



决策树

丁诚



实验课安排

- 1 数据处理与KNN
- 2-3 决策树
- 4 逻辑回归 (logistic regression) 和 感知机算法 (perception learning algorithm)
- 5 反向传播神经网络 (BPNN)
- 6 数据降维
- 7 卷积神经网络 (CNN)
- 8 循环神经网络 (RNN)
- 9 Boosting和Bagging



批改报告

原理： 30

伪代码流程图： 20

关键代码截图： 10

实验结果分析： 30

思考题： 10

创新： +10

我们不会给出具体分数，只会给ABCD的等级，并在公布等级时指出哪个部分没写好。

每次的报告我们会评2~4份优秀报告，直接加实验课总评分1~2分

实验分数组成： 签到10%， 报告50%， 期中pro15%， 期末pro25%

训练集 验证集 测试集的区别

数据类型	有无标签	作用
训练集(training set)	有	用来训练模型或确定模型参数的，如k-NN中权值的确定等。 相当于平时练习。
验证集(validation set)	有	用来确定网络结构或者控制模型复杂程度的参数，修正模型。 相当于模拟考试。
测试集(test set)	无	用于检验最终选择最优的模型的性能如何。 相当于期末考试。



上一周的问题

1. 关于模型结果，可以调库来检查自己模型的正确与否。以及AI课只会教导模型原理，不回答代码问题。
2. 关于调库，原则上只能调涉及计算的库，例如numpy, pandas等，公式的实现需要自己完成。
3. 训练集验证集测试集都是semeval划分出来的。数据的过拟合和欠拟合
(<https://www.cnblogs.com/yejintianming00/p/9338917.html>)
4. 数据集处理，词表太长可以怎么处理，stemming, stopwords, embedding等
5. 关键代码才截图
6. 测试集不需要交。之后只提供数据集，如何划分数据集以及验证效果由学生自己决定。
7. 除了期末pro，不要在意指标。考查的是对模型的理解而不是指标，不要花时间在调参上！！
保证模型没出错的情况下也不要太纠结指标为什么这么差。
8. 只收代码，不要交类似环境数据一类的东西。
9. 提交版本可以加v1, v2，例如命名格式为“学号_姓名拼音_lab1_v1.pdf”，最多提交3个版本的报告或者代码。
10. 实验最多迟交一周，补交降级
11. 不要忘记验收！！



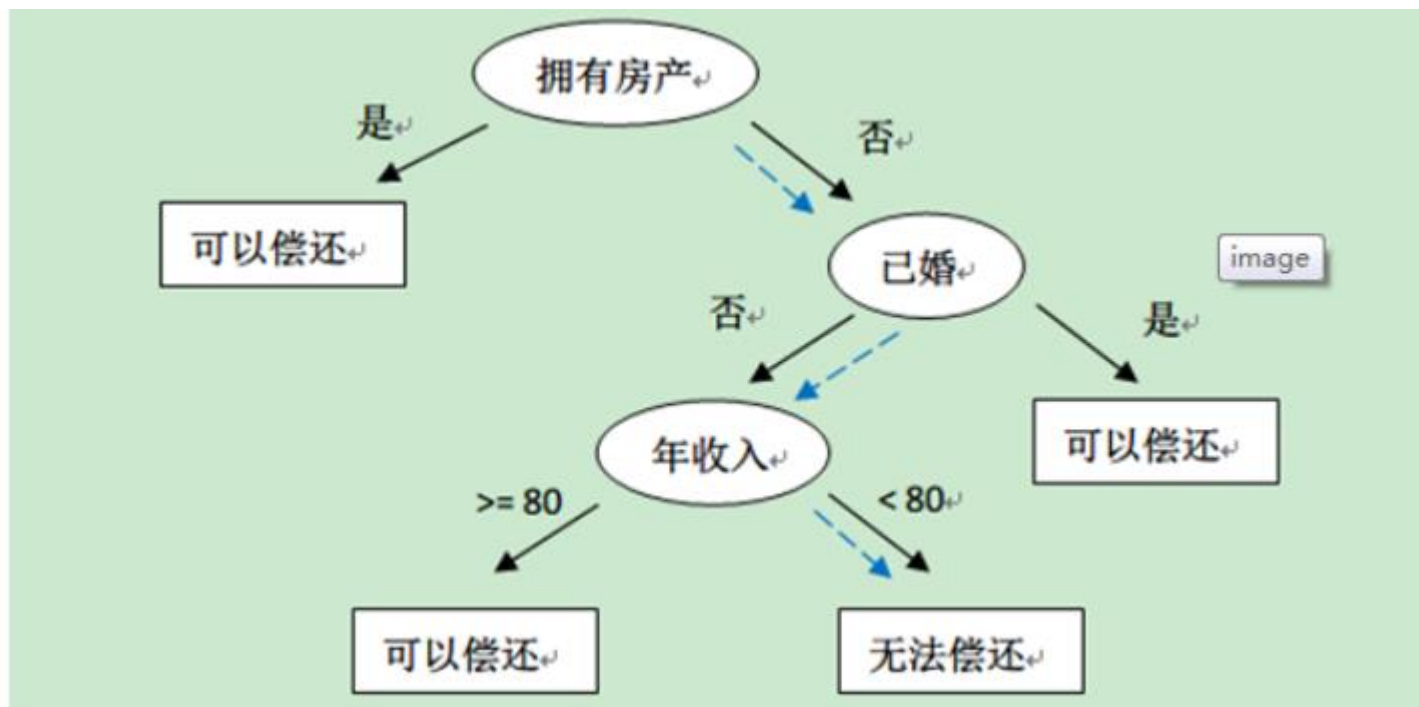
目录

- 简单回顾
- 建树步骤
- 特征选择
- 处理连续型特征
- 剪枝
- 思考题
- 实验要求



简单回顾

- 有监督 (supervised)
- 分类模型
- ID3, C4.5, CART
- 树形结构





建树步骤

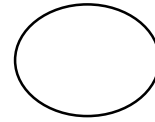
1. 初始化:
创建根结点，它拥有全部数据集和全部特征。
2. 选择特征:
遍历当前结点的数据集和特征，根据某种原则，选择一个特征。
3. 划分数据:
根据这个特征的取值，将当前数据集划分为若干个子数据集。
4. 创建结点:
为每个子数据集创建一个子结点，并删去刚刚选中的特征。
5. 递归建树:
对每个子结点，回到第2步。直到达到边界条件，则回溯。
6. 完成建树:
叶子结点采用多数投票的方式判定自身的类别。



举例

age	income	student	credit_rating	buys_computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
31...40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31...40	low	yes	excellent	yes
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
> 40	medium	no	excellent	no

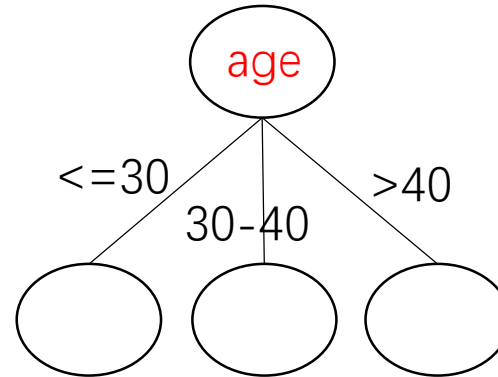
首先创建根结点：



假设选定特征age。

该特征有三种取值： ≤ 30 , 31-40, > 40 。

那么，为根结点添加三个子结点。





举例

- age='<=30'

age	income	student	credit_rating	buy_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
<=30	medium	yes	excellent	yes

- age='31-40'

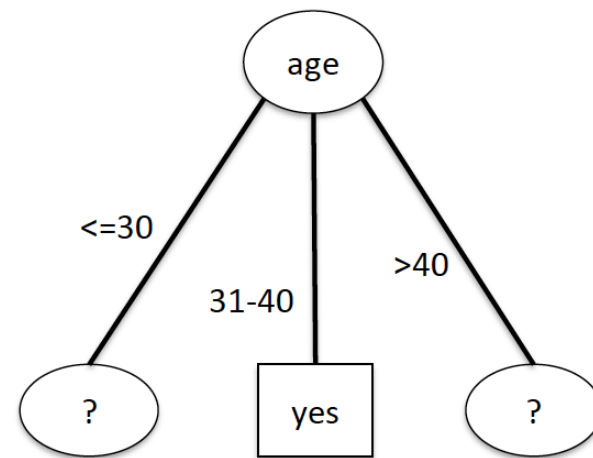
age	income	student	credit_rating	buy_computer
31-40	high	no	fair	yes
31-40	low	yes	excellent	yes
31-40	medium	no	excellent	yes
31-40	high	yes	fair	yes

- age='>40'

age	income	student	credit_rating	buy_computer
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
>40	medium	yes	fair	yes
>40	medium	no	excellent	no

如左图所示，根据age特征的不同取值，将数据集划分为三个子数据集。

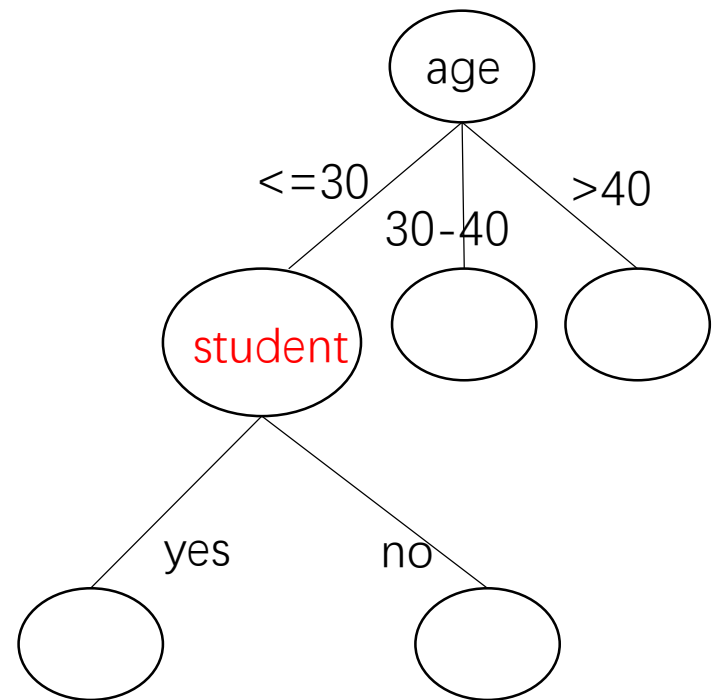
如下图所示，每个子数据集分配给一个子结点。左结点和右结点都可以继续划分，而中间结点的数据标签全为`yes`，无需划分。





举例

income	student	credit_rating	buy_computer
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes



现在对上一步中的左结点进一步划分。

该结点的数据，应当是上一步中，age特征 ≤ 30 的那些数据，即如左图所示。注意age特征已被删去。

假设从剩下的特征中，选中student特征来划分这些数据。它有yes和no两种取值，所以为左结点添加两个子结点。结果如右图所示。



建树步骤

1. 初始化:
创建根结点，它拥有全部数据集和全部特征。
2. 选择特征:
遍历当前结点的数据集和特征，根据某种原则，选择一个特征。
3. 划分数据:
根据这个特征的取值，将当前数据集划分为若干个子数据集。
4. 创建结点:
为每个子数据集创建一个子结点，并删去刚刚选中的特征。
5. 递归建树:
对每个子结点，回到第2步。直到达到边界条件，则回溯。
6. 完成建树:
叶子结点采用多数投票的方式判定自身的类别。



递归的边界条件

假设当前结点的数据集为 D ，特征集为 A

1. D 中的样本属于同一类别 C ，则将当前结点标记为 C 类叶结点。
2. A 为空集，或 D 中所有样本在 A 中所有特征上取值相同，此时无法划分。将当前结点标记为叶结点，类别为 D 中出现最多的类。
3. D 为空集，则将当前结点标记为叶结点，类别为父结点中出现最多的类。



特征选择

- 3种方法
- 使用信息增益：ID3
- 使用信息增益率：C4.5
- 使用GINI指数：CART



ID3

- 决策策略：信息增益（Information Gain）
- 步骤：

（1）计算数据集D的经验熵

$$H(D) = - \sum_{d \in D} p(d) \log p(d)$$

（2）计算特征A对数据集D的**条件熵** $H(D|A)$

$$H(D|A) = \sum_{a \in A} p(a) H(D|A = a)$$

（3）计算信息增益

$$g(D, A) = H(D) - H(D|A)$$

（4）选择**信息增益最大**的特征作为决策点



ID3举例

数据	长鼻子(x)	大耳朵(y)	是否大象(1 or 0)
A1	1	1	1
A2	0	1	0
A3	1	0	0
A4	0	0	0

1. 计算经验熵:

$$H(D) = -1/4 * \log(1/4) - 3/4 * \log(3/4)$$

2. 计算每个特征下的条件熵:

$$H(D|A="x") = (2/4) * (-(1/2) * \log(1/2) - (1/2) * \log(1/2)) \\ + (2/4) * (0 - (2/2) * \log(2/2))$$

$$H(D|A="y") = (2/4) * (-(1/2) * \log(1/2) - (1/2) * \log(1/2)) + (2/4) * 0$$

3. 计算: 信息增益

$$g(D, A="x") = H(D) - H(D|A="x")$$

$$g(D, A="y") = H(D) - H(D|A="y")$$

4. 选择信息增益最大的特征作为决策点



C4.5

- 决策策略：信息增益率（Information Gain Ratio）
- 步骤：

（1）计算特征A对数据集D的信息增益

$$g(D, A) = H(D) - H(D|A)$$

（2）计算数据集D关于特征A的值的熵SplitInfo(D, A)

$$\text{SplitInfo}(D, A) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log \left(\frac{|D_j|}{|D|} \right)$$

（3）计算信息增益率

$$g\text{Ratio}(D, A) = (H(D) - H(D|A)) / \text{SplitInfo}(D, A)$$

（4）选择信息增益率最大的特征作为决策点



C4.5举例

数据	长鼻子(x)	大耳朵(y)	是否大象(1 or 0)
A1	1	1	1
A2	0	1	0
A3	1	0	0
A4	0	0	0

1. 计算在每个特征条件下的信息增益

$$g(D, A="x")=H(D)-H(D|A="x"); \quad g(D, A="y")=H(D)-H(D|A="y")$$

2. 计算每个特征的熵

$$\text{SplitInfo}(D, A="x")=-(2/4)*\log(2/4)-(2/4)*\log(2/4)$$

$$\text{SplitInfo}(D, A="y")=-(2/4)*\log(2/4)-(2/4)*\log(2/4)$$

3. 计算信息增益率

$$\text{gainRatio}(D, A="x")=g(D, A="x")/\text{SplitInfo}(D, A="x")$$

$$\text{gainRatio}(D, A="y")=g(D, A="y")/\text{SplitInfo}(D, A="y")$$

4. 选择信息增益率最大的特征作为决策点



CART

- 决策策略：GINI系数（Gini Index，值越小表示不确定性越小）
- 步骤：
 - （1）计算特征A的条件下，数据集D的GINI系数

$$\text{gini}(D, A) = \sum_{j=1}^v p(A_j) \times \text{gini}(D_j | A = A_j)$$

其中：

$$\text{gini}(D_j | A = A_j) = \sum_{i=1}^n p_i(1 - p_i) = 1 - \sum_{i=1}^n p_i^2$$

v表示属性A的取值个数，n表示类别个数

- （2）选择GINI系数最小的特征作为决策点



CART举例

数据	长鼻子(x)	大耳朵(y)	是否大象(1 or 0)
A1	1	1	1
A2	0	1	0
A3	1	0	0
A4	0	0	0

1. 计算在每个特征的条件下，数据集的**GINI**系数

$$\begin{aligned} \text{gini}(D, A="x") &= |D_{x=1}|/|D| * \text{gini}(D_{x=1}) + |D_{x=0}|/|D| * \text{gini}(D_{x=0}) \\ &= (2/4) * [1 - (1/2)^2 - (1/2)^2] + (2/4) * [1 - (2/2)^2 - 0^2] \\ \text{gini}(D, A="y") &= |D_{y=1}|/|D| * \text{gini}(D_{y=1}) + |D_{y=0}|/|D| * \text{gini}(D_{y=0}) \\ &= (2/4) * [1 - (1/2)^2 - (1/2)^2] + (2/4) * [1 - 1^2 - 0^2] \end{aligned}$$

2. 选择**GINI**系数最小的特征作为决策点



处理连续型特征

- 以上的例子中，遇到的都是离散型特征。
- 离散型特征：取值可以看成有限集合，比如 {yes, no} 或 {high, medium, low}。
- 连续型特征：取值可以看成是一个区间，比如 $[0, 10]$ 或 $(-\infty, +\infty)$ 。
- 问题：连续型特征的取值理论上是无穷多的，这样就要求无穷多个子结点，如何处理？



处理连续型特征

- 方法：把连续型特征当作离散型处理。
- 举例：某一特征所有取值为 $\{0.15, 0.21, 0.32, 0.39, 0.53\}$ 。
- 这样的特征虽然是连续型的，但由于数据集有限，所以出现的取值也有限，所以可以当成离散型来处理。
- 用两个数的中位数来划分：0.18, 0.265, 0.355, 0.46。
- 小于0.18则为0，0.18-0.265则为1，0.265-0.355则为2，0.355-0.46则为3，大于0.46则为4。
- 这样，该特征的取值变为了 $\{0, 1, 2, 3, 4\}$ 。
- 该方法可以改进，比如只用0.15和0.53的中位数0.34来划分，则该特征只有两种取值： $\{\leq 0.34, > 0.34\}$ 。这样树的分支会更少，树的结构会更简单。



剪枝

- 作用：提升泛化性能
 - 方法：使用验证集
 - 种类：预剪枝 VS 后剪枝
-
- 网上有更多的剪枝方法，这里只讲最简单的两种，参考周志华《机器学习》



预剪枝

- 在决策树生成过程中进行。
- 对于当前的结点，判断是否应当继续划分。如果无需划分，则直接将当前结点设置为叶子结点。
- 如何判断：假设基于ID3，选择了某个特征进行划分。如果划分后，决策树在验证集上的准确率不提高，则无需划分。



后剪枝

- 先生成完整的决策树，再自底向上地对非叶结点进行考察。
- 后序遍历。
- 对于某个非叶结点，假如将它变成叶子结点，决策树在验证集上的准确率不降低，则将它变成叶子结点。



思考题

- 决策树有哪些避免过拟合的方法？
- C4.5相比于ID3的优点是什么，C4.5又可能有什么缺点？
- 如何用决策树来进行特征选择（判断特征的重要性）？



实验要求

- 实现ID3, C4.5, CART三种决策树，只提交一份代码
- 不要求实现连续型数据的处理
- 不要求实现剪枝
- 本次数据为Car_train.csv每个文件有7列，前6列为特征（都为离散型），最后一列是标签（0 or 1）。
- 请自行分好训练验证测试集（在报告里说明怎么分的），评测指标为验证集和测试集上的准确率
- 提交文件
 - 实验报告：17*****_wangxiaoming.pdf。
 - 代码：17*****_wangxiaoming.zip。如果代码分成多个文件，最好写份readme.txt。
- DDL: 09-20号 23:00:00
- 验收方式：我们会在上课前给出小数据集，然后当堂跑实验结果来验证算法的正确性。



课外拓展

- 随机森林
 - <https://blog.csdn.net/zrjdds/article/details/50133843>
 - <https://zhuanlan.zhihu.com/p/28217071>
- 提升树 (AdaBoost Decision Tree)
 - <https://blog.csdn.net/tianxiaguixin002/article/details/47701881>
- GBDT (Gradient Boosting Decision Tree)
 - <https://blog.csdn.net/w28971023/article/details/8240756>