一、基本概念

1、 支持向量：SVM 就是寻找一个最优的决策边界，距离两个类别的最近的样本最远，其中最近的样本点称为支持向量。

2、 间隔：间隔（margin）是指从 SVM **最优决策边界**到最接近它的训练样本之间的距离。

二、SVM 原理

1、自然语言描述：

SVM 是一种二类分类模型。它的基本模型是在特征空间中寻找**间隔最大化**的分离**超平面**的线性分类器。

- 当训练样本**线性可分**时，通过**硬间隔最大化**，学习一个**线性分类器**，即线性可分支持向量机；
- 当训练数据**近似线性可分**时，引入**松弛变量，通过软间隔最大化**，学习一个**线性分类器**，即线性支持向量机；
- 当训练数据**线性不可分**时，通过使用**核技巧及软间隔最大化**，学习非线性支持向量机。

*以上各种情况下的数学推到应当掌握，硬间隔最大化（几何间隔）、学习的对偶问题、软间隔最大化（引入松弛变量）、非线性支持向量机（核技巧）。*

2、数学语言描述：凸二次规划问题（原始问题目标函数）

数据集：$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$ $y_i\in\{0,1\}$

超平面：$w^Tx+b=0$

其中 $w=[w_1,w_2,\cdots,w_d]^T$ 为法向量，

$b$ 为位移，决定超平面与原点距离

间隔表示：所有样本到分割超平面最短距离

$$r=\min_i r_i=\min_i \frac{|w^Tx_i+b|}{\|w^T\|_2}=\min \frac{y_i(w^Tx_i+b)}{\|w^T\|_2}.$$

SVM 目标：最大化超平面 $\max r$

优化问题 $\Longrightarrow$

$$\max_{w,b} r.$$

$$s.t.\ \frac{y_i(w^Tx_i+b)}{\|w^T\|_2}\geq r$$

支持向量表示：$y_i(w^Tx_i+b)=1$ 的样本点

令 $\|w^T\|_2 r=1$，则问题等价为

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

$$s.t.\ y_i(w^Tx_i+b)\geq 1$$

$\Rightarrow$ 将问题进一步化简，得到最终问题

$$\min_{w,b} \|w\|_2^2$$

$$s.t.\ y_i(w^Tx_i+b)\geq 1$$

3、模型求解：构造对偶问题

（1）为什么要将求解 SVM 的原始问题转换为其对偶问题

- 对偶问题通过引入对偶变量(拉格朗日乘子)将原问题中的不等式约束转化为等式约束，将原始问题的凸二次规划问题转化为了单变量的二次规划问题，使得问题更易处理。
- 改变了问题的复杂度。由求特征向量 w 转化为求比例系数α，在原始问题下，求解的复杂度与样本的维度有关，即 w 的维度。在对偶问题下，只与样本数量有关（对应为 m）。
  - SVM 原始问题模型严重依赖于数据集的维度 d，如果维度 d 太高就会严重提升运算时间。
  - 对偶问题事实上把 SVM 从依赖 d 个维度转变到依赖 m 个数据点，考虑到在最后计算时只有支持向量才有意义，所以这个计算量实际上比 m 小很多。
- 求解更高效，因为只用求解α系数，而α系数只有在支持向量才非 0，其它全部为 0。
- 方便核函数的引入，进而推广到非线性分类问题。

（2）转化为对偶问题的过程

带约束的优化问题的通用解法为拉格朗日乘子法.

- 拉格朗日乘子法
  - 第一步：对每条不等式约束引入拉格朗日乘子 $\alpha_i \geq 0$ 得到拉格朗日函数

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

其中 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_m]^T$ 为拉格朗日乘子向量.

  - 第二步：固定 $\boldsymbol{\alpha}$，令 $L(\boldsymbol{w}, b, \boldsymbol{\alpha})$ 对 $\boldsymbol{w}$ 和 $b$ 的偏导数为零可得

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \mathbf{0} \Rightarrow \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} = 0 \Rightarrow 0 = \sum_{i=1}^{m} \alpha_i y_i \Longrightarrow \text{绝大部分（不需要惩罚的样本）} \alpha_i \text{为} 0$$

  - 第三步：回代 (将上述第二步中第一式代入第一步的拉格朗日函数中) 消去 $\boldsymbol{w}$ 和 $b$

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \text{ s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, m$$

此外，由于 $\alpha_i$ 为对偶问题的解，同时$\boldsymbol{w}$，$b$为原问题的解，因此要求求得的解还需满足$\alpha_i(y_i f(\boldsymbol{x}_i) - 1) = 0$（周志华《机器学习》附录B.1），这三个约束一起称为 KKT (Karush-Kuhn-Tucker) 条件，即要求 ($\forall\, i$)

$$\begin{cases} \alpha_i \geq 0, \\ y_i f(\boldsymbol{x}_i) - 1 \geq 0, \\ \alpha_i(y_i f(\boldsymbol{x}_i) - 1) = 0. \end{cases}$$

（3）序贯最小化算法 SMO：求解对偶问题，得到α

- 核心思想：

- **基本思路**：不断执行如下两个步骤直至收敛.
  - 第一步：选取一对需更新的变量 $\alpha_i$ 和 $\alpha_j$
  - 第二步：固定 $\alpha_i$ 和 $\alpha_j$ 以外的参数，求解对偶问题更新 $\alpha_i$ 和 $\alpha_j$.
    - 仅考虑 $\alpha_i$ 和 $\alpha_j$ 时，对偶问题的约束变为

$$\alpha_i y_i + \alpha_j y_j = -\sum_{k\neq i,j} \alpha_k y_k, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0.$$

用一个变量表示另一个变量, 回代入对偶问题可得一个单变量的二次规划, 该问题具有闭式解.

- 两变量选择问题：

a. 由于最终所有计算得到的 $\alpha_i$ 都会满足KKT 条件，因此如果存在某个 $\alpha_i$ 不满足KKT 条件，那么目标函数

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

会最快衰减. 因此，应该选择违背 KKT 条件的那些变量先更新.

b. 比较各变量所对应的目标函数值减幅的复杂度过高，因此 SMO 采用了一个启发式：使选取的两变量所对应样本之间的间隔最大. 一种直观的解释是，这样的两个变量有很大的差别，与对两个相似的变量进行更新相比，对它们进行更新会带给目标函数值更大的变化.

a. 固定其他变量后优化 $\alpha_i$, $\alpha_j$，实际上是将约束项 $\sum_{i=1} \alpha_i y_i = 0$ 转变成 $\alpha_i y_i + \alpha_j y_j = c$，这里 $c = -\sum_{k\neq i,j} \alpha_k y_k$，而进一步 $\alpha_j = c y_j - \alpha_i y_i y_j$

b. 将 $\alpha_j$ 代入式子

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \text{ s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, m$$

消去 $\alpha_j$, 上式变成关于 $\alpha_i$ 的一个二次规划问题, 仅有的约束是 $\alpha_i \geq 0$, 这样的二次规划存在封闭形式的解, 不必调用数值优化算法即可高效地计算更新后的 $\alpha_i$ 和 $\alpha_j$, 因此算法效率高.

- 如果已经求解得到 $\boldsymbol{\alpha}$，那么由 $\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$，可以计算得到 $\boldsymbol{w}$. 如何求解 $b$，暂时先放一下
- 解出 $\boldsymbol{\alpha}$, 求出 $\boldsymbol{w}$ 和 $b$, 即可得超平面所对应的模型

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b$$

三、核函数

1、核函数作用：

当样本在原始空间线性不可分时，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。在求解对偶问题时仅需计算特征向量的内积。

(1)、引入了核函数，把高维向量的内积转变成了求低维向量的内积问题。即在特征空间的内积等于它们在原始样本空间中通过核函数 K 计算的结果。

(2)、核函数是一种表征映射、实现内积逻辑关系且降低计算复杂度的一类特殊函数，定义为K(x,y)=<φ(x),φ(y)>。这里的内积是一种在高维空间里面度量数据相似度一种手段，一方面数据变成了高维空间中线性可分的数据；另一方面不需要求解具体的映射函数，只需要给定具体的核函数即可。

2、模型的改变：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boxed{k(\boldsymbol{x}_i,\boldsymbol{x}_j)} - \sum_{i=1}^{m}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, m$$

- 最后计算得到分割超平面 $\boldsymbol{f}(\boldsymbol{x}) = \sum_{i=1}^{m}\alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_j) + b$ . 这里显示出**模型最优解可通过训练样本的核函数展开**，这一展式亦称 **“**支撑向量展式**”** (support vector expansion)

3、 Mercer 定理:任何半正定的函数都可以作为核函数.也就是说只要一个对称函数所对应的核矩阵半正定，它就能作为核函数使用。通过和函数线性运算核内积运算可以得到新的核函数。

四、软间隔

1、 定义：

- 所有样本都必须划分正确，这称为“硬间隔”（hard margin).
- 允许某些样本不满足约束，这称为“软间隔”（soft margin)

2、 数学模型

如果有部分样本被错分了，那么大于等于号将不一定对每个样本都成立. 但我们可以对每一个样本设定一个参数 $\xi_i$> 0, 使得 $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1\text{-}\xi_i$ 成立. 这个 $\xi_i$ 可以表示错分的程度，称为松弛变量 (slack variables).

当 $\xi_i$ 充分大时，训练样本点 $(x_i, y_i)$ 总可以满足上述约束条件。但是，应该避免 $\xi_i$ 取太大的值，因此，在目标函数里对它进行惩罚

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \quad \text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0, i = 1, 2, \ldots, m$$

这里 $C > 0$ 是一个惩罚参数, 它是调整误差允许范围的参数，这就是常用的”软间隔支撑向量机”.

这里 $C$ 平衡最小化 $\|\boldsymbol{w}\|_2^2$ 以增大间隔和最小化 $\sum_{i=1}^{m}\xi_i$ ：

(1) $C$ 大⟶表明我们更关心的是划分的正确性，间隔可以 **“**瘦**”** 一点，但是划分错误的点要少一点. $C$ 为无穷大时, 迫使所有样本都满足 $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1$ .
(2) $C$ 小⟶表明我们想要的是更 **“**胖**”** 一点的边界，划分错误的点多一点没有关系.

3、 对偶问题

同样通过拉格朗日乘子法，可以得到软间隔支撑向量机的拉格朗日函数:

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \xi, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i$$

$\alpha_i \geq 0, \mu_i \geq 0$ 为拉格朗日乘子. 上式分别对 $\mathbf{w}, b, \xi_i$ 求偏导并置为零, 可得

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i,$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i, \quad \Longrightarrow \text{绝大部分（不需要惩罚的样本）} \alpha_i \text{为 } 0$$

$$C = \alpha_i + \mu_i.$$

将上面求得的导数代入拉格朗日函数，得到对偶问题

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

$$\text{s.t.} \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, m$$

综合约束项, 对于软间隔支撑向量机，KKT 条件要求

$$\begin{cases} 0 \leq \alpha_i \leq C, \ \mu_i \geq 0, \\ y_i f(\boldsymbol{x}_i) - 1 + \xi_i \geq 0, \\ \alpha_i(y_i f(\boldsymbol{x}_i) - 1 + \xi_i) = 0, \\ \xi_i \geq 0, \ \mu_i\xi_i = 0. \end{cases}$$