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。

1. 线性可分支撑向量机及其对偶问题

这是一个最简单的SVM算法，通过这个例子，将会认识到利用SVM算法求解一个简单的二分类问题的整个流程。后面的线性不可分问题以及核技巧，都是在基于线性可分的基础上进行的拓展。所以这个环节会介绍的比较详细，我会尽量讲清楚公式推导的部分，不要畏惧公式的繁杂，慢慢看是可以看明白的。

1.1 问题描述

给定一组训练数据集

$$T = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

该数据集包含N个样本点。每个样本点的$\mathbf{x}$加粗显示，表示这是一个向量，$\mathbf{x} \in \mathbb{R}^n$，当然如果n=1，则$\mathbf{x}$是一个标量。

在$\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)})$中的每一个维度，表示该样本点的一个特征，样本集中的每个样本点有n个维度或特征。

$y^{(i)}$表示第i个样本点的类别，$y \in \{+1, -1\}$，当$y^{(i)} = 1$，则表示$\mathbf{x}^{(i)}$是正例。当$y^{(i)} = -1$，则表示$\mathbf{x}^{(i)}$是负例。学习的目标就是要找到一个超平面，这个超平面把空间分成两个部分，使得样本集中的正负样本点分别位于各自的部分。这个超平面用方程表示为$\mathbf{w} \cdot \mathbf{x} + b = 0$，它由法相量$\mathbf{w}$和截距项$b$所决定。

需要注意的是，对于一组线性可分的数据集，存在着无穷多个超平面可以把这组数据集的正负例完全分开。我们要找的是一个最优的超平面，以下将解释什么样的超平面算最优。

1.2 最优超平面

上面介绍到，超平面由法相量$\mathbf{w}$和截距项b所决定，那么这里定义最优超平面对应的法相量和截距项分别为$\mathbf{w}^\star$和$b^\star$。则最优超平面的方程为

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

在上面的图7.1中，样本点的特征是二维的，所以利用一条直线可以将样本点分割开。用o表示正例，x表示负例。在正例中有三个点A、B、C，可以看到这三个点都被直线正确的分类了。但是仔细看这三个点到直线的距离则有所不同，C离直线的距离最近，A离得最远，B居中，这里说的距离可以理解为垂直于直线的距离。当距离越大时，表示分类的结果比较可信，当距离越小时，表示分类的结果比较不可信。为什么要说可信不可信，是因为我们获得的样本集T只是从众多样本点中随机抽样得到的，假如在运气不好的情况下，获得的样本集比较诡异，那么虽然分类直线在这个样本集上做到了完全分类正确，但是这条分类直线在其他不在样本集上的样本点，可能表现就会不佳。因为我们在抽样的时候没有做好，使得样本无法代表总体。而基于这样的样本集得到的分类直线，则不是一条好的分类直线。

在这个例子中，C点距离直线最近，表示虽然直线把C点的类别分对了，但是差一点就分错了。因为有上述抽样问题存在，我们不敢把C点刚刚好分对，而要尽可能的让它远离分类直线。通过使得这个距离最大化，得到最优的分类超平面。

为了解释最优超平面，首先要引入函数间隔和几何间隔的概念。

函数间隔

超平面的方程式是$\mathbf{w} \cdot \mathbf{x} + b = 0$，则样本点$(\mathbf{x}^{(i)}, y^{(i)})$到超平面$(\mathbf{w}, b)$的函数间隔为

$$\hat{\gamma}^{(i)} = y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$$

为什么可以这么定义，原因是对于超平面$(\mathbf{w}, b)$而言，样本点$(\mathbf{x}^{(i)}, y^{(i)})$到超平面的距离用$\mathbf{w} \cdot \mathbf{x} + b$表示，这是一个描述距离的值，它的绝对值表示距离平面的远近。这个值可正可负。如果样本点在平面的上方，即平面把它判成正例，则该值是正的，如果样本点在平面的下方，即平面把它判成负例，则该值是负的。

通过比较$\mathbf{w} \cdot \mathbf{x} + b$的符号与$y^{(i)}$的符号是否相同，可以判断平面是否正确分类。所以我们用$y^{(i)}(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)$来表示分类正确和距离平面远近，也就是上文函数间隔的定义。

在样本集T中的所有样本点中找到$\hat{\gamma}^{(i)}$最小的一个，即为超平面$(\mathbf{w}, b)$关于样本集T的函数间隔

$$\hat{\gamma} = \min_{i=1 \cdots N} \hat{\gamma}^{(i)}$$

几何间隔

上文中的分离超平面方程式$(\mathbf{w}, b) : \mathbf{w} \cdot \mathbf{x} + b = 0$，如果在方程两边同时乘以一个系数c，则不改变分离超平面。但是函数间隔却会变成原来的c倍。为了保证样本点到超平面的距离不受到系数的影响，引入几何间隔的概念，即在计算距离的时候对超平面方程式中的参数做归一化。

$$\begin{aligned} \gamma^{(i)} &= y^{(i)} \frac{(\mathbf{w} \cdot \mathbf{x}^{(i)} + b)}{\|\mathbf{w}\|} \\ &= y^{(i)} \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}^{(i)} + \frac{b}{\|\mathbf{w}\|} \right) \end{aligned}$$

则分离超平面关于样本集T的几何间隔为

$$\gamma = \min_{i=1\cdots N} \gamma^{(i)}$$

所以函数间隔和集合间隔的关系是

$$\gamma = \frac{\hat{\gamma}}{\|\mathbf{w}\|}$$
$$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|\mathbf{w}\|}$$

间隔最大化

这里的间隔指几何间隔最大化。对于一个线性可分的数据集T，存在无穷多个完全分类正确的超平面，通过使得几何间隔γ最大化，找到最优的分离超平面$(\mathbf{w}^\star, b^\star)$

我们可以把上述过程用数学方式表达出来

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

根据函数间隔与几何间隔之间的关系，我们也可以把上式写成

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

上文提到如果用到函数间隔，要注意超平面的系数$\mathbf{w}$和b按照系数c等比例增大或缩小的问题。假设将$\mathbf{w}$和b等比例的变成$c\mathbf{w}$和cb，则函数间隔变成了$c\hat{\gamma}$。带入用函数间隔表示的目标问题和约束条件中后发现没有产生任何影响。所以不失一般性，这里取函数间隔$\hat{\gamma} = 1$

所以问题变成

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

为了后续推导的方便，利用最小化$\frac{1}{\|\mathbf{w}\|}$和最大化$\frac{1}{2}\|\mathbf{w}\|^2$是等价的，于是把问题修改成

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

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

1.3 拉格朗日对偶性质

在正式求解上述凸二次规划问题之前，要先补充一下有关怎么利用对偶问题来求解原始问题的知识。

以下内容都以输入一个样本点$\mathbf{x}$为例子。

拉格朗日极小极大问题

考虑以下优化问题

$$\begin{aligned}\max_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t} \quad & c^{(i)}(\mathbf{x}) \le 0 \quad i=1,2,\cdots,k \\ & h^{(j)}(\mathbf{x}) = 0 \quad j=1,2,\cdots,l\end{aligned}$$

这是一个标准的带约束的优化问题，称为原始问题(primal problem)。下面我们要利用拉格朗日对偶性质求解该问题。

假设$f(\mathbf{x})$，$c(\mathbf{x})$和$h(\mathbf{x})$在$\mathbb{R}^n$上是连续可微的（函数存在导函数，且导函数是连续的）。引入拉格朗日函数

$$L(\mathbf{x},\alpha,\beta) = f(\mathbf{x}) + \sum_{i=1}^{k}\alpha^{(i)}c^{(i)}(\mathbf{x}) + \sum_{j=1}^{l}\beta^{(j)}h^{(j)}(\mathbf{x})$$

其中$\alpha^{(i)}$和$\beta^{(j)}$是拉格朗日乘子，且$\alpha^{(i)} \ge 0$。 拉格朗日函数$L(\mathbf{x},\alpha,\beta)$有三个变量，分别是$\mathbf{x}$，$\alpha$和$\beta$

定义关于$\mathbf{x}$的函数

$$\theta_P(\mathbf{x}) = \max_{\alpha,\beta;\alpha^{(i)}\ge 0} L(\mathbf{x},\alpha,\beta)$$

下表P表示primal原始问题的意思。可以看到，这个$\theta_P(\mathbf{x})$函数从拉格朗日函数演变而来。它只有一个变量$\mathbf{x}$，因为剩余的两个变量α和β在求$\max_{\alpha,\beta;\alpha^{(i)}\ge 0} L(\mathbf{x},\alpha,\beta)$过程中找到了能够使拉格朗日函数最大的$\alpha$和$\beta$。这里用$\alpha^\star$和$\beta^\star$表示。所以$\alpha$和$\beta$变量已经固定，即为常数。

分析$\theta_P(\mathbf{x})$的取值范围。

如果存在$\mathbf{x}$使得在原始问题中的约束条件不满足，即存在$c^{(i)}(\mathbf{x}) > 0$或$h^{(i)}(\mathbf{x}) \ne 0$，则可以通过调节$\alpha$和$\beta$使得$\theta_P(\mathbf{x})$函数达到无穷大。例如$c^{(i)}(\mathbf{x}) > 0$，则令对应的$\alpha^{(i)} \to +\infty$。如果$h^{(j)}(\mathbf{x}) \ne 0$，则根据符号的方向决定$\beta^{(j)}$的方向，使得$\beta^{(j)}h^{(j)}(\mathbf{x}) \to +\infty$，将其他拉格朗日乘子取0即可。

如果所有的$\mathbf{x}$都满足在原始问题中的约束条件，则将所有的$\alpha^{(i)}$取0，可以使得$\theta_P(\mathbf{x})$取到最大值，即$\theta_P(\mathbf{x}) = f(\mathbf{x})$。

综合考虑两种情况得到$\theta_P\mathbf{x}$的取值范围

$$\theta_P(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \mathbf{x}\ \text{satisfy the constraints} \\ +\infty, & \text{otherwise}\end{cases}$$

可以看到如果$\mathbf{x}$满足原始问题的约束条件，则$\theta_P(\mathbf{x})$可以取到$f(\mathbf{x})$。原始问题是通过调节$\mathbf{x}$对$f(\mathbf{x})$进行最小化，则通过调节$\mathbf{x}$来最小化$\theta_P(\mathbf{x})$函数，这与最小化原始问题是等价的，即

$$\min_{\mathbf{x}}\theta_P(\mathbf{x}) = \min_{\mathbf{x}}\max_{\alpha,\beta;\alpha^{(i)}\ge 0} L(\mathbf{x},\alpha,\beta)$$

则原始问题转变成了拉格朗日函数的极小极大问题，定义$p^\star = \min_{\mathbf{x}} \theta_P(\mathbf{x})$为原始问题的最优值。

拉格朗日对偶问题

在上文中提到的拉格朗日函数的极小极大问题，求解的步骤是

1. 通过调节拉格朗日乘子$\alpha^{(i)}$和$\beta^{(j)}$使得$L(\mathbf{x}, \alpha, \beta)$最大化。
2. 通过在样本集中选择$\mathbf{x}$使得$\theta_P(\mathbf{x})$函数最小。

现在我们将两个步骤的顺序交换位置

1. 先对$L(\mathbf{x}, \alpha, \beta)$，通过调节$\mathbf{x}$实现最小化，定义

$$\theta_D(\alpha, \beta) = \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta)$$

2. 再考虑调节α和β去最大化$\theta_D(\alpha, \beta)$函数，即

$$\max_{\alpha, \beta; \alpha^{(i)} \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha, \beta; \alpha^{(i)} \geq 0} \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta)$$

这是拉格朗日的极大极小问题。如果把第2步中要求$\alpha^{(i)} \geq 0$表示成约束条件，则极大极小问题可以写成约束最优化问题，即

$$\begin{aligned} \max_{\alpha, \beta} \theta_D(\alpha, \beta) &= \max_{\alpha, \beta} \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta) \\ \text{s.t} \quad \alpha^{(i)} &\geq 0 \quad i = 1, 2, \cdots, k \end{aligned}$$

这种写法称为原始问题的对偶问题。定义$d^\star = \max_{\alpha, \beta; \alpha^{(i)} \geq 0} \theta_D(\alpha, \beta)$为对偶问题的最优值。

原始问题和对偶问题的关系

因为$\theta_D(\alpha, \beta) = \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta)$，通过调节$\mathbf{x}$使得$L(\mathbf{x}, \alpha, \beta)$最小化。那么
$\theta_D(\alpha, \beta) = \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta) \leq L(\mathbf{x}, \alpha, \beta)$

同时，$\theta_P(\mathbf{x}) = \max_{\alpha, \beta; \alpha^{(i)} \geq 0} L(\mathbf{x}, \alpha, \beta)$，通过调节$\alpha$和$\beta$使得$L(\mathbf{x}, \alpha, \beta)$最大化。那么
$\theta_P(\mathbf{x}) = \max_{\alpha, \beta; \alpha^{(i)} \geq 0} L(\mathbf{x}, \alpha, \beta) \geq L(\mathbf{x}, \alpha, \beta)$

联立两式可得$\theta_D(\alpha, \beta) \leq \theta_P(\mathbf{x})$。即$\theta_D$关于$\alpha$和$\beta$的函数不超过$\theta_P$关于$\mathbf{x}$的函数。因此$\theta_D(\alpha, \beta)$的最大值也不会大于$\theta_P(\mathbf{x})$的最小值。

所以$\max_{\alpha, \beta; \alpha^{(i)} \geq 0} \theta_D(\alpha, \beta) \leq \min_{\mathbf{x}} \theta_P(\mathbf{x})$，即有以下关系。

$$d^\star = \max_{\alpha, \beta; \alpha^{(i)} \geq 0} \min_{\mathbf{x}} L(\mathbf{x}, \alpha, \beta) \leq \min_{\mathbf{x}} \max_{\alpha, \beta; \alpha^{(i)} \geq 0} L(\mathbf{x}, \alpha, \beta) = p^\star$$

说明了原始问题的最优值不小于对偶问题的最优值。由于我们要用求解对偶问题的最优值来求解原始问题的最优值，所以要保证$d^\star = p^\star$。接下来要介绍满足原始问题最优解等于对偶问题最优解的情况。

KKT条件

在原始问题和对偶问题中，$f(\mathbf{x})$和$c^{(i)}(\mathbf{x})$是凸函数，$h^{(j)}(\mathbf{x})$是仿射函数。仿射函数的定义在这个链接可以找到。https://baike.baidu.com/item/%E4%BB%BF%E5%B0%84%E5%87%BD%E6%95%B0/9276178?fr=aladdin

不等式约束$c^{(i)}(\mathbf{x})$是严格可行的，即存在$\mathbf{x}$使得所有的$c^{(i)}(\mathbf{x}) < 0 \quad i = 1, 2, \cdots, k$。

则存在$x^\star$是原始问题的最优解，$\alpha^\star$和$\beta^\star$是对偶问题的最优解的充分必要条件是$x^\star$、$\alpha^\star$和$\beta^\star$要满足以下KKT条件

$$
\begin{aligned}
&\nabla_{\mathbf{x}} L(\mathbf{x}^\star, \alpha^\star, \beta^\star) = 0 \\
&\nabla_{\alpha} L(\mathbf{x}^\star, \alpha^\star, \beta^\star) = 0 \\
&\nabla_{\beta} L(\mathbf{x}^\star, \alpha^\star, \beta^\star) = 0 \\
&(\alpha^{(i)})^\star c^{(i)}(\mathbf{x}^\star) = 0 \quad i = 1, 2, \cdots, k \\
&c^{(i)}(\mathbf{x}^\star) \le 0 \quad i = 1, 2, \cdots, k \\
&(\alpha^{(i)})^\star \ge 0 \quad i = 1, 2, \cdots, k \\
&h^{(j)}(\mathbf{x}^\star) = 0 \quad j = 1, 2, \cdots, l
\end{aligned}
$$

其中，前三个条件是要求$f(\mathbf{x})$，$c(\mathbf{x})$和$h(\mathbf{x})$在$\mathbb{R}^n$上是连续可微的，因此存在对$x^\star$、$\alpha^\star$和$\beta^\star$的偏导数，且偏导数为0。

第五和第七个条件是原始问题中的约束条件要求满足的。第六个条件是在引入拉格朗日时的拉格朗日乘子$\alpha^{(i)}, i = 1, 2, \cdots, k$要满足大于等于0的条件。第四个条件称为KKT对偶互补条件，可以看到$(\alpha^{(i)})^\star$和$c^{(i)}(\mathbf{x}^\star)$两项中至少要有一项等于0，在之后的推导中会利用到这个性质。

总结：在求解带约束的原始问题时，可以通过拉格朗日对偶性质，将其转换成对偶问题。通过求解对偶问题的最优解$\alpha^\star$和$\beta^\star$，反推原始问题的最优解$x^\star$，检查原始问题和对偶问题的最优解是否满足KKT条件。如果是，则$p^\star = d^\star = L(\mathbf{x}^\star, \alpha^\star, \beta^\star)$

利用对偶问题来求解原始问题的知识介绍到这，接下来介绍怎样在最大化间隔问题上利用这个性质。

1.4 凸二次问题求解

拉格朗日乘子法

回顾在1.2节的截尾提到的凸二次规划问题，我们称之为原始问题。利用1.3节介绍的拉格朗日对偶性质，可以把问题转变成对偶问题。 定义拉格朗日函数

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

其中N个$\alpha^{(i)}, i = 1 \cdots N$组成拉格朗日乘子向量$\alpha = (\alpha^{(1)}, \alpha^{(2)}, \cdots, \alpha^{(N)})$，即样本点的个数与拉格朗日乘子个数相同。

通过拉格朗日乘子，把N个不等式约束条件，转变到目标函数中的部分。得到了拉格朗日函数L。

利用拉格朗日对偶性质，原始问题的对偶问题是极大极小问题，即$\max_{\alpha;\alpha^{(i)} \ge 0} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$，接下来分两步求解这个极大极小问题。

1. 求内层的$\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$

将拉格朗日函数$L(\mathbf{w}, b, \alpha)$分别求对$\mathbf{w}$和b的偏导数，令它们等于0，即

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

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

第一个式子是对$\mathbf{w}$向量求偏导数，等式右边是0向量，第二个式子是对标量b求偏导数，等式右边是数字0。整理上面两个式子可得

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

$$\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} = 0$$

将其代回拉格朗日函数。 向量$\mathbf{w} \in \mathbb{R}^n$。将$\mathbf{w}$展开可以得到

$$\begin{aligned}
\mathbf{w} &= \{w_1, w_2, \cdots, w_n\} \\
&= \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} \mathbf{x}^{(i)} \\
&= \{\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_1^{(i)}, \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_2^{(i)}, \cdots, \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_n^{(i)}\}
\end{aligned}$$

所以拉格朗日函数第一项中

$$\begin{aligned}
\|\mathbf{w}\|^2 &= w_1^2 + w_2^2 + \cdots + w_n^2 \\
&= (\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_1^{(i)})^2 + (\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_2^{(i)})^2 + \cdots + (\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} x_n^{(i)})^2 \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} x_1^{(i)} x_1^{(j)} + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} x_2^{(i)} x_2^{(j)} + \cdots + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} x_n^{(i)} x_n^{(j)} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} (x_1^{(i)} x_1^{(j)} + x_2^{(i)} x_2^{(j)} + \cdots + x_n^{(i)} x_n^{(j)}) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})
\end{aligned}$$

拉格朗日函数的第二项

$$\begin{aligned}
\sum_{i=1}^{N} \alpha^{(i)} y^{(i)} (\mathbf{w} \cdot \mathbf{x}^{(i)} + b) &= \sum_{i=1}^{N} \alpha^{(i)} y^{(i)} ((\sum_{j=1}^{N} \alpha^{(j)} y^{(j)} \mathbf{x}^{(j)}) \cdot \mathbf{x}^{(i)} + b) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha^{(i)} \alpha^{(j)} y^{(i)} y^{(j)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})
\end{aligned}$$

所以拉格朗日函数整理后可写成

$$
\begin{aligned}
L(\mathbf{w}, b, \alpha) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}(\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}(\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}) + \sum_{i=1}^{N}\alpha^{(i)} \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}(\mathbf{x}^{(i)}\cdot\mathbf{x}^{(j)}) + \sum_{i=1}^{N}\alpha^{(i)}
\end{aligned}
$$

此时得到的拉格朗日函数为最小值，即

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

在这个表达式中，$\mathbf{w}$和b消失了，因为它们被α、y和$\mathbf{x}$表示了。

2. 求$\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$关于$\alpha$的极大值。

   回顾1.3节，定义$\theta_D(\alpha) = \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$，对$\theta_D(\alpha)$求关于$\alpha$的极大值

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

   如果把负号拿掉，则目标函数变成

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

   在原始问题中，目标函数$\frac{1}{2}\|\mathbf{w}\|^2$和约束条件$y^{(i)}\left(\mathbf{w}\cdot\mathbf{x}+b\right)-1$是凸函数，且$y^{(i)}\left(\mathbf{w}\cdot\mathbf{x}+b\right)-1$严格可行，则存在$\mathbf{w}^\star$和$b^\star$是原始问题的解，$\alpha^\star$是对偶问题的解。下面介绍$(\mathbf{w}^\star, b^\star)$和$\alpha^\star$的关系。

原始问题与对偶问题的解的关系

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。

$$
\begin{aligned}
&\nabla_{\mathbf{w}} L(\mathbf{w}^\star, b^\star, \alpha^\star) = \mathbf{w}^\star - \sum_{i=1}^{N}(\alpha^{(i)})^\star y^{(i)}\mathbf{x}^{(i)} = 0 \\
&\nabla_{b} L(\mathbf{w}^\star, b^\star, \alpha^\star) = -\sum_{i=1}^{N}(\alpha^{(i)})^\star y^{(i)} = 0 \\
&(\alpha^{(i)})^\star(y^{(i)}(\mathbf{w}^\star\cdot\mathbf{x}^{(i)} + b^\star) - 1) = 0, i = 1, 2, \cdots, N \\
&y^{(i)}(\mathbf{w}^\star\cdot\mathbf{x}^{(i)} + b^\star) - 1 \geq 0, i = 1, 2, \cdots, N \\
&(\alpha^{(i)})^\star \geq 0, i = 1, 2, \cdots, N
\end{aligned}
$$

前两个条件是令偏导数等于0得到的解，代回偏导数满足条件。第三个是KKT对偶互补条件。第四个条件是原始问题中的约束条件。最后一个是拉格朗日乘子要满足的条件。

由第一个条件可得

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

可以看到$\mathbf{w}^\star$是由N个含有拉格朗日乘子$(\alpha^{(i)})^\star$的项相加得到。如果所有的拉格朗日乘子都等于0，则$\mathbf{w}^\star = \mathbf{0}$，而0向量不是原始问题的解，说明不可能所有的$(\alpha^{(i)})^\star$都等于0。

假设存在$(\alpha^{(j)})^\star > 0$，由KKT条件中的对偶互补条件可得，

$$y^{(j)}(\mathbf{w}^\star \cdot \mathbf{x}^{(j)} + b^\star) - 1 = 0$$

表明第j个样本点使得原始问题的不等式约束取到等号。

两边同时乘以$y^{(j)}$得到$(y^{(j)})^2(\mathbf{w}^\star \cdot \mathbf{x}^{(j)} + b^\star) = y^{(j)}$。因为$(y^{(j)})^2 = 1$，

$$\mathbf{w}^\star \cdot \mathbf{x}^{(j)} + b^\star = y^{(j)}$$

将$\mathbf{w}^\star$带入$\mathbf{w}^\star \cdot \mathbf{x}^{(j)} + b^\star = y^{(j)}$，整理后得到

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

根据$\mathbf{w}^\star = \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} \mathbf{x}^{(i)}$和$b^\star = y^{(j)} - \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} (\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)})$可知，我们可以用求解对偶问题得到的解$\alpha^\star$去表达原始问题的解$(\mathbf{w}^\star, b^\star)$。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

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

根据$\mathbf{w}^\star = \sum_{i=1}^{N} (\alpha^{(i)})^\star y^{(i)} \mathbf{x}^{(i)}$可知，分类超平面只与$(\alpha^{(i)})^\star > 0, i = 1, 2, \cdots, N$对应的样本点$(\mathbf{x}^{(i)}, y^{(i)})$有关，而$\alpha^{(i)})^\star = 0$对应的样本点对$(\mathbf{w}^\star, b^\star)$没有贡献。我们称$(\alpha^{(i)})^\star > 0$为支撑向量。

其实很好理解，根据KKT对偶互补条件可知，$(\alpha^{(i)})^\star > 0$，则$y^{(i)}(\mathbf{w}^\star \cdot \mathbf{x}^{(i)} + b^\star) = 0$，则$\mathbf{w}^\star \cdot \mathbf{x}^{(i)} + b^\star = \pm 1$，表明该样本点距离分离超平面的距离等于1，即点在分离超平面上。

以上是有关线性可分支撑向量机的内容。