常用分类算法及原理

一． 基本概念

找出数据中一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到某个给定的类别。它可以应用到客户的分类、客户的属性和特征分析、客户满意度分析、客户的购买趋势预测等

二． 常用分类算法

Bayes（贝叶斯）

贝叶斯分类法是基于贝叶斯定定理的统计学分类方法。它通过预测一个给定的元组属于一个特定类的概率，来进行分类。朴素贝叶斯分类法假定一个属性值在给定类的影响独立于其他属性的 —— 类条件独立性

贝叶斯（Bayes）分类算法是一类利用概率统计知识进行分类的算法，如朴素贝叶斯（Naive Bayes）算法。这些算法主要利用 Bayes 定理来预测一个未知类别的样本属于各个类别的可能性，选择其中可能性最大的一个类别作为该样本的最终类别。由于贝叶斯定理的成立本身需要一个很强的条件独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。为此就出现了许多降低独立性假设的贝叶斯分类算法，如 TAN（Tree Augmented Na?ve Bayes)算法，它是在贝叶斯网络结构的基础上增加属性对之间的关联来实现的。

通常，事件 A 在事件 B 的条件下的概率，与事件 B 在事件 A 的条件下的概率是不一样的；然而，这两者是有确定的关系，贝叶斯定理就是这种关系的陈述。
贝叶斯定理是指概率统计中的应用所观察到的现象对有关概率分布的主观判断（即先验概率）进行修正的标准方法。当分析样本大到接近总体数时，样本中事件发生的概率将接近于总体中事件发生的概率。

作为一个规范的原理，贝叶斯法则对于所有概率的解释是有效的；然而，频率主义者和贝叶斯主义者对于在应用中概率如何被赋值有着不同的看法：频率主义者根据随机事件发生的频率，或者总体样本里面的个数来赋值概率；贝叶斯主义者要根据未知的命题来赋值概

率。

贝叶斯统计中的两个基本概念是先验分布和后验分布：
先验分布。总体分布参数 θ 的一个概率分布。贝叶斯学派的根本观点，是认为在关于总体分布参数 θ 的任何统计推断问题中，除了使用样本所提供的信息外，还必须规定一个先验分布，它是在进行统计推断时不可缺少的一个要素。他们认为先验分布不必有客观的依据，可以部分地或完全地基于主观信念。后验分布。根据样本分布和未知参数的先验分布，用概率论中求条件概率分布的方法，求出的在样本已知下，未知参数的条件分布。因为这个分布是在抽样以后才得到的，故称为后验分布。贝叶斯推断方法的关键是任何推断都必须且只须根据后验分布，而不能再涉及样本分布。
贝叶斯公式为
$P(A\cap B)=P(A)*P(B|A)=P(B)*P(A|B)$
$P(A|B)=P(B|A)*P(A)/P(B)$
其中：
1、P(A)是 A 的先验概率或边缘概率，称作"先验"是因为它不考虑 B 因素。
2、P(A|B)是已知 B 发生后 A 的条件概率，也称作 A 的后验概率。
3、P(B|A)是已知 A 发生后 B 的条件概率，也称作 B 的后验概率，这里称作似然度。
4、P(B)是 B 的先验概率或边缘概率，这里称作标准化常量。
5、P(B|A)/P(B)称作标准似然度。
贝叶斯法则又可表述为：
后验概率=(似然度*先验概率)/标准化常量=标准似然度*先验概率
P(A|B)随着 P(A)和 P(B|A)的增长而增长，随着 P(B)的增长而减少，即如果 B 独立于 A 时被观察到的可能性越大，那么 B 对 A 的支持度越小。

贝叶斯公式为利用搜集到的信息对原有判断进行修正提供了有效手段。在采样之前，经济主体对各种假设有一个判断（先验概率），关于先验概率的分布，通常可根据经济主体的经验判断确定(当无任何信息时，一般假设各先验概率相同)，较复杂精确的可利用包括最大熵技术或边际分布密度以及相互信息原理等方法来确定先验概率分布

SVM（支持向量机）

支持向量机把分类问题转化为寻找分类平面的问题，并通过最大化分类边界点距离分类平面的距离来实现分类。

支持向量机的优缺点
优点

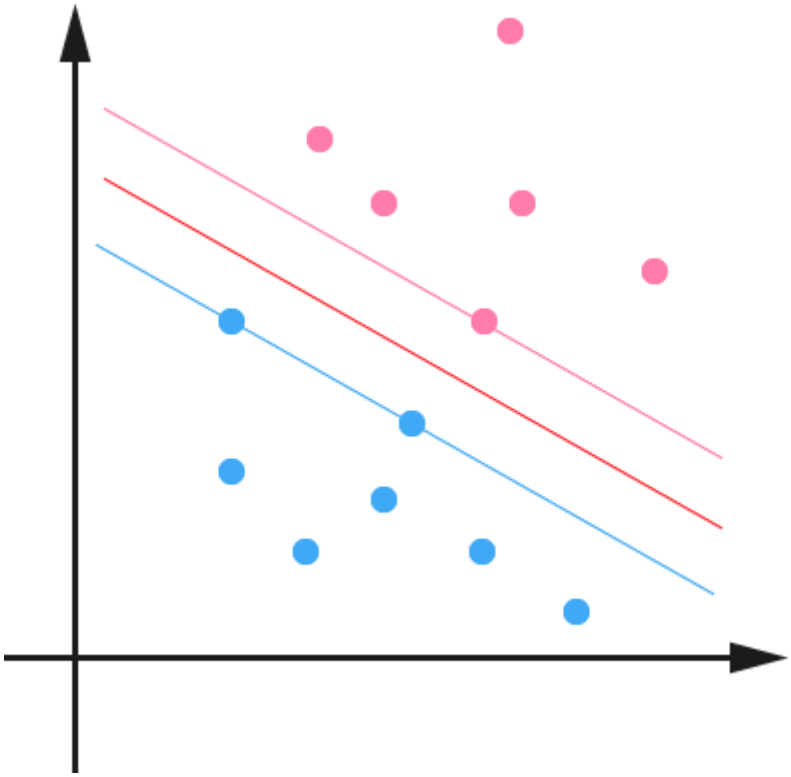
可以解决小样本下机器学习的问题。
提高泛化性能。
可以解决高维、非线性问题。超高维文本分类仍受欢迎。
避免神经网络结构选择和局部极小的问题。

缺点

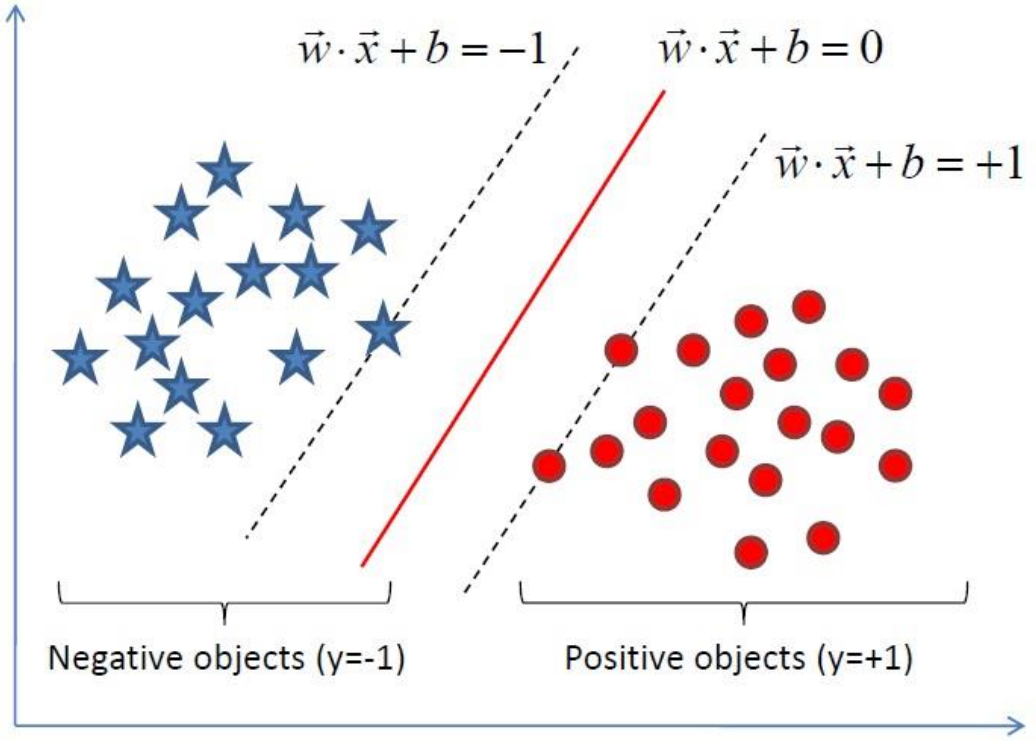
缺失数据敏感。
内存消耗大，难以解释。
运行和调差略烦人。

支持向量机的公式

转自研究者 July: SVM 的求解，先导出 12||w||2，继而引入拉格朗日函数，转化为单一因子对偶变量 a 的求解。如此求 w.b 与 a 的等价，而求 a 的解法即为 SMO。把求分类函数 f(x)=ω*x+b 的问题转化为求 w,b 的最优化问题，即凸二次规划问题，妙。



从上图我们可以看出，这条红色的线（超平面）把红色的点和蓝色的点分开了。超平面一边的点对应的 y 全部是-1，而另外一边全部是 1。
接着我们可以令分类函数：f(x)=ωTx+b。显然 x 是超平面上的点时，f(x)=0。那么我们不妨要求所有满足 f(x)<0 的点，其对应的 y 等于-1，而 f(x)>0 则对应的 y=1 的数据点。（我盗用了很多图。。。）



回忆之前的目标函数:
max1||ω||s.t.,yi(ωT+b)⩾1,i=1,...,n
这个问题等价于
max1||ω||2s.t.,yi(ωT+b)⩾1,i=1,...,n

很显然这是一个凸优化的问题，更具体的，它是一个二次优化问题—目标函数是二次的，约束条件是线性的。这个问题可以用任何现成的 QP（Quadratic Programming）优化包解决。但是因为这个问题的特殊性，我们还可以通过 Lagrange Duality 变换到对偶变量的优化问题，找到一种更加行之有效的方法求解。首先我们给每一个约束条件加上一个

Lagrange mutiplier,我们可以将它们融合到目标函数中去。 L(ω,b,a)=12||ω||2−∑

ni=1α(yi(wTxi+b)−1)，然后我们令 θ(ω)=maxai⩾0L(ω,b,a)容易验证，当某个约束条件不满

足时，例如 yi(wTxi+b)<1，那么我们显然有 θ(w)=∞。而当所有约束条件都满足时，则有 θ(ω)=12||ω||2，亦即我们最初要最小化的量。那么我们现在的目标函数就变成了:
minw,bθ(ω)=minω,bmaxai⩾0L(ω,b,α)=p*，并且我们有 d*⩽p*，因为最大值中最小的一个一定要大于最小值中最大的一个。总之 p*提供了一个第一个问题的最优值 p*的一个下界。在满足 KKT 条件时，二者相等，我们可以通过求解第二个问题来求解第一个问题。
先让 L 关于 ω 和 b 最小化，我们分别把 L 对 w 和 b 求偏导:

∂L∂ω=0⟹ω=∑i=1nαiyixi

∂L∂b=0⟹∑i=1nαiyi=0

再带回 L 得到：

$$L(\omega,b,a)=12\sum i,j=1n\alpha i\alpha jyiyjxTixj-\sum i,j=1n\alpha i\alpha jyiyjxTixj-b\sum i=1n\alpha iyi+\sum i=1n\alpha i=\sum i=1n\alpha i-12\sum i,j=1n\alpha i\alpha jyiyjxTixj$$

此时我们得到关于 dual variable α 的优化问题：

$$\max\alpha\sum ni=1\alpha i-12\sum ni,j=1\alpha i\alpha jyiyjxTixjs.t.,\alpha i\geqslant 0,i=1,...,n\sum ni=1\alpha iyi=0$$

这个问题存在高效的算法，不过求解过程就不在这里介绍了。对于一个数据点进行分类时，我们是把 x 带入到 f(x)=wTx+b 中，然后根据其正负号来进行类别划分的。把 $\omega=\sum ni=1\alpha iyixi$ 代入到 f(x)=wTx+b，我们就可以得到 $f(x)=\sum ni=1\alpha iyi<xi,x>+b$，这里的形式的有趣之处在于，对于新点 x 的检测，只需要计算它与训练数据点的内积即可。

为什么非支持向量的 α 等于零呢？因为对于非支持向量来说，$L(\omega,b,a)=12||\omega||2-\sum ni=1\alpha(yi(wTxi+b)-1)$中的，(yi(wTxi+b)−1)是大于 0 的，而且 αi 又是非负的，为了满足最大化，αi 必须等于 0