**基于WEKA的数据挖掘算法学习（四）**

# 支持向量机的理论知识

基于训练集D在样本空间找到一个划分超平面，将不同的类别的样本分开。但这种划分超平面可能很多，我们希望产生的分类结果是最鲁棒的，对未见示例的泛化能力最强。

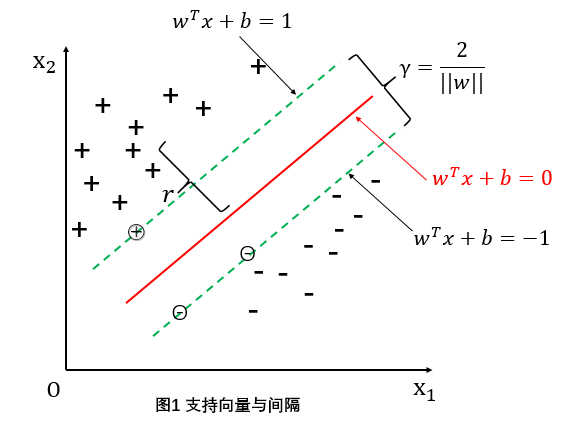
## 1.1间隔与支持向量

在样本空间中，划分超平面通过如下线性方程来描述：

其中为法向量，决定方向；b为位移项，决定超平面与原点之间的距离。

假设超平面（w,b）能将训练集样本正确分类，令

当距离超平面最近的训练样本点使上式等号成立，它们就被称为“支持向量”，两个异类支持向量到超平面的距离之和为，称为间隔，



如图一支持向量与间隔可以清晰的看出超平面中的关系。

为了找到“最大间隔”的划分超平面，需要找出约束参数w与b，使得γ最大。最大化间隔仅需最大化，等价于最小化，于是得到支持向量机的基本型：

## 1.2对偶问题

### 1.2.1拉格朗日乘子法

超平面模型是一个凸二次规划问题，使用高效的拉格朗日乘子法得其“对偶问题”，该问题的拉格朗日函数可写为，

令对w，b的偏导为零带入消去w和b可解得对偶问题。

### SMO算法

基本思路：固定之外的所有参数，然后求上的极值，由于固定之外的其他变量，则可由其他变量导出。因此，SMO 每次选择两个变量和，并固定其他参数，然后求解对偶问题更新后的和。

SMO算法之所以高效，在于固定其他参数后，仅优化两个参数的过程能做到非常高效。

## 1.3核函数

核函数的问题的引入在于，现实任务中，原始空间内也许并不能存在一个能正确划分两类样本的超平面，对于这类问题，可以从原始空间中映射到更高维的特征空间，使得样本在这个特征空间内线性可分。

求解后可得到，

其中的就是核函数，特征空间的好坏对支持向量机的性能至关重要，而核函数隐式的定义了这个特征空间，因此“核函数的选择”成为了支持向量机最大变数。

# 贝叶斯分类器的理论知识

## 2.1贝叶斯决策论

贝叶斯决策论是概率论框架下实施决策的基本方法，考虑基于概率和误判损失来选择最优的类别标记。最小化的分类错误率的贝叶斯最优分类器为，

,c∈y

即每个样本x，选择能使后验概率最大的类别标记。因此，首先要获得后验概率。

首先考虑条件分布有，

再基于贝叶斯定理，可写为，

于是估计的问题就转化为如何基于训练数据D来估计先验和似然。

## 2.2朴素贝叶斯分类器

估计后验概率的主要困难在于：类条件概率是所有属性上的联合概率，难以从有限的训练样本直接估计而得。于是，引入朴素贝叶斯分类器采用“属性条件独立性假设”。由于属性条件独立性假设，可得朴素贝叶斯分类器的表达式，

针对离散型属性，直接使用占训练集的比例来估计，而对连续属性考虑概率密度函数，比如正态分布，可以找第c类样本在第i个属性上取值的均值和方差，从而根据概率密度函数求得概率。

需要注意的地方：可能出现某个属性值在训练集中未与某个类同时出现的情况，引入拉普拉斯修正可以避免因训练集样本不充分的而导致概率估值为0的问题。

## 2.3半朴素贝叶斯分类器

朴素贝叶斯采用了属性条件独立性假设，简化了求解，但在实际任务中很难实现，于是引入了半朴素贝叶斯分离器。基本思路是适当考虑一部分属性间的相互依赖信息，不需进行完全联合概率密度，也不至于彻底忽略了比较强的属性依赖关系。

1、SPODE：假设所有属性都依赖于同一属性，通过交叉验证等模型选择方法来确定超父类属性。

2、TAN：最大权生成树，保留了强关联属性之间的依赖性。

3、AODE：AODE尝试将每个属性作为超父类构建SPODE，然后将具有足够训练数据支撑的SPODE集成起来作为最终结果。

## 2.4贝叶斯网络

也称为“信念网”，借助有向无环图来刻画属性之间的依赖关系，使用条件概率表来描述属性的联合概率分布。个人理解为在某个有向无环图中，在父节点条件下，求某个属性的概率值。变量间也有不同的典型依赖关系，贝叶斯网络结构有效表达了属性间的条件独立性。

# 3.weka算法实现

## 3.1数据集的准备

1、UCI完备数据集：segment.csv

2、UCI不完备数据集：hepatitis.csv

3、UCI不平衡数据集： abalone.csv

## 3.2 SMO

选择五折交叉数据验证，不断调整Kernel（核函数种类）参数进行数据测试，结果如下：

当Kernel= the polynomial kernel（多项式核函数）时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.890 | 0.890 | 0.890 | 0.963 | 89.07% | 0.66s |
| hepatitis | 0.674 | 0.671 | 0.662 | 0.656 | 67.096% | 0.03s |
| abalone | 0.218 | 0.249 | 0.194 | 0.720 | 24.922% | 7.64s |

当Kernel= The RBF kernel（[RBF径向基函数]高斯核函数）时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.721 | 0.638 | 0.632 | 0.897 | 63.80 % | 0.63s |
| hepatitis | 0.677 | 0.606 | 0.533 | 0.569 | 60.64% | 0.05s |
| abalone | 0.092 | 0.202 | 0.124 | 0.646 | 20.158% | 41.52s |

当Kernel= the universal kernel.（通用内核函数）时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.888 | 0.886 | 0.886 | 0.957 | 88.57 % | 0.48s |
| hepatitis | 0.705 | 0.703 | 0.698 | 0.692 | 70.322% | 0.04s |
| abalone | 0.253 | 0.273 | 0.238 | 0.725 | 27.268% | 49.1s |

从跑出来的数据结果来看，abalone.csv的效果最差，开始以为是数据处理有问题，后来发现处理结果正确，确实是有29个类，且由于样本数据集有4000+，建模时间花费相当大，然后再一折一折的验证，花费时间很大。开始使用了magic04.csv，将近两万的数据量，跑一个数据集差不多要跑两三个小时，就换成了segment.csv。此外，从结果可以知道，选择训练集合适的核函数对于支持向量机至关重要，“核函数的选择”成为了支持向量机最大变数。

## 3.3NaiveBayes

选择五折交叉数据验证，进行数据测试，结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.765 | 0.781 | 0.765 | 0.946 | 78.095% | 0.04s |
| hepatitis | 0.712 | 0.710 | 0.704 | 0.757 | 70.96% | 0s |
| abalone | 0.201 | 0.238 | 0.207 | 0.710 | 23.82% | 0.09s |

从实验结果来看，完备数据集segment.csv效果最佳，abalone.csv仍然不好。与SMO数据结果对比，时间效率整体优越程度相当大，基本上立马可以跑出结果。

## 3.4 BayesNet

选择五折交叉数据验证，保持searchAlgorithm=K2，调整Estimator参数进行数据测试，结果如下：

当Estimator= BMAE Estimator时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.838 | 0.833 | 0.835 | 0.968 | 83.33% | 0.05s |
| hepatitis | 0.638 | 0.619 | 0.581 | 0.633 | 61.93% | 0.04s |
| abalone | 0.181 | 0.240 | 0.197 | 0.711 | 23.98% | 0.23s |

当Estimator= simpleEstimator时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.828 | 0.829 | 0.826 | 0.967 | 82.85% | 0.01s |
| hepatitis | 0.623 | 0.613 | 0.580 | 0.632 | 61.29% | 0.01s |
| abalone | 0.181 | 0.240 | 0.197 | 0.711 | 23.98% | 0.07s |

保持searchAlgorithm=hillClimber，调整Estimator参数进行数据测试，结果如下：

当Estimator= BMAE Estimator时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.838 | 0.833 | 0.835 | 0.968 | 83.33% | 0.02s |
| hepatitis | 0.638 | 0.619 | 0.581 | 0.633 | 61.93% | 0.02s |
| abalone | 0.181 | 0.240 | 0.197 | 0.711 | 23.98% | 0.22s |

当Estimator= simpleEstimator时：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 数据集 | Precision | Recall值 | F值 | AUC值 | Correctly | Time |
| segment | 0.828 | 0.829 | 0.826 | 0.967 | 82.85% | 0.03s |
| hepatitis | 0.623 | 0.613 | 0.580 | 0.632 | 61.29% | 0.01s |
| abalone | 0.181 | 0.240 | 0.197 | 0.711 | 23.98% | 0.07s |

可以看出，更换估计策略（Estimator）对跑出来数据结果稍有影响，而searchAlgorithm对结果的影响甚至可以忽略不计。

# 4.本周学习心得体会

本周学习了《机器学习 周志华版》中的支持向量机和贝叶斯决策的理论知识。支持向量机中推导难度较大，建立完支持向量机基本型后，主要使用拉格朗日乘子法来求解了对偶问题；贝叶斯决策论主要解决基于训练数据集来估计先验概率问题，通过属性独立得到朴素贝叶斯，通过部分属性依赖得到半朴素贝叶斯。总之就是逐渐往符合实际训练数据集的算法做调整。

实验数据测试主要感觉就是SMO花费时间相当大。然后跑数据时调参有时可以对实验结果有很大改进（如SMO核函数的选择），有时却没什么改变（如贝叶斯中估计决策estimitor的选择）。