QG 工作室数据挖掘小组实验报告

实习生： 许继元　　导师：　　日期： 2020 年 07 月 22 日

实验名称：数据挖掘理论学习第一阶段

已完成内容：

1. DecisionTree 算法
2. SVM 算法
3. AdaBoost 算法
4. RandomForest 算法

未完成内容：暂无

未完成原因：暂无

需要帮助：暂无

实验总结

知识点总结：

DecisionTree 算法：

简介：

决策树算法可用作分类，也可用作回归，属于监督学习。

类型：

基于三种划分指标，决策树算法可分为 ID3 算法（信息增益）、C4.5 算法（信息增益率）和 CART 算法（基尼指数）。

流程：

输入：训练集 D、属性集 A

过程：1.将所有的训练数据都放在根结点中

2.选择一个当前的最优属性，将根结点的数据分割成子集

3.对每个子集，选择该子集的最优属性，得到子集的子集

4.递归执行，直到各个子集都有较好的分类时结束

5.剪枝处理

实践：在 lenses、iris 和西瓜数据集上分别实现了三种算法。

优点：

- 良好的解释性及可视化性
- 数据预处理少
- 支持多输出
- 模型好坏易验证
- 支持连续变量

缺点：

- 决策树生成容易过拟合
- 模型生成不稳定，易受小错误样本影响
- 贪心搜索容易陷入局部最优
- 不支持非线性逻辑，例如 XOR

SVM 算法：

简介：

支持向量机是分类与回归分析中的一种监督学习算法，也是一种二分类模型，其基本模型定义为特征空间上间隔最大的线性分类器，且基于最大间隔分隔数据，可转化为求解凸二次规划的问题。

流程：

在 SVM 中，我们试图找到处于两类样本正中间的划分超平面。而距离超平面最近的几个样本点称为支持向量，两个异类支持向量到超平面的距离之和称为间隔，在 SVM 中我们希望实现最大间隔。

1. 利用拉格朗日乘子法求二次规划问题，求解出最大间隔超平面对应的模型，解出拉格朗日乘子后即可求解模型

2. 为了避免二次规划问题随着训练样本增加，计算开销的增加问题，使用 SMO 算法计算出对应最优解的拉格朗日乘子

3. 为了提高模型的泛化能力，引入松弛变量，允许某些样本误分类，求解软间隔支持向量机模型

3.1. 通过拉格朗日乘子法把 m 个约束转换 m 个拉格朗日乘子，得到该问题的拉格朗日函数。

3.2. 分别对参数求偏导，代入拉格朗日函数得到对偶问题。

3.3. 使用 SMO 算法求解对偶问题，解出所有样本对应的拉格朗日乘子，进而求解出模型。

4. 为了处理非线性划分，引入核函数，避免计算高维空间中的内积。

优点：

- SVM 的模型只与占训练数据少部分的支持向量有关，故 SVM 不直接依赖数据分布，所得的划分超平面不受某一类点的影响；
- 即使数据类别不平衡比较严重，SVM 也不需做相应处理。
- 具有较好的鲁棒性

缺点：

- SVM 算法对大规模训练样本难以实施
- SVM 无法直接解决多分类问题
- SVM 的输出无概率意义

AdaBoost 算法：

简介：

AdaBoost 是一种迭代算法，其思想是针对同一个训练集训练不同的弱分类器，然后集成这些弱分类器，构成一个强分类器。

流程：

基于基学习器的线性组合来最小化指数损失函数。

输入：训练集、基学习算法、训练轮数

过程：

1. 给定训练样本集 S，其中 X 和 Y 分别对应于正例样本和负例样本；T 为训练的最大循环次数；
2. 初始化样本权重为 1/n ，即为训练样本的初始概率分布；
3. 第一次迭代：
   (1) 训练样本的概率分布相当，训练弱分类器；
   (2) 计算弱分类器的错误率；
   (3) 选取合适阈值，使得误差最小；
   (4) 更新样本权重；

经 T 次循环后，得到 T 个弱分类器，按更新的权重叠加，最终得到的强分类器。

优点：

- 很好的利用了弱分类器进行级联
- 可以将不同的分类算法作为弱分类器
- 具有很高的精度
- 充分考虑的每个分类器的权重

缺点：

- 弱分类器数目不太好设定
- 数据不平衡会导致分类精度下降
- 训练比较耗时

RandomForest 算法

简介：

随机森林是 Bagging 的一个扩展变体。RF 在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机属性选择，且采用随机采样。

流程：

1. 随机采样
2. 在指定特征个数下选取最优特征
3. 构造决策树
4. 创建随机森林

优点：

- 训练可以高度并行化
- 由于可以随机选择决策树节点划分特征，这样在样本特征维度很高的时候，仍然能高效的训练模型
- 在训练后，可以给出各个特征对于输出的重要性
- 由于采用了随机采样，训练出的模型的方差小，泛化能力强
- 对部分特征缺失不敏感

缺点：

- 在某些噪音比较大的样本集上，随机森林模型容易陷入过拟合
- 取值划分比较多的特征容易对随机森林的决策产生更大的影响，从而影响拟合的模型的效果

遇到问题：	解决过程：
二次规划、凸优化、对偶问题以及拉格朗日乘子法的基础概念不够理解，在推导 SVM 算法过程中，难以理解《机器学习》中的推导思路，易遗忘。	查阅资料了解概念，找到更详细的数学推导解析的博客，并做笔记记录数学推导过程，过程中练习 latex 数学公式的编写。

导师评价				
实验分数	知识掌握情况	代码编写能力	建议	评价日期