SVM入门

传统的机器学习、统计的机器学习、结构风险、小样本、VC维

SVM的特点

SVM简介

- 支持向量机(Support Vector Machine)是Cortes和Vapnik于1995年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。
- 支持向量机方法是建立在统计学习理论的VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力或称泛化能力）。

统计的机器学习和传统的机器学习

- 统计机器学习能够精确的给出学习效果，能够解答需要的样本数等等一系列问题。与统计机器学习的精密思维相比，传统的机器学习基本上属于摸着石头过河，用传统的机器学习方法构造分类系统完全成了一种技巧，一个人做的结果可能很好，另一个人差不多的方法做出来却很差，缺乏指导和原则。

什么是VC

- 所谓VC维是对函数类的一种度量，可以简单的理解为问题的复杂程度，VC维越高，一个问题就越复杂。正是因为SVM关注的是VC维，后面我们可以看到，SVM解决问题的时候，和样本的维数是无关的（甚至样本是上万维的都可以，这使得SVM很适合用来解决文本分类的问题，当然，有这样的能力也因为引入了核函数）。

经验风险

- 机器学习本质上就是一种对问题真实模型的逼近（我们选择一个我们认为比较好的近似模型，这个近似模型就叫做一个假设），真实模型一定是不知道的。既然真实模型不知道，那么我们选择的假设与问题真实解之间究竟有多大差距，我们就没法得知。
- 这个与问题真实解之间的误差，就叫做风险（更严格的说，误差的累积叫做风险）。我们选择了一个假设之后（更直观点说，我们得到了一个分类器以后），真实误差无从得知，但我们可以用某些可以掌握的量来逼近它。最直观的想法就是使用分类器在样本数据上的分类的结果与真实结果（因为样本是已经标注过的数据，是准确的数据）之间的差值来表示。这个差值叫做经验风险Remp(w)。

经验风险最小化

- 传统的机器学习：**基于经验风险最小化，但是泛化能力非常差。**
- 原因：选择了一个足够复杂的分类函数（它的VC维很高），能够精确的记住每一个样本，但对样本之外的数据一律分类错误。
- 经验风险最小化原则：适用的大前提是经验风险要确实能够逼近真实风险才行，但实际上能逼近么？答案是不能，因为样本数相对于现实世界要分类的文本数来说简直九牛一毛，经验风险最小化原则只在这占很小比例的样本上做到没有误差，当然不能保证在更大比例的真实文本上也没有误差。

泛化误差界

- 统计学习因此而引入了泛化误差界的概念，就是指真实风险应该由两部分内容刻画：
  - 一是经验风险，代表了分类器在给定样本上的误差；
  - 二是置信风险，代表了我们在多大程度上可以信任分类器在未知文本上分类的结果。
    - 很显然，第二部分是没有办法精确计算的，因此只能给出一个估计的区间，也使得整个误差只能计算上界，而无法计算准确的值（所以叫做泛化误差界，而不叫泛化误差）。

置信风险->结构风险

- 置信风险与两个量有关：
  - 一是样本数量，显然给定的样本数量越大，我们的学习结果越有可能正确，此时置信风险越小；
  - 二是分类函数的VC维，显然VC维越大，推广能力越差，置信风险会变大。
- 泛化误差界的公式为：
- $R(w) \leq Remp(w) + \Phi(n/h)$
  - 公式中R(w)就是真实风险，Remp(w)就是经验风险，Φ(n/h)就是置信风险。
  - **统计学习的目标从经验风险最小化变为了寻求经验风险与置信风险的和最小，即结构风险最小。**

SVM的特点：

- 基于最小化结构风险的算法；
- 小样本：
  - 并不是说样本的绝对数量少（实际上，对任何算法来说，更多的样本几乎总是能带来更好的效果），而是说与问题的复杂度比起来，SVM算法要求的样本数是相对比较少的。
- 非线性：
  - 是指SVM擅长应付样本数据线性不可分的情况，主要通过松弛变量（也有人叫惩罚变量）和核函数技术来实现，这一部分是SVM的精髓，

SVM与文本分类

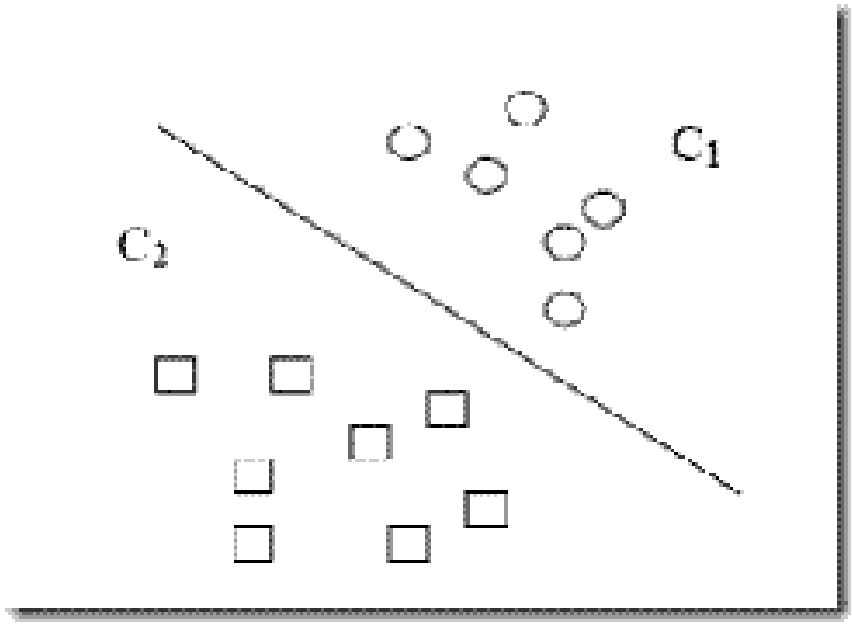
- 高维模式识别是指样本维数很高，例如文本的向量表示，如果没有经过另一系列文章（《文本分类入门》）中提到过的降维处理，出现几万维的情况很正常，其他算法基本就没有能力应付了，SVM却可以，主要是因为SVM 产生的分类器很简洁，用到的样本信息很少（仅仅用到那些称之为“支持向量”的样本，此为后话），使得即使样本维数很高，也不会给存储和计算带来大麻烦（相对照而言，kNN算法在分类时就要用到所有样本，样本数巨大，每个样本维数再一高，这日子就没法过了.....）。

线性分类器(一定意义上,也可以叫做感知机) 是最简单也很有效的分类器形式.在一个线性分类器中,可以看到SVM形成的思路,并接触很多SVM的核心概念.

PART 1: 线性分类器

1. 线性可分

- C1和C2是要区分的两个类别，在二维平面中它们的样本如上图所示。中间的直线就是一个分类函数，它可以将两类样本完全分开。
- **一般的，如果一个线性函数能够将样本完全正确的分开，就称这些数据是线性可分的，否则称为非线性可分的。**
- **中间的分类函数不是唯一的。**



2. 线性函数和超平面

- 线性函数：在一维空间里就是一个点，在二维空间里就是一条直线，三维空间里就是一个平面，可以如此想象下去，如果不关注空间的维数，这种线性函数还有一个统一的名称——**超平面（Hyper Plane）**！
- g(x)=wx+b：X为样本向量
  - 实值函数
  - 阈值
  - 离散型的输出结果

分类间隔：分类器的评价指标

- 文本分类的例子
  - 训练样本Di=(xi,yi)：
    - xi就是文本向量（维数很高），yi就是分类标记。
    - 在二元的线性分类中，这个表示分类的标记只有两个值，1和-1（用来表示属于还是不属于这个类）。
  - 样本点到某个超平面的间隔：
    - δi=yi(wxi+b) = |wxi+b|=|g(xi)|

$$\delta_i = \frac{1}{\|w\|}|g(x_i)|$$

几何间隔：解析几何中点xi到超平面g(xi)=0的欧式距离公式

几何间隔：样本点到超平面的距离

- 一个点的集合（就是一组样本）到某个超平面的距离为此集合中离超平面最近的点的距离。

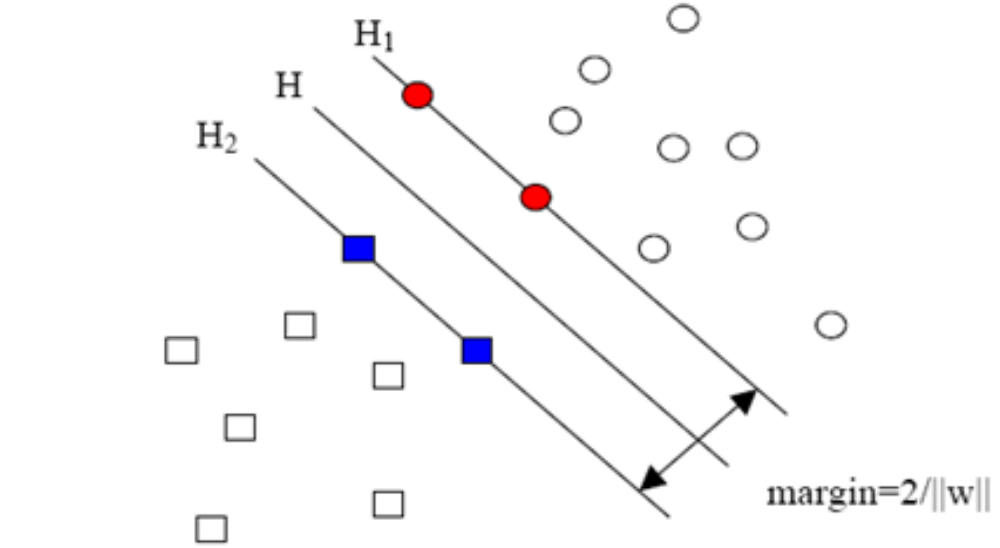


图2　线性可分情况下的最优分类线

误分次数与几何间隔

- 误分次数：代表分类器的误差。误分次数的上界由几何间隔决定！
- 几何间隔δ越大的解，它的误差上界越小。
- **因此最大化几何间隔成了我们训练阶段的目标。**

$$\text{误分次数} \le \left(\frac{2R}{\delta}\right)^2$$

  - δ是样本集合到分类面的间隔；
  - R=max ||xi||　i=1,...,n，即R是所有样本中（xi是以向量表示的第i个样本）向量长度最长的值（也就是说代表样本的分布有多么广）

最大化集合间隔：寻找最小的||w||

$$\delta_{几何} = \frac{1}{\|w\|}|g(x)|$$

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

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

$$\text{subject to} \quad y_i[(wx_i)+b]-1 \geq 0 \quad (i=1,2,\ldots,l)$$ (l是样本数)

基于可行域的寻最优解

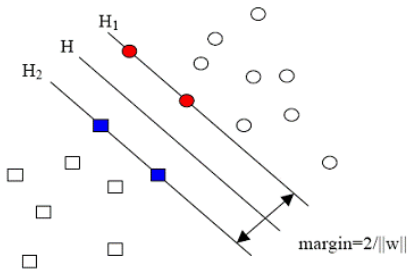


图2 线性可分情况下的最优分类线

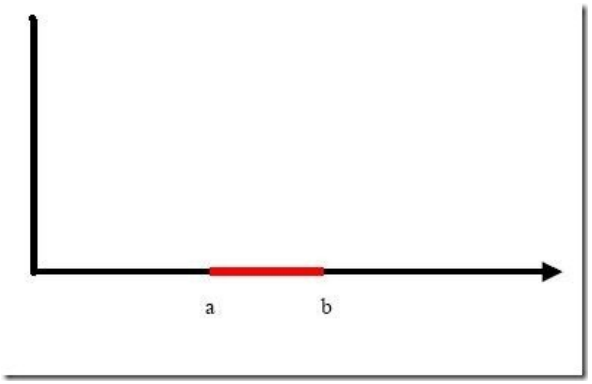
支持向量

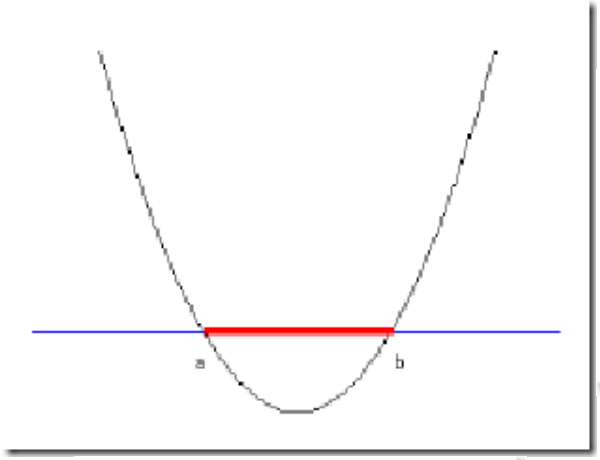
- 支持向量：可行区域上的边界点
- 凸集：一个点的集合，其中任取两个点连一条直线，这条线上的点仍然在这个集合内部。
- 二次规划：（能够找到最优解的规划）
  - 自变量就是w，而目标函数是w的二次函数，所有的约束条件都是w的线性函数因此这是一个二次规划的寻优。（Quadratic Programming，QP），
  - 由于它的可行域是一个凸集，因此它是一个凸二次规划。

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

$$\text{subject to} \quad y_i[(wx_i)+b]-1 \geq 0 \quad (i=1,2,\ldots,l)$$ (l是样本数)

核函数：升维后的线性可分





$$g(x) = c_0 + c_1 x + c_2 x^2$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}$$

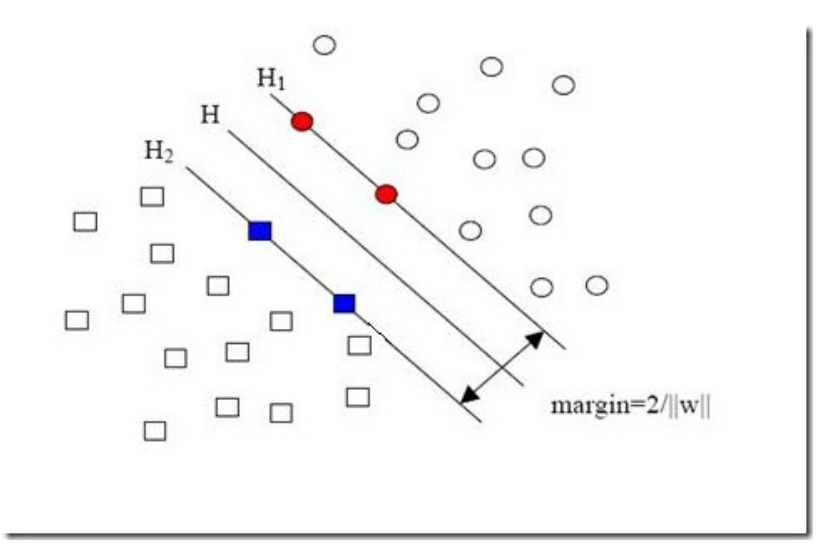
f(y)=<a,y>

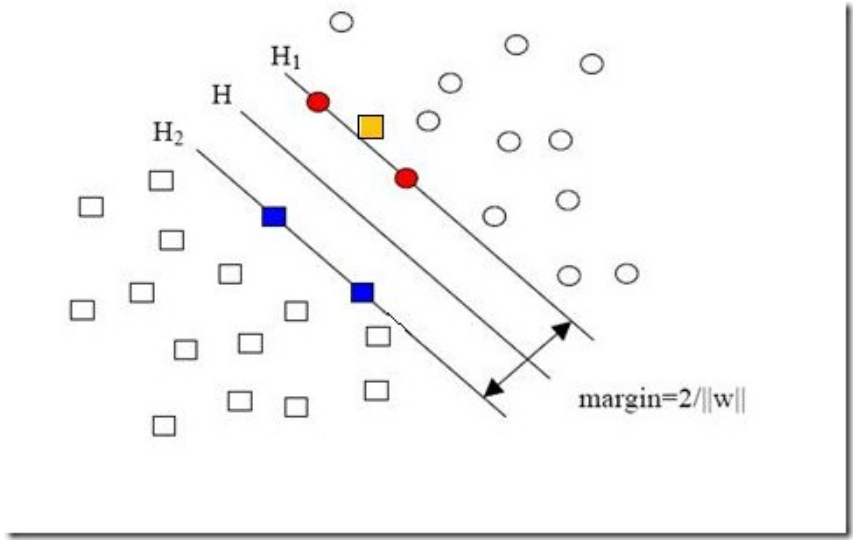
g(x)=f(y)=ay

- 解决线性不可分问题的基本思路——
  - 向高维空间转化，使其变得线性可分。
  - 找到x映射到y的方法

g(x)=K(w,x)+b （低纬空间下将x输入到K核函数中，得到与f(x')一样的值
f(x′)=<w′,x′>+b

松弛变量-计算机的容噪方法





$y_i[(wx_i)+b]\geq 1$ (i=1,2, . . . ,l) (l是样本数)

$y_i[(wx_i)+b]\geq 1-\zeta_i$ (i=1,2, . . . ,l) (l是样本数)

$\zeta_i \geq 0$

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i$$

subject to $y_i[(wx_i)+b]\geq 1-\zeta_i$ (i=1,2, . . . ,l) (l是样本数) (式1)

$\zeta_i \geq 0$

SVM的后续

- 如何选则核函数？
- 如何让二值的SVM完成多类分类任务？
- 如何使用SVM