算法介绍文档

版本 1.0

目录

[1 随机森林 5](#_Toc461186531)

[1.1 算法简介 5](#_Toc461186532)

[1.2 参数说明 5](#_Toc461186533)

[1.3 调参方法 6](#_Toc461186534)

[2 Lasso 6](#_Toc461186535)

[2.1 算法简介 6](#_Toc461186536)

[2.2 参数说明 7](#_Toc461186537)

[2.3 调参方法 7](#_Toc461186538)

[3 SVM 8](#_Toc461186539)

[3.1 算法简介 8](#_Toc461186540)

[3.2 参数说明 8](#_Toc461186541)

[3.3 调参方法 9](#_Toc461186542)

[4 NaiveBayes 10](#_Toc461186543)

[4.1 算法简介 10](#_Toc461186544)

[4.2 参数说明 10](#_Toc461186545)

[4.3 调参方法 11](#_Toc461186546)

[5 CART 11](#_Toc461186547)

[5.1 算法简介 11](#_Toc461186548)

[5.2 参数说明 12](#_Toc461186549)

[5.3 调参方法 12](#_Toc461186550)

[6 LogisticRegression 13](#_Toc461186551)

[6.1 算法简介 13](#_Toc461186552)

[6.2 参数说明 13](#_Toc461186553)

[7 AdaBoost 13](#_Toc461186554)

[7.1 算法简介 14](#_Toc461186555)

[7.2 参数说明 14](#_Toc461186556)

[8 Stacking 14](#_Toc461186557)

[8.1 算法简介 14](#_Toc461186558)

[8.2 参数说明 14](#_Toc461186559)

[9 神经网络 15](#_Toc461186560)

[9.1 算法简介 15](#_Toc461186561)

[9.2 参数说明 15](#_Toc461186562)

# 随机森林

## 算法简介

随机森林是一种组合分类器，其中的每个分类器都是一颗决策树。个体决策树在每个结点使用随机选择的属性决定划分。更准确的说，每一棵树都依赖于独立抽样，并与森林中所有树具有相同分布的随机向量的值。分类时，每棵树都投票并且返回得票最多的类。

决策树是一种基本的分类和回归方法。其内部结点（非叶结点）表示在一个属性上的测试，每个分支表示该测试的一个输出，而每个叶结点存放一个类标号。常见的属性选择度量有：

* 信息增益







* 增益率





* 基尼指数





## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -I | 组成森林的树的棵树 | 一般为20-500 |
| -K | 构建单棵树随机选取的特征数目 | 一般为原特征数基于2的对数+1 |
| -depth | 单棵树的最大深度 | 如果为0，则不限制深度 |
| -num-slots | 并行时使用的CPU核数 | 如果为1，则不使用并行加速策略 |
| -S | 随机数种子 |  |

## 调参方法

默认参数：-I/50/-K/8/-S/12/-depth/0/-num-slots/4

I：影响模型效果与性能，随着该参数的增大，黑样本的Recall与Precision都会有一定程度的提升，同时模型训练速度下降（在num-slots参数一致时）。

K：影响模型效果，随着该参数的增大，黑样本的Recall与Precision都会有较为明显的提升。

depth：影响模型效果，随着该参数的增加，黑样本的Recall与Precision会有显著增加，之后可能会出现过拟合现象，随着参数的继续增加，黑样本的Recall与Precision反而下降。

num-slots：影响模型性能，随着该参数的增大，模型训练速度加快。

S：影响模型效果，但影响较为有限。

SampleRate：影响模型效果与性能，随着该参数的增大，黑样本的Recall将下降，但Precision将会上升，变化幅度较为显著，同时模型训练速度将会下降。

# Lasso

## 算法简介

Lasso由Tibshirani于1996年提出，旨在构造一个惩罚函数获得一个精炼的模型，最终通过确定一些指标的系数为0，实现指标集合精简的目的是一种处理具有复共线性数据的有偏估计。其基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而能够产生某些严格等于0的回归系数，得到解释力较强的模型。Lasso算法可以用如下最优化问题来表述：

在限制的情况下，求使得残差平方和达到最小的回归系数的估值。

当将Lasso算法与逻辑回归结合在一起，相关问题便转化为求解如下无约束最优问题：

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -S | SVM的类型 | 0--L2-regularized logistic regression (primal)  1--L2-regularized L2-loss support vector classification (dual)  2--L2-regularized L2-loss support vector classification (primal)  3--L2-regularized L1-loss support vector classification (dual)  4--support vector classification by Crammer and Singer  5--L1-regularized L2-loss support vector classification  6--L1-regularized logistic regression  7--L2-regularized logistic regression (dual) |
| -Z | 开启数据的标准化 | 如果数据集没有做标准化，则必须开启 |
| -E | 终止迭代最小误差 | 一般为0.01 |
| -C | 损失函数惩罚系数 | 可选范围从0.0001到10000，一般取10-100 |
| -W | 第i个类的权重 | 一般固定一个类为1，其余类取相应的比值 |

## 调参方法

默认参数：-S/6/-E/0.1/-C/10/-W/1 15/-Z

E：影响模型效果与性能，随着该参数的减小，黑样本的Recall与Precision都会上升，之后可能会出现过拟合现象，随着参数的继续减小，黑样本的Recall与Precision反而下降，同时模型训练速度也将会下降。

C：影响模型效果与性能，随着该参数的增大，黑样本的Recall将上升，Precision将下降，变化幅度较小。

W：影响模型效果，随着该参数（比值）的增大，黑样本的Recall将上升，但Precision将会下降，变化幅度较为显著。

SampleRate：影响模型效果与性能，随着该参数的增大，黑样本的Recall将下降，但Precision将会上升，变化幅度较为显著，同时模型训练速度将会下降。

# SVM

## 算法简介

支持向量机，因其英文名为support vector machine，故一般简称SVM，是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

SVM的主要思想可以概括为以下两点：

* 它是针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转化为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能。
* 它基于结构风险最小化理论之上在特征空间中建构最优分割超平面，使得学习器得到全局最优化，并且在整个样本空间的期望风险以某个概率满足一定上界。

SVM算法可以用如下最优化问题来表述：

在的约束条件下，求解，等价于求解。

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -S | SVM的类型 | 0 = C-SVC  1 = nu-SVC  2 = one-class SVM  3 = epsilon-SVR  4 = nu-SVR |
| -K | 核函数类型 | 0 = linear: u'\*v  1 = polynomial: (gamma\*u'\*v + coef0)^degree  2 = radial basis function: exp(-gamma\*|u-v|^2)  3 = sigmoid: tanh(gamma\*u'\*v + coef0) |
| -D | 多项式核最高次项的次数 | 使用此参数前提是 – K参数为1，一般选择1-11 |
| -G | 核函数的gamma参数 | 使用此参数前提是-K=2,对于高斯核函数（-K/2），通常可选择下面几个数的倒数：0.1，0.3，0.8，1.6，3.2，6.4 ，12.8 |
| -R | 多项式核和sigmoid核的coef0  参数 | 一般选0.2，0.4，0.6，0.8，1 |
| -C | C-SVC, epsilon-SVR, nu-SVR中损失函数惩罚系数 | 可选范围从0.0001到10000，一般取10-100 |
| -N | nu-SVC, one-class SVM, nu-SVR中 | 使用one-class SVM时指定此参数 |
| -Z | 开启数据的标准化 | 除非原始数据已经进行过标准化，否则必须开启。 |
| -J | 关闭数据的nominal to binary转化 | 若输入数据没有名义型变量可关闭此功能，否则必须开启。 |
| -V | 关闭数据的缺失值处理 | 如果数据没有缺失值可以关闭此功能。 |
| -M | 设置cache大小，单位是MB | 最高不能超过机器允许的内存大小 |
| -E | 设置终止迭的最小误差 | 一般为0.01 |
| -H | 关闭shrinking | 若训练太慢，可以尝试关闭 |
| -W | 第i个类的权重 | 一般固定一个类为1，其余类取相应的比值 |
| -B | 输出每个类的概率估计 |  |
| -seed | 随机数种子 |  |

## 调参方法

默认参数：-K/2/-G/1/-C/20/-W/1 25/-M/500/-Z

C：影响模型效果与性能，随着该参数的增大，黑样本的Recall将上升，Precision将下降，变化幅度较小。

M：影响模型性能，随着该参数的增大，训练速度会提升。

E：影响模型效果与性能，随着该参数的减小，黑样本的Recall与Precision都会上升，之后可能会出现过拟合现象，随着参数的继续减小，黑样本的Recall与Precision反而下降，同时模型训练速度也将会下降。

H：影响模型性能，开启时可以提高训练速度。

W：影响模型效果，随着该参数（比值）的增大，黑样本的Recall将上升，但Precision将会下降，变化幅度较为显著。

seed：影响模型效果，但影响较为有限。

SampleRate：影响模型效果与性能，随着该参数的增大，黑样本的Recall将下降，但Precision将会上升，变化幅度较为显著，同时模型训练速度将会下降。

# NaiveBayes

## 算法简介

贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

朴素贝叶斯发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率，其假定属性之间相互独立，因此理论上具有最小的误差率，但在实际应用中并非如此。同时，其所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。

根据贝叶斯定理，对一个分类问题，给定样本特征x，样本属于类别y的概率为：

假设特征向量x有M个维度，依据朴素假设，进行全概率公式展开，则有：

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -K | 对于连续型变量，使用核密度估计代替高斯函数估计 |  |
| -D | 对于连续型变量进行有监督离散化 | 如果原始数据存在连续型变量，最好开启离散化 |

## 调参方法

默认参数：-K

SampleRate：影响模型效果与性能，随着该参数的增大，黑样本的Recall将下降，但Precision将会上升，变化幅度较为显著，同时模型训练速度将会下降。

# CART

## 算法简介

决策树是一种常用的机器学习算法模型，它可以用于分类、回归等多种场景。在分类问题中，表示基于属性对实例进行分类的过程，可以简单地认为是if-then规则的集合，也可以认为是在属性空间与类空间上的条件概率分布。其根据特定策略（属性选择度量）将样本数据划分为不同的数据子集，进而形成特定的树结构。分类过程即依据上述策略，自根结点开始选择合适分支递归向下搜索，直至叶结点，从而确定实例所属类别。常见的三种属性选择度量为：

* 信息增益







* 增益率





* 基尼指数





这里使用的决策树算法为基于增益率的C4.5算法。

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -U | 使用不剪枝树 | 使用不剪枝的树，使用该参数后树将任意生长，可能导致过拟合。如果作为弱模型使用，不推荐使用此参数，如果作为强模型使用，要防止过拟合。 |
| -C | 设置剪枝阈值 |  |
| -M | 每个叶子上的数据条数 |  |
| -R | REP剪枝策略 | 使用reduced error pruning剪枝策略，即在满足剪枝条件时，先执行剪枝操作，当剪枝确实降低了错误率，才真正执行剪枝，否则不剪枝，推荐使用 |
| -N | 设置REP剪枝策略时的fold数目 | 当制定了-R参数时使用，一般取3-10 |
| -B | 树的每个节点只使用二分裂 | 推荐使用 |
| -S | 关闭子树提升 | 不推荐关闭。 |
| -A | 使用拉普拉斯平滑 | 推荐使用 |
| -J | 关闭连续属性离散化 | 如果数据中存在连续值，必须使用。 |
| -Q | 随机数种子 |  |
| -depth | 树深度 |  |

## 调参方法

默认参数：-M/5/-R/-B/-L

C：影响模型效果，随着该参数的增加，黑样本的Recall将会上升，但Precision却会下降，变化幅度较为显著。

M：影响模型效果，随着该参数的增加，黑样本的Recall将会下降，但Precision却会上升。

N：影响模型效果与性能，随着该参数的增加，黑样本的Recall与Precision将上升，变化幅度较为显著，同时模型训练速度将会下降。

Q：影响模型效果，但影响较为有限。

depth：影响模型效果，随着该参数的增加，黑样本的Recall与Precision会有显著增加，之后可能会出现过拟合现象，随着参数的继续增加，黑样本的Recall与Precision反而下降。

SampleRate：影响模型效果与性能，随着该参数的增大，黑样本的Recall将下降，但Precision将会上升，变化幅度较为显著，同时模型训练速度将会下降。

# LogisticRegression

## 算法简介

逻辑回归（Logistic Regression）是机器学习中的一种分类模型，主要用于二分类问题，所以使用了sigmoid函数：

其所做的假设为：

基于最大似然估计，有：

求解：

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -R | L2正则化参数 |  |
| -M | 最大迭代次数 |  |

# AdaBoost

## 算法简介

Adaboost是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。使用adaboost分类器可以排除一些不必要的训练数据特征，并放在关键的训练数据上面。

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -P | 前级训练使用的数据比例 | 一般指定90 |
| -S | 随机数种子 |  |
| -I | 迭代次数 |  |
| -W | 要提升的模型的完整名字 | 这里的模型一定要是弱分类模型，也就是logist/naivebayes/CART/SVM等，不包括randomforest |
| -- | 要提升模型的参数 | 基础模型的参数附在--/后面，格式与之前基本模型参数设置方法一致 |

# Stacking

## 算法简介

Stacking讲的是非线性组合学习器的概念，其具体的过程如下：

1.划分训练数据集为两个不相交的集合。

2. 在第一个集合上训练多个学习器。

3. 在第二个集合上测试这几个学习器

4. 把第三步得到的预测结果作为输入，把正确的回应作为输出，训练下一层的学习器。

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -X | 交叉验证的fold数 | 一般范围是3-10 |
| -S | 随机数种子 |  |
| -M | 第二级模型的完整名称 |  |

# 神经网络

## 算法简介

神经网络由大量的节点（或称神经元）之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。

人工神经网络具有四个基本特征：

* 非线性
* 非局限性
* 非常定性
* 非凸性

单个神经元可以把一个n维向量空间用一个超平面分割成两部分，该超平面的方程为：

那么若A为神经元的输入向量，f为其激励函数，则其输出t可表示为：

## 参数说明

|  |  |  |
| --- | --- | --- |
| 参数名称 | 参数解释 | 备注 |
| -L | 学习速率 | 在0-1之间 |
| -S | 随机数种子 |  |
| -M | 动量 |  |
| -N | 最大迭代次数 |  |
| -V | 验证数据集比例 | 在0-100之间。 |
| -E | 终止迭代次数上限 |  |
| -B | 关闭数据的nominal to binary转化 | 若输入数据没有名义型变量可关闭此功能，否则必须开启。 |
| -C | 关闭数据的标准化 | 除非原始数据已经进行过标准化，否则必须开启。 |
| -H | 隐含层结构 | 隐含层数目和每层神经元数，每层之间用#分割，数字代表该层层神经元个数。其中  'a' = (attribs + classes) / 2,  'i' = attribs,  'o' = classes,  't' = attribs+ classes) |
| -D | 开启decay |  |