支持向量机的学习

给定两个类别的训练样本集 $D=\left\{\left(\mathbf{x}_1,z_1\right),\ldots,\left(\mathbf{x}_n,z_n\right)\right\}$， $\mathbf{x}_i$ 为训练样本的特征矢量， z_i 是样本的类别标识：

$$z_i=\begin{cases}+1, & \mathbf{x}_i\in\omega_1\\ -1, & \mathbf{x}_i\in\omega_2\end{cases}$$

线性可分的情况

我们先来讨论样本集 D 是线性可分的情况。作为最优分类超平面来说，首先需要能够对样本集 D 正确分类，亦即需要满足：

$$z_i\left(\mathbf{w}^t\mathbf{x}_i+w_0\right)>0，\quad \forall i=1,\ldots,n \tag{6.23}$$

由于：

$$z_i\left(\mathbf{w}^t\mathbf{x}_i+w_0\right)\geq\min_{1\leq j\leq n}\left[z_j\left(\mathbf{w}^t\mathbf{x}_j+w_0\right)\right]=b_{\min}>0$$

$b_{\min}$ 是训练样本集中距离分类超平面最近样本的函数间隔。当权值矢量 $\mathbf{w}$ 和偏置 w_0 同时乘上一个正数 $1/b_{\min}$ 时，对应超平面的位置是不会发生变化的，因此（6.23）式的条件可以重写为：

$$z_i\left(\mathbf{w}^t\mathbf{x}_i+w_0\right)\geq 1，\quad \forall i=1,\ldots,n \tag{6.24}$$

这样做的好处是通过适当调整权值矢量和偏置，使得最优超平面到样本集的函数间隔 b 变为了 1，亦即支持面与分类界面之间的函数间隔为 1。

最优分类超平面的第 2 个条件是要使得与样本集之间的几何间隔 γ 最大，在函数间隔为 1 的条件下有：

$$\gamma=\frac{1}{\left\|\mathbf{w}\right\|}$$

这样我们就得到了训练样本集线性可分条件下学习最优分类超平面的优化准则和约束条件：

原始优化问题

$$\min_{\mathbf{w},w_0} J_{SVM}\left(\mathbf{w},w_0\right)=\frac{1}{2}\left\|\mathbf{w}\right\|^2 \tag{6.25}$$

约束：

$$z_i\left(\mathbf{w}^t\mathbf{x}_i+w_0\right)\geq 1，\quad i=1,\ldots,n$$

这是一个典型的线性不等式约束条件下的二次优化问题。在支持向量机的学习算法中，一般并不是直接求解这个原始问题，而是转而求解与其等价的对偶问题。附录 B.6 介绍了求解约束优化问题的 Lagrange 乘数法，下面我们从另外一个角度来分析一下约束优化的对偶问题。

在说明对偶问题之前，先来看定义在矢量 $\mathbf{u}$ 和 $\mathbf{v}$ 上的函数 $f\left(\mathbf{u},\mathbf{v}\right)$ 的 min-max 和 max-min 问题。

$$\text{min-max 问题：}\begin{cases}f^*\left(\mathbf{u}\right)=\max_{\mathbf{v}} f\left(\mathbf{u},\mathbf{v}\right)\\ \min_{\mathbf{u}} f^*\left(\mathbf{u}\right)=\min_{\mathbf{u}}\max_{\mathbf{v}} f\left(\mathbf{u},\mathbf{v}\right)\end{cases}$$

max-min 问题：$\begin{cases} f^*(\mathbf{v}) = \min\limits_{\mathbf{u}} f(\mathbf{u},\mathbf{v}) \\ \max\limits_{\mathbf{v}} f^*(\mathbf{v}) = \max\limits_{\mathbf{v}} \min\limits_{\mathbf{u}} f(\mathbf{u},\mathbf{v}) \end{cases}$

可以证明上述两个问题如果有解存在的话，必在同一点取得最优解[7]，亦即：

$$\min_{\mathbf{u}} \max_{\mathbf{v}} f(\mathbf{u},\mathbf{v}) = \max_{\mathbf{v}} \min_{\mathbf{u}} f(\mathbf{u},\mathbf{v}) = f(\mathbf{u}^*,\mathbf{v}^*) \tag{6.26}$$

也就是说函数 $f(\mathbf{u},\mathbf{v})$ 先对 $\mathbf{u}$ 取最小值再对 $\mathbf{v}$ 取最大值，还是先对 $\mathbf{v}$ 取最大值再对 $\mathbf{u}$ 取最小值的结果是一样的，两个优化问题的求解顺序可以颠倒。

下面根据（6.25）的原始优化问题构造 Lagrange 函数：

$$L(\mathbf{w},w_0,\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i\left[z_i\left(\mathbf{w}^t\mathbf{x}_i + w_0\right) - 1\right] \tag{6.27}$$

其中 $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_n)^t$，$\alpha_i \ge 0$ 是针对（6.25）优化问题中每个约束不等式引入的 Lagrange 系数。

考虑 Lagrange 函数关于矢量 $\boldsymbol{\alpha}$ 的最大化问题：$\max\limits_{\boldsymbol{\alpha}\ge\mathbf{0}} L(\mathbf{w},w_0,\boldsymbol{\alpha})$，当 $z_i\left(\mathbf{w}^t\mathbf{x}_i + w_0\right) > 1$ 时，Lagrange 函数在 $\alpha_i = 0$ 处取得最大值；而当 $z_i\left(\mathbf{w}^t\mathbf{x}_i + w_0\right) = 1$ 时，α_i 可以大于 0。总之当 Lagrange 函数取得最大值时，（6.27）求和式中的两个乘积项 α_i 和 $z_i\left(\mathbf{w}^t\mathbf{x}_i + w_0\right) - 1$ 必有一项为 0，因此：

$$\max_{\boldsymbol{\alpha}} L(\mathbf{w},w_0,\boldsymbol{\alpha}) = J_{SVM}(\mathbf{w},w_0) = \frac{1}{2}\|\mathbf{w}\|^2$$

这样（6.25）的原始优化问题就等价于一个 min-max 问题。考虑到（6.26）式，这个问题也等价于一个 max-min 问题：

$$\min_{\mathbf{w},w_0} J_{SVM}(\mathbf{w},w_0) = \min_{\mathbf{w},w_0} \max_{\boldsymbol{\alpha}} L(\mathbf{w},w_0,\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \min_{\mathbf{w},w_0} L(\mathbf{w},w_0,\boldsymbol{\alpha})$$

首先计算 Lagrange 函数针对 $\mathbf{w}$ 和 w_0 的最小化问题：

$$\frac{\partial L(\mathbf{w},w_0,\boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i \tag{6.28}$$

$$\frac{\partial L(\mathbf{w},w_0,\boldsymbol{\alpha})}{\partial w_0} = -\sum_{i=1}^{n}\alpha_i z_i = 0 \quad \rightarrow \quad \sum_{i=1}^{n}\alpha_i z_i = 0 \tag{6.29}$$

将(6.28)和(6.29)重新代入 Lagrange 函数：

$$\begin{aligned} L(\mathbf{w},w_0,\boldsymbol{\alpha}) &= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i\left[z_i\left(\mathbf{w}^t\mathbf{x}_i + w_0\right) - 1\right] \\ &= \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i\right)^t\left(\sum_{i=1}^{n}\alpha_i z_i \mathbf{x}_i\right) - \sum_{i=1}^{n}\left\{\alpha_i z_i\left(\sum_{j=1}^{n}\alpha_j z_j \mathbf{x}_j\right)^t \mathbf{x}_i + \alpha_i z_i w_0 - \alpha_i\right\} \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j z_i z_j \mathbf{x}_i^t\mathbf{x}_j - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j z_i z_j \mathbf{x}_i^t\mathbf{x}_j - w_0\sum_{i=1}^{n}\alpha_i z_i + \sum_{i=1}^{n}\alpha_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j z_i z_j \mathbf{x}_i^t\mathbf{x}_j \end{aligned}$$

此时，Lagrange 函数只与优化矢量 $\boldsymbol{\alpha}$ 有关，而与 $\mathbf{w},w_0$ 无关。因此，可以由 Lagrange 函数针对 $\boldsymbol{\alpha}$ 的最大化，同时考虑（6.29）式的约束，得到原始问题的对偶优化问题：

对偶优化问题

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j z_i z_j \mathbf{x}_i^t \mathbf{x}_j \qquad (6.30)$$

约束：

$$\alpha_i \geq 0, \quad i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i z_i = 0$$

原始优化问题和对偶优化问题都是典型的线性不等式约束条件下的二次优化问题，求解两者中的任何一个都是等价的。但 SVM 算法一般求解的是对偶问题，因为它有如下两个特点：

1、 对偶问题不直接优化权值矢量 $\mathbf{w}$，因此与样本的特征维数 d 无关，只与样本的数量 n 有关。当样本的特征维数很高时，对偶问题更容易求解；
2、 对偶优化问题中，训练样本只以任意两个矢量内积的形式出现，因此只要能够计算矢量之间的内积，而不需要知道样本的每一维特征就可以进行优化求解。

以上两个特点使得我们可以很容易地将“核函数”引入到算法中，实现非线性的 SVM 分类。