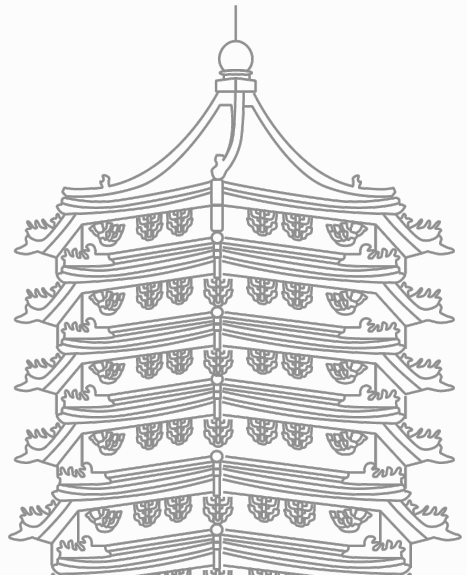
2021 决策树期末大作业

助教：高翘楚

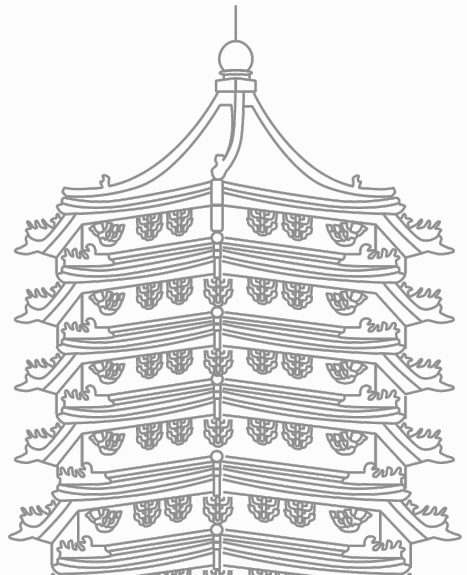
Tel：18811317822

Email：gaoqiaochu@pku.edu.cn

CONTENTS

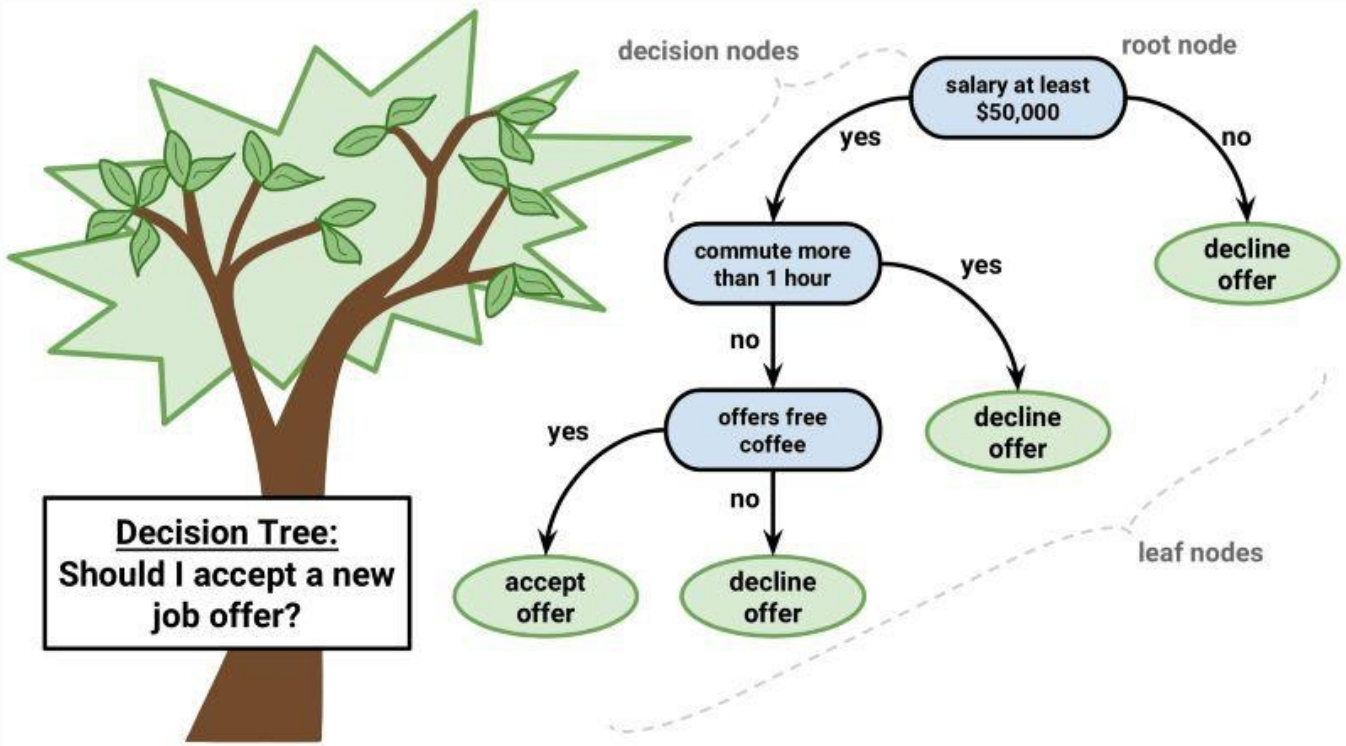


PART 01

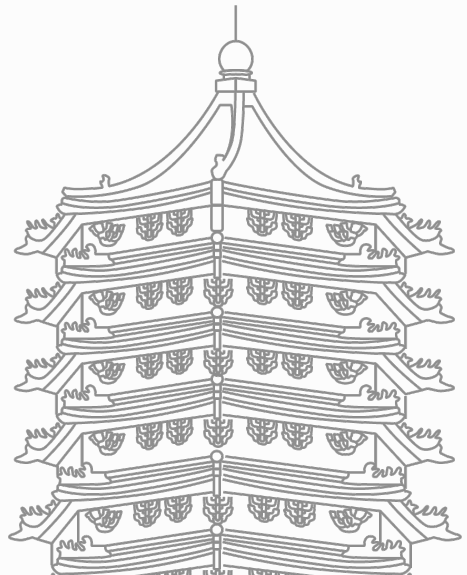


01 决策树简介

- 决策树是一类被广泛使用的**分类**与**回归**算法，它通过对问题相关的特征进行一系列测试，并且利用这些测试的结果确定最终查询样本的属类（分类任务）或是预测值（回归任务）
- 决策树由**根结点**，**中间结点**，**叶子结点**和分支结点构成，每一个非叶子结点都对应着针对一个特征的测试，被测试特征拥有的可能取值数目决定了其拥有的向下分支的数目。每一个叶子结点都对应了一种预测结果。



PART 02



02 决策树构建

2.1 信息增益（Information Gain）

- 信息增益是一种对于某个特征所包含的信息量的度量。具体而言，某个特征的信息增益等于**原数据集的熵**减去对该特征进行测试并根据测试结果**分割原数据集后各个子数据集的熵**的加权和。
- 数学模型为：

$$IG(d,D) = H(t,D) - rem(d,D)$$

其中原数据集的熵：

$$H(t,D) = -\sum_{l \in levels(t)} P(t=l) \times \log_2 P(t=l)$$

各个子数据的熵的加权和：

$$rem(d,D) = \sum_{l=level(d)} \frac{|D_{d=l}|}{|D|} \times H(t, D_{d=l})$$

02 决策树构建

2.1 信息增益（Information Gain）

（1）计算原数据集按照目标属类进行划分的信息熵 $H(t,D)$；

（2）对于每一个特征，根据其不同的取值，将数据集划分为若干子数据集，计算每个子数据集内部按照目标属类进行划分的信息熵 $H(t,D_{d=l})$，并对各个子数据集的结果进行加权求和；

（3）原数据集的信息熵减去（2）中得到的加权和得到每个特征的信息增益。

02 决策树构建

2.1 信息增益（Information Gain）

例：

ID	SUSPICIOUS WORDS	UNKNOWN SENDER	CONTAINS IMAGES	CLASS
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	spam
693	false	true	true	ham
782	false	false	false	ham
976	false	false	false	ham

（1）计算整个数据集根据目标变量划分后的熵：

$$H(t,D) = -\sum_{l\in\{spam,ham\}} P(t=l)\times \log_2 P(t=l)$$

$$= -[P(t=spam)\times \log_2 P(t=spam) + P(t=ham)\times \log_2 P(t=ham)]$$

$$= -\left[\frac{3}{6}\times \log_2 \frac{3}{6} + \frac{3}{6}\times \log_2 \frac{3}{6}\right] = 1 \text{ bit}$$

02 决策树构建

2.1 信息增益（Information Gain）

例：

ID	SUSPICIOUS WORDS	UNKNOWN SENDER	CONTAINS IMAGES	CLASS
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	spam
693	false	true	true	ham
782	false	false	false	ham
976	false	false	false	ham

（2）计算以某个特征对D进行划分之后各子数据集的熵的加权平均

以suspicious words 为例

$$rem(\text{Words, D}) = \left(\frac{|D_{\text{Words=true}}|}{|D|}\times H(t, D_{\text{Words=true}}) + \frac{|D_{\text{Words=false}}|}{|D|}\times H(t, D_{\text{Words=false}})\right)$$

$$= \left\{\frac{3}{6}\times\left[-\sum_{l\in\{spam,ham\}} P(t=l)\times\log_2 P(t=l)\right] + \frac{3}{6}\times\left[-\sum_{l\in\{spam,ham\}} P(t=l)\times\log_2 P(t=l)\right]\right\}$$

$$= \left\{\frac{3}{6}\times\left[-\left(\frac{3}{3}\times\log_2\frac{3}{3}+\frac{0}{3}\times\log_2\frac{0}{3}\right)\right] + \frac{3}{6}\times\left[-\left(\frac{0}{3}\times\log_2\frac{0}{3}+\frac{3}{3}\times\log_2\frac{3}{3}\right)\right]\right\} = 0\ bit$$

02 决策树构建

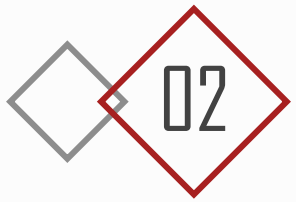
2.1 信息增益（Information Gain）

例：

ID	SUSPICIOUS WORDS	UNKNOWN SENDER	CONTAINS IMAGES	CLASS
376	true	false	true	spam
489	true	true	false	spam
541	true	true	false	spam
693	false	true	true	ham
782	false	false	false	ham
976	false	false	false	ham

（3）计算信息增益IG

$$IG(Words, D) = H(t, D) - rem(Words, D) = 1 - 0 = 1bits$$



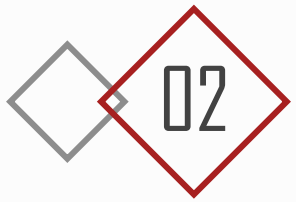
决策树构建

2.1 信息增益（Information Gain）

小结：

信息增益的物理意义：信息增益的物理意义类似于互信息，某个特征的信息增益越大表示将原数据集按照这个特征进行划分之后得到的子数据集的熵的数学期望越小，或者说，**针对该特征的测试能够提供最多的信息量**。

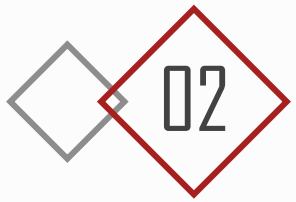
因此，信息增益适合于进行特征选择：需要决定根据某个特征对数据集进行划分时，选择具有最大信息增益的特征即可。



决策树构建

2.2 基于信息增益的决策树构建算法——ID3算法

- ID3算法是一个**递归的、深度优先的**决策树构造算法，整个决策树的构造过程依赖于树的**深度优先遍历**。
- 首先，根据信息增益选择一个特征进行测试，并将该特征对应的测试作为决策树的根节点
- 训练数据集根据该测试进行划分，对于每一个测试结果都产生一个子数据集，对于每一个子数据集，都将其作为根结点的一个子结点
- 递归：对于根结点的每一个子结点不断重复：每个子结点都是一个子树（subtree）的根结点，其对应的数据集是父节点的数据集的一个子集，同时它在根据信息增益进行特征选择时不用考虑在父节点中已经测试过的特征。
- 停止条件：如果子数据集中所有样本都属于同一类，或是子数据集样本数为0，或是子数据集能够使用的特征集合为空集。



02 决策树构建

2.2 基于信息增益的决策树构建算法——ID3算法

Algorithm 1 ID3 Algorithm

Input:

set of descriptive features $\boldsymbol{d}$
set of training instances $\mathcal{D}$

Output: decision tree T

if all the instance in $\mathcal{D}$ have the same target level C **then**
    return decision tree T cosisting of single leaf node with label C
end if
if $\boldsymbol{d} = \emptyset$ **then**
    return decision tree T cosisting of single leaf node with label of the the majority target level in $\mathcal{D}$
end if
if $\mathcal{D} = \emptyset$ **then**
    return decision tree T cosisting of single leaf node with the label of the the majority target level of the dataset of the its parent node
end if
$\boldsymbol{d}[best] \leftarrow \arg\max_{d \in \mathcal{D}} \ IG(d, \mathcal{D})$.
make a new node $Node_{d[best]}$ and label it with $\boldsymbol{d}[best]$.
partition dataset $\mathcal{D}$ using $\boldsymbol{d}[best]$.
$\boldsymbol{d} = \boldsymbol{d} - \boldsymbol{d}[best]$
for each partition $\mathcal{D}_i$ of $\mathcal{D}$ **do**
    ID3($\boldsymbol{d}$, $\mathcal{D}_i$)
end for

决策树构建

2.3 改进的特征选择指标

- 信息增益是经典的特征选择指标，但是它也存在缺点：信息增益倾向于选择具有更多可能取值的特征，因为如果一个特征具有很多取值，那么根据这个特征进行数据集划分之后各个子数据集的大小很可能都比较小，而比较小的数据集更可能是纯度高（低熵）的，于是信息增益较大。由于信息增益会偏爱具有更多取值的特征，从而导致其选择了一些不相关特征进行测试，后续的研究提出使用信息增益比例（Information Gain Ratio, IGR）来进行特征选择。IGR的表达为：

$$IGR(d,D)=\frac{IG(d,D)}{-\sum_{l\in level(D)}P(d=l)\times\log_2 P(d=l)}$$

- 分母是按照特征d对数据集进行分割的熵。根据IGR的表达式，某个特征d如果具有很多取值，那么$IG(d,D)$可能较大，与此同时分母也会较大，因此IGR对于IG偏爱多取值特征这一缺陷有一定的修正作用。

PART 03

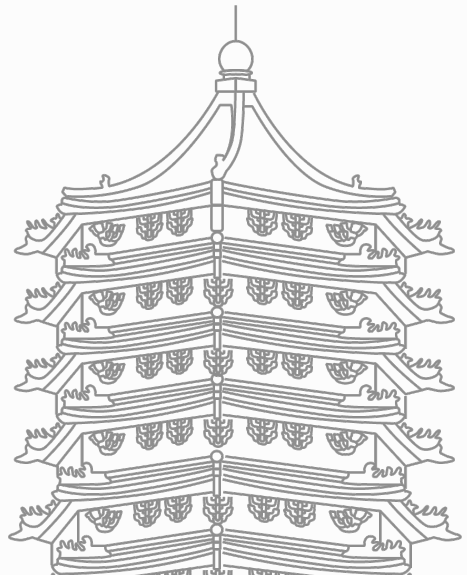
01 决策树简介

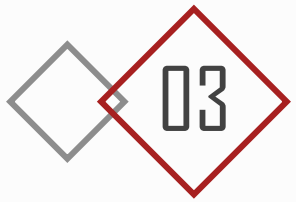
02 决策树构建

03 决策树剪枝

04 数据集准备

05 作业提交





决策树构建

3.1 剪枝的原因

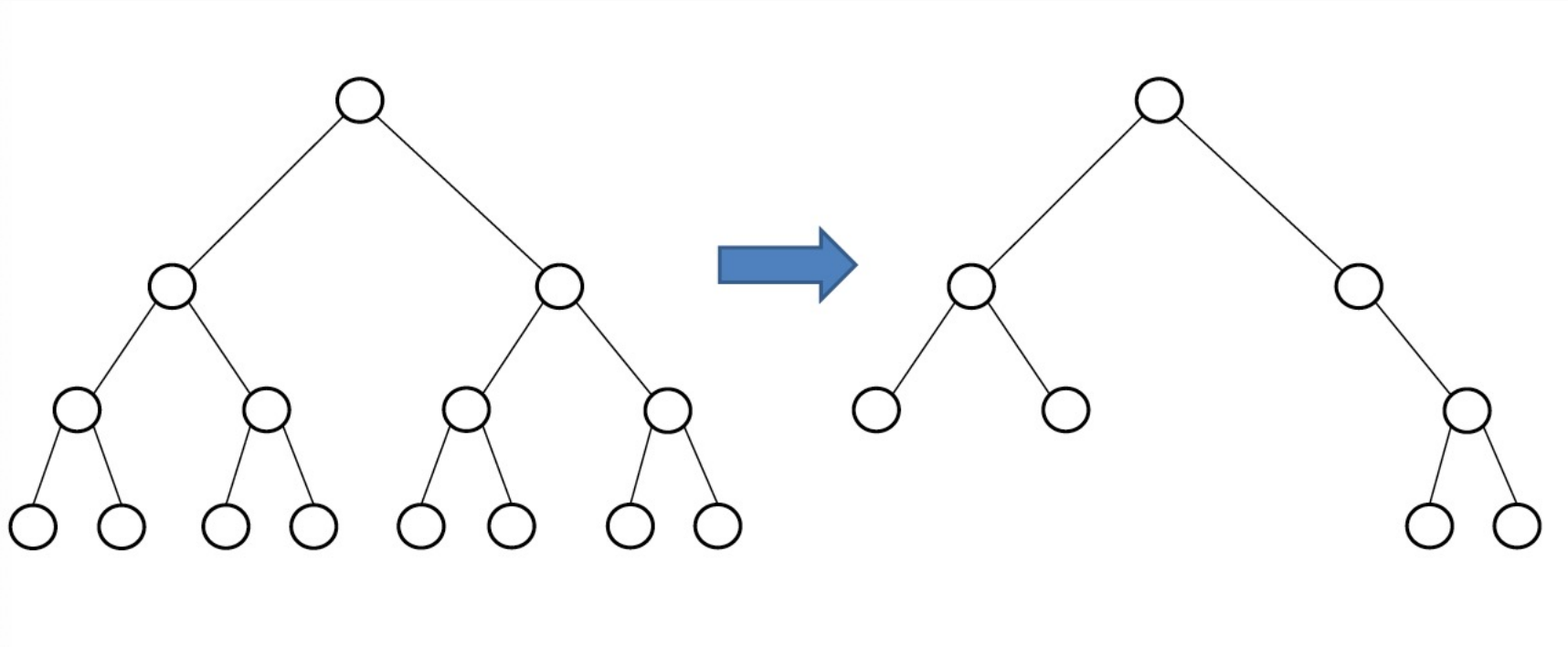
对抗过拟合

- 由于决策树构建算法递归地划分数据集的特性，天然地会导致产生很多叶子结点，这些叶子结点中只包含一些噪声样本。为了解决这个问题，决策树剪枝成为广泛使用地防止决策树模型过拟合的方法。显然，剪枝后的决策树不能完美地拟合训练数据集，但是却**滤除了一些噪声样本**，从而提高了模型的鲁棒性与泛化能力。

决策树构建

3.2 剪枝的含义

- 将决策树的一个子树除了其根结点之外全部从决策树中移除，从而其根结点变为剪枝之后的决策树中的一个叶子结点



3.3 剪枝的策略

- **早停止策略**

早停止策略是比较简单的剪枝方式，也被叫做预剪枝（pre-pruning）。早停止的判断标准可以是**最大树深度**，每个叶子结点包含的**最少样本数目**，**最小信息增益**等等。

- **后剪枝策略**

在后剪枝策略中，首先根据标准决策树构建算法，如上节中的ID3等，构建一棵未剪枝的决策树。随后从完整的决策树的叶子结点开始，使用**自底向上**（bottom-up）的方式进行剪枝，剪枝的过程即删除叶子结点的过程。

3.3 剪枝的策略

- 基于验证集错误率的剪枝方法：

设当前决策树为DT，它的某个叶子结点为L_i，将**验证数据集**在当前决策树DT下进行测试，得到测试错误率R_1。将叶子结点L_i从当前决策树DT中剔除，得到剪枝后的决策树DT_{pruned}。将验证数据集在剪枝后的决策树DT_{pruned}下测试，得到测试错误率R_2，若$R1 \geq R_2$，则将叶子结点L_i从当前决策树DT中删去。上述过程从决策树最底层的叶子结点开始，使用**层次优先**的遍历方式遍历决策树的各个结点，并在每个结点出都考虑是否将其**对应的子树**从决策树中删除，直到达到根结点为止。

PART 04

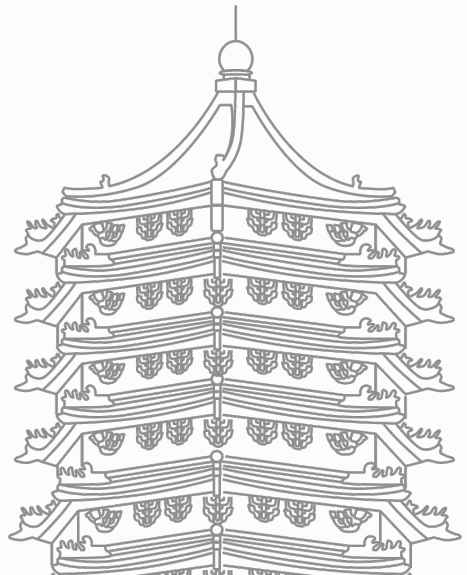
01 决策树简介

02 决策树构建

03 决策树剪枝

04 数据集准备

05 作业提交



04 数据集准备

数据集分析实例

- 汽车评估数据集：一共包含6个分类特征和四个类别，1728个测试实例

以下是数据集的大致分布情况

CLASS	N	N[%]
unacc	1210	70.023
acc	384	22.222
good	69	3.993
v-good	65	3.762

还可以分析各个特征值的取值分布

Feature	buying	maint
Value and count	med': 432	vhigh': 217
	high': 350	high': 216
	low': 83	med': 216
		low': 216

04 数据集准备

数据集来源

- 数据集的特征都是类别（categorical）特征，且问题都是分类问题
- UCI数据集官网： http://archive.ics.uci.edu/ml/index.php

可以使用一下三个数据集，也可以在UCI官网或者其它来源自行寻找数据集；如自选数据集，请在报告中注明数据集基本信息，将数据集一并打包上传

- 井字游戏数据集：985个样本，9个分类特征，二分类问题
  http://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame
- 心理实验数据集：625个样本，4个分类特征，分成三类
  http://archive.ics.uci.edu/ml/datasets/Balance+Scale
- 幼儿园数据集：12960个样本，8个分类特征，分成五类
  http://archive.ics.uci.edu/ml/datasets/Nursery

PART 05

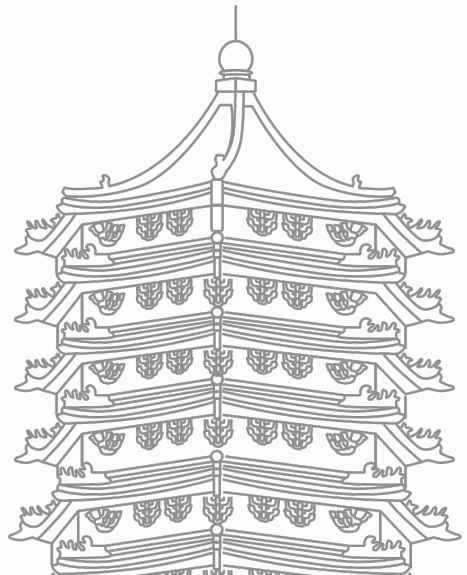
01 决策树简介

02 决策树构建

03 决策树剪枝

04 数据集准备

05 作业提交



05 作业提交

5.1 代码和实验

- 本次作业需要提交一份**报告**以及**源代码文档**。使用的**编程语言不限**，要求最终的大作业代码文档中包含一个**ReadMe文档**，文档中详细描述**使用方式以及预期结果**。
- 代码要求能够在其他机器上正常运行，正常运行的含义是指：如果采用Python等脚本类语言，需要在提交的大作业代码文档的ReadMe中详细注明**入口脚本以及命令行参数的设置方式**；如果采用C++等静态语言，要求尽可能将所需要的第三方库函数以源代码形式放在最终的代码文档中，并通过CMakeList或MakeFile等方式明确编译方法，以**保证在其他机器上不修改CMakeList或MakeFile的条件下编译出可执行程序**。此外，如果有些同学编写的是C++的动态链接库配合脚本语言，则需要详细写明**动态链接库的编译与安装方法**。

05 作业提交

5.2 作业内容

基础部分

- 给出对于数据集特征的分析，例如样本分布等
- 实现基于**信息增益**（IG）与**信息增益比例**（IGR）的决策树构造算法，比较性能区别并试分析之；
- 对基于IG和IGR构造的决策树实现**预剪枝策略**，比较二者对于最终测试准确率的影响。

附加部分（Optional）

- 对基于IG和IGR构造的决策树实现**后剪枝策略**，比较二者对于最终测试准确率的影响。

报告中需要包括在**三个数据集**上的测试结果和分析

05 作业提交

5.3 注意事项

- 本次作业不允许调用现成的库或者API完成决策树的构造与测试过程，所有的核心算法必须自己独立完成，能用的库只能是线性代数类的库以及文件读写的库。如果发现直接调用现成库函数或是抄袭开源代码者一律按照0分处理。
- 完成附加部分可以获得一定的加分，加分额度视对附加题的完成度决定，最多不超过10分。
- 本次作业鼓励大家报告构造决策树所需要的时间，如果在构造决策树过程中使用了特殊的数据结构或使用了底层优化技术对程序进行了加速，欢迎大家在报告中着重强调，助教会根据报告中对内容核实其报告内容的可信度以及有效性，如果确有速度上的提升，亦可酌情加分。

05 作业提交

5.4 Tips

在完成决策树的基线算法后

- 可使用python的sklearn库中的决策树模块验证算法正确性
- 可参考西瓜书第4章决策树的内容逐步验证算法正确性

05 作业提交

5.5 提交

提交途径

- 教学网——教学内容——2021决策树大作业

提交内容

- 全部源码文件、相关文档、pdf格式报告
- 全部内容打包为zip文件上传
- zip文件命名为：学号_姓名_决策树大作业

截止日期：6月16日 24：00

2021

Thanks for Listening