第八章 支持向量机SVM

一、基本思想

1. 分类样本的基本思想
   当我们试图找到不同类别的最大超平面的时候SVM算法非常有用
   针对两个特征所构成的超平面，输入空间中只有两类点，里面有许多条分界线将我们的类别划分，但是我们如何找到全部里面最好的划分直线或者超平面呢
   针对二维空间来说，就是找到一条直线满足$Ax + By + C = 0$
   针对n维空间来说，就是找到一个超平面满足$w^T x + b = 0$
2. **支持向量机的优化目标——最大化间隔**
   如何确定正中间的超平面：使用支持向量机，通过求解间隔最大化来确定最佳超平面
   选择距离两侧分类点最大间隔超平面的原因：选择距离两侧分类点距离最大的超平面，即选择”正中间“，容忍性好，鲁棒性高，泛化能力强，如果我们选择了一个有较低间隔分类器，那么它对于新数据就有很大的可能误判
   在线性分类器的超平面周围间隔范围内不存在任何训练数据点
   针对不同间隔的分类直线，我们要尽可能选择最大的间隔的，以消除误差和一些过拟合现象

   针对我们要选择距离最近数据点距离最大的超平面，这个超平面可以将不同类别样本点分开，如果这个超平面存在，它就被称为最大间隔超平面或硬间隔
3. 基本概念
   - 超平面：是在特征空间中用来分别不同数据点的决策边界，在线性分类中，它就是一个线性方程
   - **支持向量**：即与超平面距离最近的数据点，它们在决定超平面及边距方面起着关键作用。将超平面平移刚好与其中的数据点相交对应数据点叫做支持向量。例如在二维平面上，平移分割超平面的最优直线得到刚好划分两个类别的的边界时，在平行线上的数据点，叫做支持向量
   - 边际：支持向量与超平面之间的距离。 支持向量机算法的主要目标是最大化边缘值。边缘值越宽，表示分类性能越好。
   - 线性可分：一个训练集线性可分是指对训练集样本$\{(x_i, y_i)|i = 1, 2, \ldots, N\}$，对于指定的$(W, b)$，使得对于任意的$i = 1, 2, \ldots, N$，都有若$y_i = +1$，则$W^T x_i + b >= 0$，若$y_i = -1$，则$W^T x_i + b <= 0$
4. SVM的类别：
   - 硬间隔：是指将不同类别数据点恰当分开并且没有任何错误分类
   - 软间隔：当数据无法完美地被划分或者包含越界的点，SVM使用软边界允许一些错误分类
   - 使用RBF核函数的非线性支持向量机：一种用在支持向量机的方法，使得支持向量机可以使用线性分类器分类非线性数据

二、硬间隔——LSVM

硬间隔不允许出现任何错误分类

1. **SVM模型建立**
   任何超平面可以被写作满足以下方程的数据点的集合：

$$W^Tx+b=0$$

   根据点到超平面的距离公式得到

$$\frac{w^Tx+b}{||w||}>=d\ ,y=1$$

$$\frac{w^Tx+b}{||w||}<=d\ ,y=-1$$

   所以说针对两个边界直线的约束可以这么表示

$$W^TX+b>=1\ \ if\ y_i=+1$$

$$W^TX+b<=-1\ \ if\ y_i=-1$$

   这两个公式综合起来可以表示为

$$y_i(W^TX+b)>=1$$

   所以说这个优化的方程可以变为如下模型：
   目标函数（里面的$\frac{1}{2}$仅仅为了求导方便）：

$$min\ \frac{1}{2}||w||^2$$

   约束条件：

$$s.t.\ y_i(wx_i+b)>=1$$

   目标函数是最小化$||w||^2$相当于最大化间隔，约束确保所有点都会被正确分类
   针对上述模型需要列举以下事实：
   - $w^Tx+b=0$与$aw^Tx+ab=0$是同一个平面，其中$a\in N^+$
   - 点到平面的公式

$$d=\frac{|w_1x_0+w_2x_0+b|}{\sqrt{(w_1^2+w_2^2)}}$$

     由此可以得到向量x_0到超平面$W^Tx+b=0$的距离为

$$d=\frac{|W^Tx_0+b|}{||w||}$$

   - 我们可以用a去缩放(w,b)，得到(aw,ab)，因为这表示的是同一个平面，最终使得最终在支持向量x_0上的值$|W^Tx_0+b|=1$，此时支持向量与平面的距离就为

$$d = \frac{1}{||w||}$$

因此我们就转化为求$||w||$最大化问题

2. **SVM模型的求解**

这是一个凸优化问题：目标函数是二次项，限制条件是一次性，最后导致的结果是要么无解要么是极小值

针对这个问题可以使用拉格朗日乘数法解决，借助拉格朗日乘数可以将其转化为无约束问题求解，针对上述问题我们可以建立拉格朗日函数

$$max_{a>=0}min_{w,b}L(w,b,\alpha) = \frac{1}{2}w \cdot w - \sum_i \alpha_i y_i (w \cdot x_i + b) + \sum_i \alpha_i$$

分别对w和b求偏导，并且令其为0，得到

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

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

将w代回拉格朗日函数并且简化得到

$$L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

约束为： $\alpha_i >= 0, \sum_i \alpha_i y_i = 0$

其中与支持向量所对应的拉格朗日乘子α_i不为0

只有当支持向量改变的时候超平面才会发生移动，如果我们删除其他的点超平面不会发生变化

支持向量的特别之处在于，他们是确定最大间隔超平面的训练数据点，因此他们决定了超平面的形状

如果移动了支持向量，重新训练SVM这会使得结果超平面发生变化，轻微移动其他点则不会引发这种变化。

SVM的解主要是由支持向量决定，而不是所有训练样本，如果移除的是非支持向量，超平面不会发生改变

如果数据完全线性可分那么就不需要使用松弛变量，理论上数据完全线性可分使用硬间隔SVM足够，不必使用松弛变量，但是实际上数据可能会有噪声，使用松弛变量可以提高模型的鲁棒性，防止过拟合

SVM的对偶形式优化问题只依赖样本的内积x_i，x_j，和标签y_i就可以求解α，从而得到完整的SVM模型

三、软间隔

1. 软间隔的目的

如果我们的数据不是线性可分的怎么办呢，数据不能线性可分由于：

- 他们的潜在模式不是线性可分的

- 噪声已经污染了数据，导致数据所产生的潜在模式呈现出线性不可分的排列
  使用经典的SVM已经不足以处理前者的问题，但是针对后者我们可以使用软间隔来改善经典的SVM方法

2. 引入松弛变量的SVM建模：
   一般来说在训练数据中间隔大小和错误数量存在一种权衡关系
   通过引入松弛变量来放松其中的约束条件，松弛变量被引进，来允许某些约束条件得以放宽，这里对于每一个数据点引入了一个松弛变量ξ，位于间隔内的点的数量需要尽可能少，因此我们也对松弛变量加以限制，乘了一个惩罚因子C，它代表对松弛变量的惩罚力度不能让ξ特别大。这个惩罚因子C是预先设置的
   松弛变量允许输入可以更加靠近超平面，甚至在错误的一边，但是这里针对松弛变量对于目标函数有一个惩罚，后面的一项叫做正则项

$$min(\frac{1}{2}w^Tw + C\sum_{i=1}^{n}\xi_i), C > 0$$

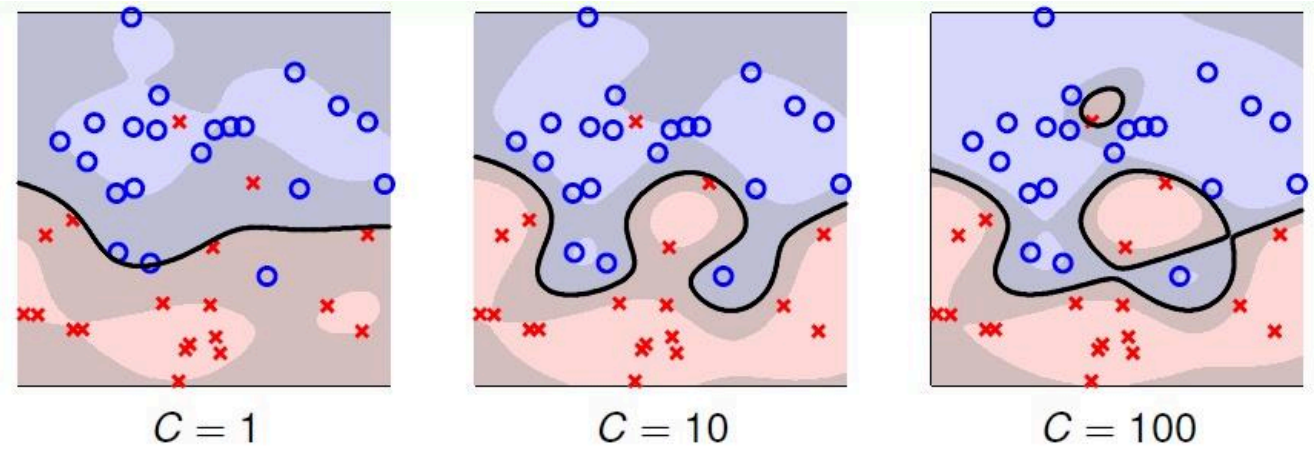
   约束条件：

$$\xi_i >= 0,\ y_i(w^Tx_i + b) >= 1 - \xi_i,\ i = 1, \ldots, n$$

   其中ξ = 0代表该样本完全遵守规矩
   0<ξ<1代表样本略微违规，样本仍然在正确的一侧，但是距离超平面小于1
   ξ >1代表样本已经跨越超平面，被错误分类

3. C取值比较

	C取较小值	**C取较大值**
目的	SVM变得非常松散，可能会错误分类一些数据点，来获得一个简化的解决方案	SVM会变得非常严格，严格保持松弛变量特别小，尝试将所有数据点划分在超平面正确的一侧
间隔	大	小
模型	简单	复杂
问题	欠拟合	过拟合
边界	平坦	蜿蜒曲折



四、核技巧

初始的输入空间可以被映射到一些更高维度的特征空间，当训练数据是可分的，如此我们仍然可以将其转化为线性可分问题来解决

1. 核函数
   定义了一个高纬映射$\phi(x)$，将一个低维的向量x映射到高纬空间$\phi(x)$
   将原始空间内的向量作为输入，在特征空间中计算它们的点积结果

   $$k(x,z) =< \phi(x), \phi(z) >$$

   使用SVM做非线性分类的时候不需要显式地将数据映射到高维特征空间，也就是说即使理论上说我们需要将数据映射到一个高纬度甚至说无限延伸的空间中，但是我们不需要真正计算这个$\varphi(x)$，换句话说我们不需要知道它的显式表达式，我们仅仅需要知道

   $$K(x_1,x_2) = \phi(x_1)^T\phi(x_2)$$

   我们只需要计算核函数就等价于在高维空间中计算内积，因为核函数一定对应某个特征映射，但是我们不需要也不必知道$\phi(x)$的样子
   总结得到：
   已知 K，一定能找到一个特征空间 H 和一个映射 φ(x)，使得
   $K(x,z) = \langle \phi(x), \phi(z) \rangle$
   任何合法核函数 $K(x,z)$都隐含着一个特征空间 H 和映射 φ(x)，
   即使 φ 无法写出，也依然存在，你也不需要写。
2. SVM中核函数的替换
   由上面我们使用拉格朗日乘子法，推导得到SVM的对偶形式，因此我们可以从对偶问题出发，根据$x \to \phi(x)$的映射关系，原有的对偶问题的目标函数就会变成

   $$max_\alpha(\sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(x_i,x_j))$$

   对应超平面就会变成

   $$f(x) = \sum_i \alpha_i y_i K(x_i,x) + b$$

   推理得到的非线性支持向量机

   $$f(x) = sign(\sum_i \alpha_i y_i K(x_i,x) + b)$$

3. $K(x_1,x_2) = \phi(x_1)^T\phi(x_2)$的充要条件
   - 交换性：$K(x_1,x_2) = K(x_2,x_1)$
   - 半正定性：对任意常数C_i，向量X_i有：

   $$\sum_{i=1}^N\sum_{j=1}^N C_iC_jK(x_i,x_j) >= 0$$

4. 核函数
   - 高斯核：

$$K(x_1, x_2) = exp(-\frac{||x_1 - x_2||^2}{2\sigma^2})$$

   - 多项式核：

$$(x_1^T x_2 + 1)^d$$

五、多分类支持向量机

1. 一对一方法
   将多分类问题划分为多个二分类问题，针对每个类别设置一个二分类器，假设有K个类别则需要训练$\frac{K(K+1)}{2}$个分类器，每个分类器针对两个类别
   我们需要一个超平面将每个二分类划分开，忽视第三类别数据点，只使用当前两个类别的数据
   很大程度上受到数据集不均衡问题影响很少，但是计算上更花费资源
2. 一对剩余方法
   我们需要找到一个超平面一次把一个类与其他类别划分开，假设有K个类别则针对每个类使用一个分类器，总共训练K个分类器
   容易受到样本不均衡的影响，实现简单
   样本不均衡：一个类别的数据远远小于其他类别样本数总和

六、总结

1. 核化支持向量机的优缺点：

优点	缺点
他们在一系列数据集上表现很好	效率随着训练数据集的大小增加而降低
它们具有很强的通用性：可以指定不同的内核函数，或者还可以为特定的数据类型定义自定义内核。	需要对输入数据进行仔细的规范化处理，并对参数进行调整优化。
它们适用于高维和低维数据。 在数据有限的情况下效果显著。	计算量大。支持向量机的计算成本可能较高，尤其是在处理大型数据集时。训练所需的时间和内存需求会随着训练样本数量的增加而显著增加。

2. 支持向量中的核函数
   - 线性不可分的数据：在现实世界中，很多数据集并不是线性可分的，直接应用线性 SVM 无法找到有效的分类边界
   - 映射到高维空间：将原始数据通过某种非线性映射，转换到一个更高维的空间处理线性不可分数据，但计算的时候会出现维度灾难的问题，这种映射可以是显式的，也可以是隐式的

- 计算效率（核技巧）：核函数可以隐式的计算两个样本的内积，而不需要实际执行映射，核函数在原空间的值等于在高维空间中样本的内积，从而大大简化了计算流程，同时也解决了维度灾难问题