SVM,AdaBoost,LR三种学习策略对比

1.SVM

应用场景

1. 文本和超文本的分类；
2. 用于图像分类；
3. 用于手写体识别；

优势

1. 分类效果好；
2. 可以有效地处理高维空间的数据；
3. 可以有效地处理变量个数大于样本个数的数据；
4. 只是使用了一部分子集来进行训练模型，所以SVM模型不需要太大的内存；
5. 可以提高泛化能力；
6. 无局部极小值问题；

缺点

1. 无法处理大规模的数据集，因为该算法需要较长的训练时间；
2. 无法有效地处理包含噪声太多的数据集；
3. SVM模型没有直接给出概率的估计值，而是利用交叉验证的方式估计，这种方式耗时较长；
4. 对缺失数据非常敏感；
5. 对于非线性问题，有时很难找到一个合适的核函数。

适用情况

1. 数据的维度较高；
2. 需要模型具有非常强的泛化能力；
3. 样本数据量较小时；
4. 解决非线性问题；

不适用情况

1. 数据集的数据量过大；
2. 数据集中的含有噪声；
3. 数据集中的缺失较多的数据；
4. 对算法的训练效率要求较高；

选用前提

1. 该项目所提供的样本数据相对较少；
2. 该问题是属于非线性问题；
3. 数据集经过“独热编码”后，维度较高；

2.AdaBoost

应用场景

1. 用于二分类或多分类问题；
2. 用于特征选择；
3. 多标签问题；
4. 回归问题；

优点

1. 精度非常高；
2. 可以与各种方法构建子分类器，AdaBoost算法提供一种计算框架；
3. 弱分类器的构造方法比较简单；
4. 算法易于理解，不用做特征筛选；
5. 不易发生过拟合。
6. 易于编码；

缺点

1. AdaBoost算法的迭代次数不好设定，需要使用交叉验证的方式来进行确定；
2. 数据集的不平衡分布导致分类器的分类精度下降；
3. 训练比较耗费时间；
4. 对异常值比较敏感；

适用情况

1. 用于解决二分类问题；
2. 解决大类单标签问题；
3. 处理多类单标签问题；
4. 处理回归相关的问题。

不适用情况

1. 数据集分布非常不均匀；
2. 数据集中含有较多的异常值；
3. 对算法的训练的效率要求较高；

选用前提

1. 该数据集可以归属为多标签分类问题；
2. 数据集中异常值较少；
3. 对算法模型的准确率要就较高；

3.LR

LR是很多分类算法的基础组件，它的好处是输出值自然地落在0到1之间，并且有概率意义。因为LR本质上是一个线性的分类器，所以处理不好特征之间相关的情况。

应用场景

1. 用于分类：适合做很多分类算法的基础组件。
2. 用于预测：预测事件发生的概率（输出）。
3. 用于分析：单一因素对某一个事件发生的影响因素分析（特征参数值）。

使用条件

1. 当数据线性可分
2. 特征空间不是很大的情况
3. 不在意新数据的情况
4. 后续会有大量新数据的情况。

优点：

1. 从整体模型来说，模型清洗，背后的概率推导经得住推敲；
2. 从输出值来说，输出值自然落在0到1之间，并且有概率意义；
3. 从模型参数来说，参数代表每个特征对输出的影响，可解释性强；
4. 从运行速度来说，实施简单，非常高效（计算量小、存储占用低），可以在大数据场景中使用；

5. 从过拟合角度来说，解决过拟合的方法很多，如L1、L2正则化；
6. 从多重共线性来说，L2正则化就可以解决多重共线性问题；

缺点：

1. （特征相关情况）因为它本质上是一个线性的分类器，所以处理不好特征之间相关的情况；
2. （特征空间）特征空间很大时，性能不好；
3. （预测精度）容易欠拟合，预测精度不高；

4.综合对比

1.SVM只考虑分类面附近的局部的点，即支持向量，LR则考虑所有的点，与分类面距离较远的点对结果也起作用，虽然作用较小。

SVM中的分类面是由支持向量控制的，非支持向量对结果不会产生任何影响。LR中的分类面则是由全部样本共同决定。线性SVM不直接依赖于数据分布，分类平面不受一类点影响；LR则受所有数据点的影响，如果数据不同类别strongly unbalance，一般需要先对数据做balancing。

2、在解决非线性分类问题时，SVM采用核函数，而LR通常不采用核函数。

分类模型的结果就是计算决策面，模型训练的过程就是决策面的计算过程。在计算决策面时，SVM算法中只有支持向量参与了核计算，即kernel machine的解的系数是稀疏的。在LR算法里，如果采用核函数，则每一个样本点都会参与核计算，这会带来很高的计算复杂度，所以，在具体应用中，LR很少采用核函数。

3、SVM不具有伸缩不变性，LR则具有伸缩不变性。

SVM模型在各个维度进行不均匀伸缩后，最优解与原来不等价，对于这样的模型，除非本来各维数据的分布范围就比较接近，否则必须进行标准化，以免模型参数被分布范围较大或较小的数据影响。LR模型在各个维度进行不均匀伸缩后，最优解与原来等价，对于这样的模型，是否标准化理论上不会改变最优解。但是，由于实际求解往往使用迭代算法，如果目标函数的形状太"扁"，迭代算法可能收敛得很慢甚至不收敛。所以对于具有伸缩不变性的模型，最好也进行数据标准化。

4、SVM损失函数自带正则项，因此，SVM是结构风险最小化算法。而LR需要额外在损失函数上加正则项。

所谓结构风险最小化，意思就是在训练误差和模型复杂度之间寻求平衡，防止过拟合，从而达到真实误差的最小化。未达到结构风险最小化的目的，最常用的方法就是添加正则项。

> 参考资料：
>
> https://www.zhihu.com/question/26726794
>
> https://blog.csdn.net/zkl99999/article/details/80907083?utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7Edefault-6.control&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7Edefault-6.control