中国地质大学(武汉)

机器学习第二次作业

姓 名：叶宇涛

专 业：计算机科学与技术

学 号：20191000595

指导老师：刘超

目录

[中国地质大学(武汉) 1](#_Toc99392108)

[机器学习第一次作业 1](#_Toc99392109)

[[3.3]实现对率回归，给出西瓜数据集3.0a结果。 3](#_Toc99392110)

[编程题目理解 3](#_Toc99392111)

[对率回归算法原理阐述 3](#_Toc99392112)

[算法设计思路 3](#_Toc99392113)

[实验流程、测试结果及分析 3](#_Toc99392114)

[代码结构，核心代码简要分析 3](#_Toc99392115)

[本次实验解决的主要问题，主要收获 3](#_Toc99392116)

[编码及内容撰写中的参考来源 3](#_Toc99392117)

# [3.3] 任意选择4个UCI数据集，对基于信息增益划分选择(ID3)、基于基尼指数划分选择(CART)，基于对率回归划分选择的决策树算法(包括未剪枝、预剪枝、后剪枝三种)进行实验比较。

## 编程题目理解

题目要求选择4个UCI数据集，分别进行基于信息增益、基尼指数、对率回归的决策树算法的实现，并且还要实现未剪枝、预剪枝、后剪枝三种不同的算法。对于不同的决策树，改变的是划分的依据，也就是信息增益、基尼指数这一类指标的不同。在这里选择了以下数据集：Iris、wine、breast\_cancer、diabetes。都是UCI中的分类数据集，并且没有少数据和异常数据的情况。其中，主要利用iris数据集进行测试。

## 决策树算法原理阐述

决策树是一种分类模型。其输入是带有标签的数据，输出是一颗决策树。其非叶节点代表的是逻辑判断；叶节点代表的是分类的子集。决策树算法原理是通过训练数据形成if-then的判断结构。从树的根节点到叶节点的每一条路径构成一个判断规则。我们需要选择合适的特征作为判断节点，可以快速的分类，减少决策树的深度。最理想的情况是，通过特征的选择把不同类别的数据集贴上对应类标签，树的叶子节点代表一个集合，集合中数据类别差异越小，其数据纯度越高。

决策树划分伪代码如下：

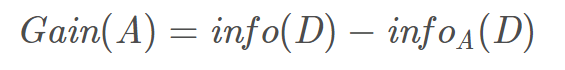


图表 1 决策树算法伪代码

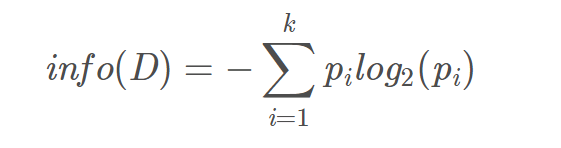
其中，根据数据特征函数的不同，可以划分不同的算法。比较经典的算法有ID3、C4.5、Cart算法。

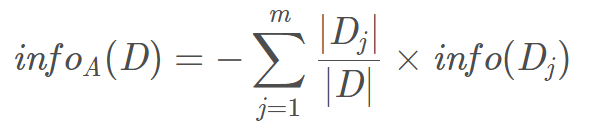
ID3：

ID3使用的数据特征函数是信息增益，给出公式如下：



其中，





计算所有类别的信息增益，选择信息增益最大的作为分类节点，这就是ID3算法的划分依据。

CART：

CART算法用基尼指数作为划分指标：

这代表每个类别中的两个样本不同的概率，也可以依据这个判断样本的纯度。

基于对率回归：

基于对率回归算法的指标就是利用对数几率回归的算法划分样本，对于一个树节点中的样本，根据全部特征可以进行对率转换为如下形式：

然后通过梯度下降算法求得对率回归的一组参数，通过这组参数，可以将该节点样本值分为两类。实现的决策树应如下图所示：



图表 2 基于对率回归算法

## 算法设计思路

首先需要构建一棵决策树，因此

## 实验流程、测试结果及分析

## 代码结构，核心代码简要分析

## 本次实验解决的主要问题，主要收获

## 编码及内容撰写中的参考来源