$$L(\mathbf{w},b,\xi,\alpha,\mu)=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{N}\xi^{(i)}-\sum_{i=1}^{N}\alpha^{(i)}(y^{(i)}(\mathbf{w}\cdot\mathbf{x}^{(i)}+b)-1+\xi^{(i)})-\sum_{i=1}^{N}\mu^{(i)}\xi^{(i)}$$

其中拉格朗日乘子$\alpha^{(i)} \geq 0, \xi^{(i)} \geq 0, i=1,2,\cdots,N$ 根据1.4节，分两步走求解该原始问题的对偶问题，即极大极小问题。

首先求$L(\mathbf{w},b,\xi,\alpha,\mu)$对$\mathbf{w},b,\xi$求极小，它们的偏导数分别是

$$\begin{aligned}
&\nabla_{\mathbf{w}} L(\mathbf{w},b,\xi,\alpha,\mu) = \mathbf{w} - \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} = 0 \\
&\nabla_{b} L(\mathbf{w},b,\xi,\alpha,\mu) = -\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0 \\
&\nabla_{\xi^{(i)}} L(\mathbf{w},b,\xi,\alpha,\mu) = C - \alpha^{(i)} - \mu^{(i)} = 0, \quad i=1,2,\cdots,N
\end{aligned}$$

整理得

$$\begin{aligned}
&\mathbf{w} = \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} \\
&\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0 \\
&C - \alpha^{(i)} - \mu^{(i)} = 0, \quad i=1,2,\cdots,N
\end{aligned}$$

代回拉格朗日函数，整理得

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

整理的过程和1.4节中的一致。 将$\min_{\mathbf{w},b,\xi} L(\mathbf{w},b,\xi,\alpha,\mu)$定义成$\theta_D(\alpha,\mu)$，求$\theta_D(\alpha,\mu)$对$\alpha$的极大

$$\begin{aligned}
\max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}) + \sum_{i=1}^{N} \alpha^{(i)} \\
\text{s. t.} \quad & \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0 \\
& \alpha^{(i)} \geq 0, \quad i=1,2,\cdots,N \\
& \mu^{(i)} \geq 0, \quad i=1,2,\cdots,N \\
& C - \alpha^{(i)} - \mu^{(i)} = 0, \quad i=1,2,\cdots,N
\end{aligned}$$

可以把$\alpha^{(i)}, \mu^{(i)}, C$的约束条件改写成$0 \leq \alpha^{(i)} \leq C$。目标函数中添加一个负号，最大化问题改写成最小化问题，于是变成了如下的对偶问题的形式

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}) - \sum_{i=1}^{N} \alpha^{(i)} \\
\text{s. t.} \quad & \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0 \\
& C \geq \alpha^{(i)} \geq 0, i=1,2,\cdots,N
\end{aligned}$$

我们通过求解对偶问题得到了最优解$\alpha^{\star} = \{(\alpha^{(1)})^{\star}, (\alpha^{(2)})^{\star}, \cdots, (\alpha^{(N)})^{\star}\}^T \in \mathbb{R}^N$。根据KKT条件可以推导出$\alpha^{\star}$与$(\mathbf{w}^{\star}, b^{\star})$的关系。

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

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

$$\nabla_\xi L(\mathbf{w}, b, \xi, \alpha, \mu) = C - \alpha^\star - \mu^\star = 0$$

$$(\alpha^{(i)})^\star (y^{(i)} (\mathbf{w}^\star \cdot \mathbf{x}^{(i)} + b^\star) - 1 + (\xi^{(i)})^\star) = 0, \quad i = 1, 2, \cdots, N$$

$$(\mu^{(i)})^\star (\xi^{(i)})^\star = 0$$

$$y^{(i)} (\mathbf{w}^\star \cdot \mathbf{x}^{(i)} + b^\star) - 1 + (\xi^{(i)})^\star \geq 0, \quad i = 1, 2, \cdots, N$$

$$(\xi^{(i)})^\star \geq 0, \quad i = 1, 2, \cdots, N$$

$$(\alpha^{(i)})^\star \geq 0, \quad i = 1, 2, \cdots, N$$

$$(\mu^{(i)})^\star \geq 0, \quad i = 1, 2, \cdots, N$$

其中前三个条件为求拉格朗日函数对$\mathbf{w}, b, \xi$的偏导数等于0得到的。第四和第五个函数分别来自原始问题中的两个约束条件，通过引入拉格朗日乘子α和μ引入到目标函数时，要满足的KKT对偶互补条件。第六和第七个条件是原始问题的约束条件。最后两个条件为引入的拉格朗日乘子需要满足的条件。

通过整理这些KKT条件可得

$$\mathbf{w}^\star = \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} \mathbf{x}^{(i)}$$

$$b^\star = y^{(j)} - \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})$$

可以看到原始问题的最优解$(\mathbf{w}^\star, b^\star)$通过拉格朗日乘子$\alpha^\star$表示出来了。

根据$\mathbf{w}^\star \cdot \mathbf{x} + b^\star = 0$可以把分离超平面表示成

$$\sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} (\mathbf{x}^{(i)} \cdot \mathbf{x}) + b^\star = 0$$

这是线性可分的SVM的对偶形式。可以看到分离超平面依赖于训练样本$\mathbf{x}^{(j)}$和输入的$\mathbf{x}$的内积。 类别决策函数则为将样本点带入分类超平面的方程式后得到的结果的符号，写成

$$f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} (\mathbf{x}^{(i)} \cdot \mathbf{x}) + b^\star)$$

注意：在利用$b^\star = y^{(j)} - \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})$求解$b^\star$时，要在$\alpha^\star = \{(\alpha^{(1)})^\star, (\alpha^{(N)})^\star, \cdots, (\alpha^{(N)})^\star\}$中选择$0 < (\alpha^{(j)})^\star < C$，则可以求出多个$b^\star$。在实际情况中，$b^\star$可以取平均值。

2.3 支撑向量

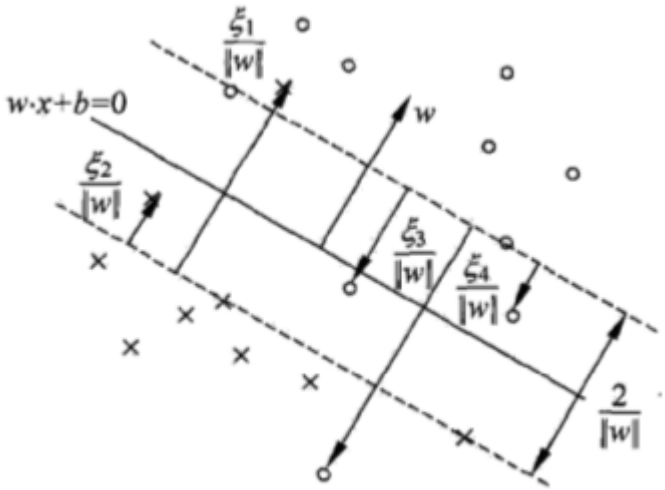

图 7.5　软间隔的支持向量

图中画o的点，有在间隔边界上，有在分离超平面与间隔之间，有在分离超平面上，也有误分在分离超平面另外一侧。这些点都会影响分离超平面的选择，按照定义，它们被称为支撑向量。支撑向量的特点是，对应的拉格朗日乘子$\alpha^{(i)} > 0$。

若$\alpha^{(i)} < C$，则根据KKT的第三和第五个条件可知$\xi^{(i)} = 0$。说明这个样本点不需要做间隔的松弛，即它在分类超平面的间隔上。 若$\alpha^{(i)} = C$，则根据KKT的第三个条件可知$\mu^{(i)} = 0$，则第五个条件告知$\xi^{(i)}$要分情况讨论。 $\xi^{(i)}$在引入的时候是以满足$\xi^{(i)} \geq 0$为约束的。

当$\xi^{(i)} = 0$表示样本点不需要做间隔松弛，即它在分离超平面的间隔上。

当$1 > \xi^{(i)} > 0$表示样本点在分离超平面和间隔之间。

当$\xi^{(i)} = 1$表示样本点在分离超平面上。

当$\xi^{(i)} > 1$表示样本点在跨过了分离超平面，跑到了错误的一侧。