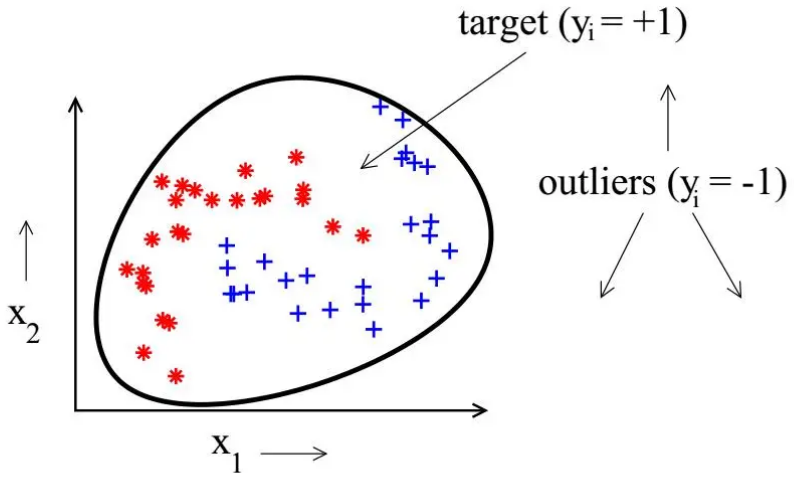

one class classification

马鹤宁
中国科学院大学 工学硕士

关注她

123 人赞同了该文章

在科研生涯中，导师要求做一个区分器，奈何只能得到一个类别的数据集，百思不得其解，多方查找，了解one class classification，随在此对此类问题进行详细介绍。

分类问题，例如二分类和多分类，由于多分类问题都可以解体成多个二分类问题，所以，一般来说，二分类问题被看做是基本的分类问题，在这里就拿二分类问题为例。在一个二分类问题中，元素被分成两个类， ω_1 和 ω_2，被分别标记成+1和-1，或者称为正类和负类，也就是训练集合中的每一个元素 $x_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,d})$ 分属于集合 $y_i \in \{-1, +1\}$ 。分类问题的目标是要从训练集合中学习到一个函数 $y = f(x)$ ，这个函数能够对于一个新的给定的元素x进行预测出其所属类别，而且要尽可能准确。

在二分类里，理想情况下，需要要求训练集中的每类元素的数量巨大且几乎相等，即使在现实世界中，会出现正负类样本失衡的情况，所幸，众多策略已经被提出来解决此类问题，有点偏题，不多赘述。但是如果训练集中仅仅只有一类数据，那么要如何测试新的数据呢？并如何判断它是否与训练数据相似？这就引入了 one class classification。在one-class classification中，仅仅只有一类的信息是可以用于训练，其他类别的(总称为outlier)信息是缺失的，也就是区分两个类别的边界线是通过仅有的一类数据的信息学习得到的。

举例说明，假设有一个apple-pear的数据集，每个元素包含有两个特征，宽度width和重量weight，所属类别为苹果或梨。那么训练数据集的每一个元素可以表示成一个2维特征空间里的一个点，下图中的红色星星点表示苹果，蓝色加号点表示梨。虚线圈表示整个数据集。图1中的黑色实线将苹果和梨完美地进行区分，在二分类里面，类别仅仅包括苹果和梨，那么所有的元素要么是苹果要么是梨，不可能是其他类别。对于one class classification，将apple-pear看做一个整体，为一个类apple-pear，其他的不在这个数据集范围内（图1虚线圈）的则属于outlier，outlier包含非（苹果和梨）的其他所有类别。也就是说对于一个新的元素，假如它在虚线圈中，则说明这个元素的类别为apple-pear，加入处于虚线圈外，则说明这个元素的类别既不是苹果也不是梨，至于是什么，不清楚。

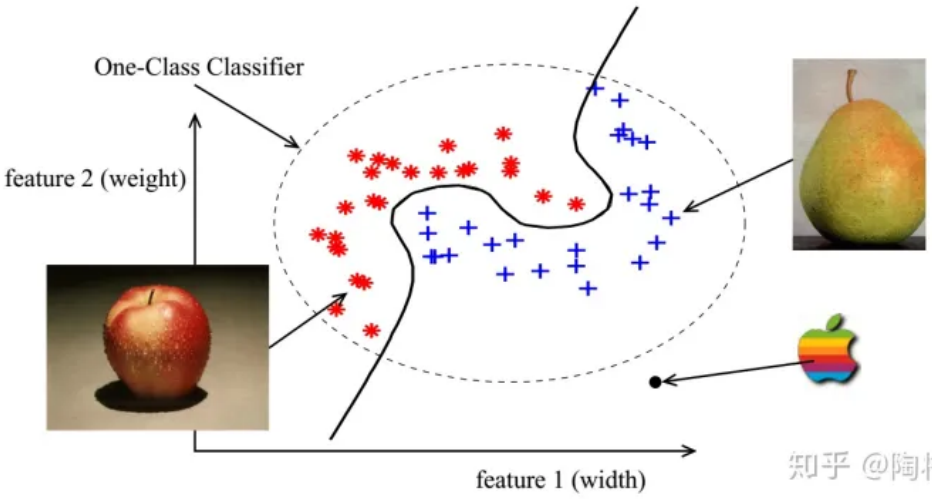

图1: 二分类和one class classification 比较

support vector data description(SVDD)

在one-calss classification中，边界应该全方位包围数据，我们定义一个对数据封闭的边界模型 $f(x;w)$：一个超球面。定义这个超球面的中心点为 a，半径为R，我们需要这个超平面能够包含所有的训练数据，那么也就是说训练集的经验风险损失（empirical error）为0，那也就是说所有数据所有元素 x_i 到超球面的中心a的距离应该严格小于半径R $||x_i - a||^2 \leq R^2$ 。但是为了使得模型更加健壮，也就是说允许一部分outlier存在一定概率被错认为数据集中的元素，那么经验损失就不必须是为0. 引入一个松弛变量slack variable $\xi_i \ (\xi_i > 0) \forall i$，则损失函数则既包含经验风险，还包含结构风险，如下所示：

$$\varepsilon(R, a, \xi) = R^2 + C\sum_i \xi_i$$

限制条件为

$$||x_i - a||^2 \leq R^2 + \xi_i$$
$$\xi_i > 0 \ \ \forall i$$

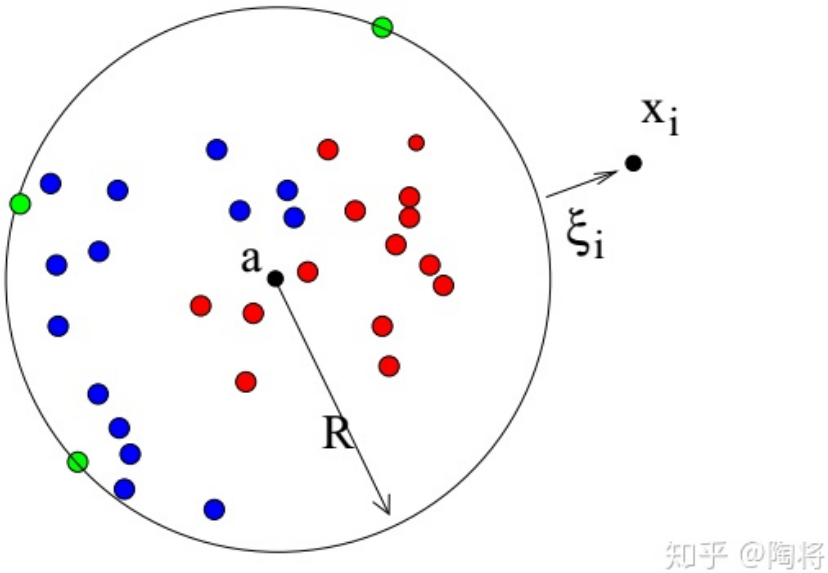

图2：包含数据集的超球面

引入拉格朗日乘子Lagrange multiplier，构建拉格朗日函数为：

$$L(R, a, \xi, \alpha, \gamma) = R^2 + C\sum_i \xi_i - \sum_i \alpha_i \left\{R^2 + \xi_i - (x_i \cdot x_i - 2a \cdot x_i + a \cdot a)\right\} - \sum_i \gamma_i \xi_i$$

拉格朗日乘子 $\alpha_i \geq 0$ 和 $\gamma_i \geq 0$

对每一个变量进行求导，可得到：

$$\frac{\partial L}{\partial R} = 0: \qquad \sum_i \alpha_i = 1$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0: \qquad \mathbf{a} = \frac{\sum_i \alpha_i \mathbf{x}_i}{\sum_i \alpha_i} = \sum_i \alpha_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial \xi_i} = 0: \qquad \gamma_i = C - \alpha_i, \quad \forall_i$$

由于 $\alpha_i = C - \gamma_i$，而且 $\alpha_i \geq 0$； $\gamma_i \geq 0$，由此推出一个新的限制条件 $0 \leq \alpha_i \leq C \, \forall i$

$$\begin{aligned} L(R, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= R^2 - \sum_i \alpha_i R^2 + C \sum_i \xi_i - \sum_i \alpha_i \xi_i - \sum_i \gamma_i \xi_i \\ &\quad + \sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i - 2 \sum_i \alpha_i \mathbf{a} \cdot \mathbf{x}_i + \sum_i \alpha_i \mathbf{a} \cdot \mathbf{a} \qquad \text{(A.8)} \\ &= 0 + 0 + \sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i - 2 \sum_i \alpha_i \sum_j \alpha_j \mathbf{x}_j \cdot \mathbf{x}_i + 1 \cdot \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \\ &= \sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x}_i - \sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \qquad \text{(A.9)} \end{aligned}$$

所以经过上述的推导，最后的损失函数带上限制条件如下面的公式所示：

$$L(R, a, \xi, \alpha, \gamma) = \sum_i \alpha_i x_i \cdot x_i - \sum_{i,j} \alpha_i \alpha_j x_i \cdot x_j$$
$$s.t. \, 0 \leq \alpha_i \leq C \, \forall i$$

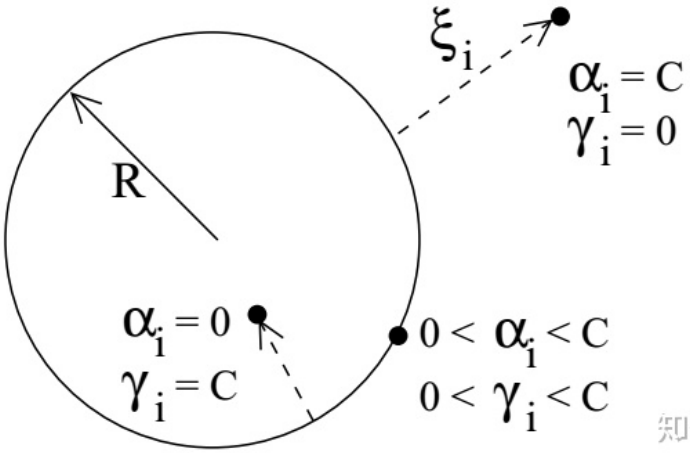

图3：对在超球面不同位置的两个拉格朗日乘子的取值

v-support vector classifier

上述的SVDD算法定义了一个包围数据的超球面，形成一个封闭的边界。而v-SVC的基本思想是放置一个能够将数据dataset和原点origin用最大间隔值(maximal margin)分开的超平面，如图4所示。定义超平面为 ω，数据为 x_i，最大间隔为 ρ，超平面 ω 将数据 x_i 和原点以最大间隔 ρ 分隔，公式表示为：

$$\omega \cdot x_i \geq \rho - \xi_i, \xi_i \geq 0, \forall_i$$

对于要最小化的目标函数，则定义为：

$$min\left(\frac{1}{2}||\omega||^2-\rho+\frac{1}{vN}\sum_i\xi_i\right)$$

其中正则化参数 $v\in(0,1)$ 是一个用户定义的参数，指示能够接受的数据的百分比。这也是v-SVC的由来。

对于一个新的测试数据 z ,判别函数为：

$$f_{v-SVC}(z,\omega,\rho)=I(w\cdot z\leq\rho)$$

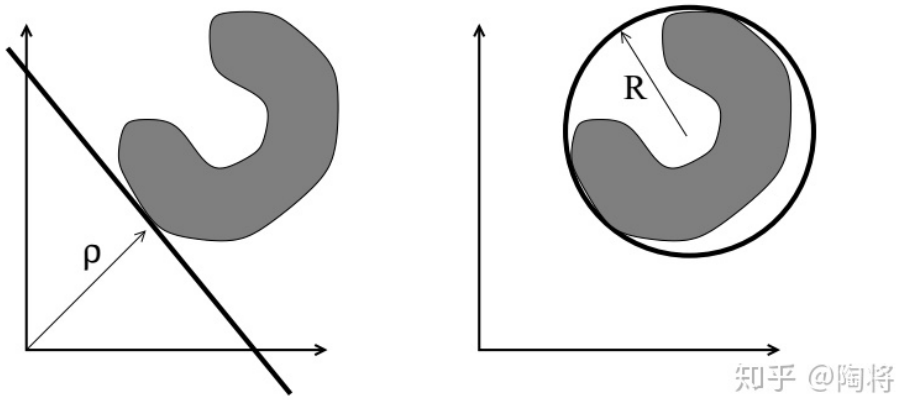

图4：v-SVC and SVDD

v-SVC 在二分类SVM的基础上就很好理解，对S支持向量机不熟悉的，请移步：

逻辑回归LR vs 支持向量机 SVM
blog.csdn.net/weixin_42111770/article/details/82703417#

参考文献：

[1] . one class classification

发布于 2019-03-02 11:09

机器学习　SVM　分类算法

8 条评论 默认 最新

H Yang
还有一些情况是只有positive class 和 unlabeled数据，使用AUC maximization的方法具有理论保证，可以看看我们的工作“A Robust AUC Maximization Framework With Simultaneous Outlier Detection and Feature Selection for Positive-Unlabeled Classification”
ieeexplore.ieee.org/doc...
2019-04-18 回复 5

浅层学习
AUC-based method不会造成测试集泄露吗?
03-28 回复 喜欢

厨师也开挖掘机
之前在工程上遇到过类似的问题，如果训练数据只有一类数据，无异常点，判断测试数据是否是异常点也属于这个"一分类"问题？？感觉你讲的与隶属度有点相似啊
2019-04-10 回复 1

于工不移山
您好，这种异常问题您是用什么思路解决的？最近也遇到了类似的问题
2022-02-26 回复 喜欢

马鹤宁 作者
嗯嗯，就是一分类问题呀
2019-04-10 回复 喜欢

不要辣的
请问下如何可避免过拟合?
2019-07-03 回复 1

arosy
请问是什么样的单类数据呀?
2021-09-12 回复 喜欢

Maya

一篇文章让你弄懂到底什么是classpath

classpath其实就是一个路径而已，我们经常在spring的配置文件中这样写： <property name="configLocation" value="classpath:mybatis/... '>这样配置...

yuan

Python基础介绍 | Class类

自从改用Python做数据挖掘以来，我就很少用面向对象的内容了，那啥是面向对象呢？ 可以这么理解，如果我们能把代码组合成一个一个可以重复使用的类别，那么这能使程序的可复用性更高，以后...

恒仔

Python入门 class类的继承

面向对象的编程带来的主要好处之一是代码的重用，实现各种重用的方法之一是通过继承机制。继承完全可以理解成类之间的父类和子类型关系。 继承概念：继承是类与类的一种关系，是一种子类与...

木头人

class类文件结构（这是一篇非常枯燥的文章）

我们知道，Java文件编译后会产生一个字节码文件（ .class文件），本文介绍字节码文件的文件结构相关内容。前言Java诞生之初就宣称的“一次编译，处处运行”的性质一直是Java的一大特点，而J...

一只小白