北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Data Analytics with R
语言数据分析

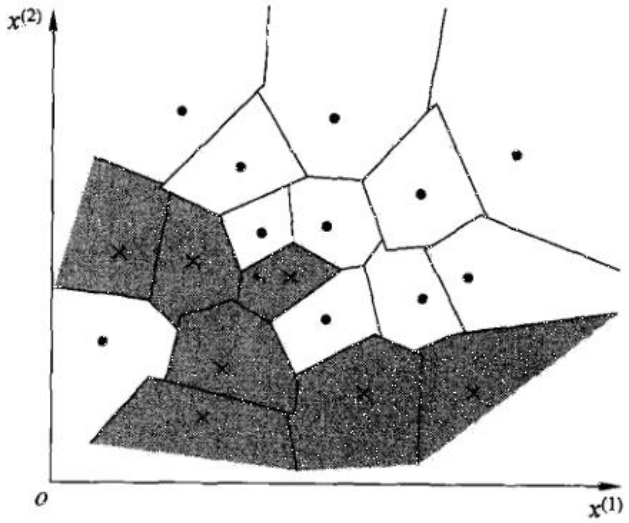
既是世间法、自当有分别

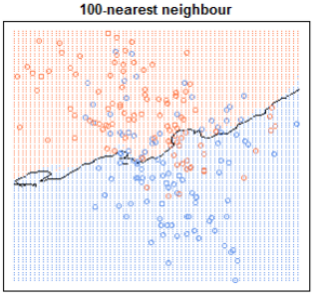
艾新波 / 2018 • 北京

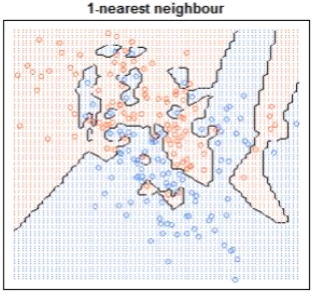
课程体系

- R语言数据分析
  - 上部：论道 
    - 第1章 气象万千、数以等观
    - 第2章 所谓学习、归类而已
    - 第3章 格言联璧话学习
    - 第4章 源于数学、归于工程
  - 中部：执具 
    - 第5章 工欲善其事必先利其器
    - 第6章 基础编程
    - 第7章 数据对象
  -
    - 第8章 人人都爱tidyverse
    - 第9章 最美不过数据框
  - 下部 博术 
    - 第10章 观数以形
    - 第11章 相随相伴、谓之关联
    - 第12章 既是世间法、自当有分别
    - 第13章 方以类聚、物以群分
    - 第14章 庐山烟雨浙江潮

近邻法：空间划分的角度

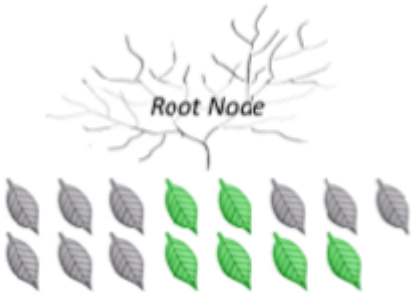


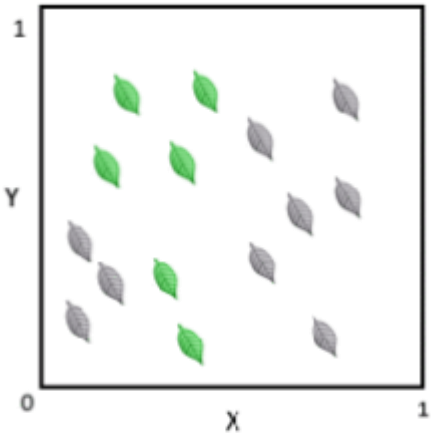




左图引自: 李航. 统计学习方法. 北京：清华大学出版社, pp.38

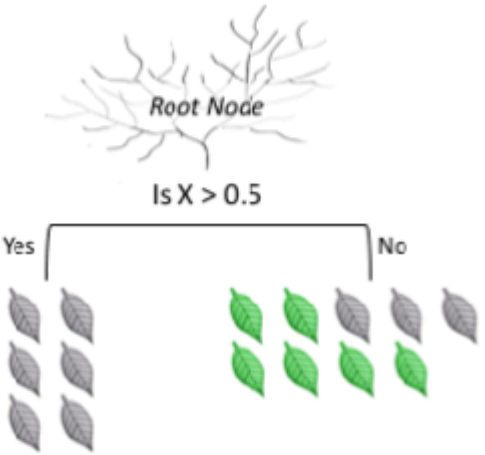
另一种空间划分的方法

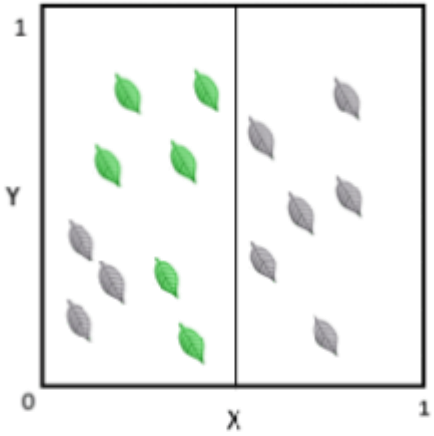




https://algobeans.com/

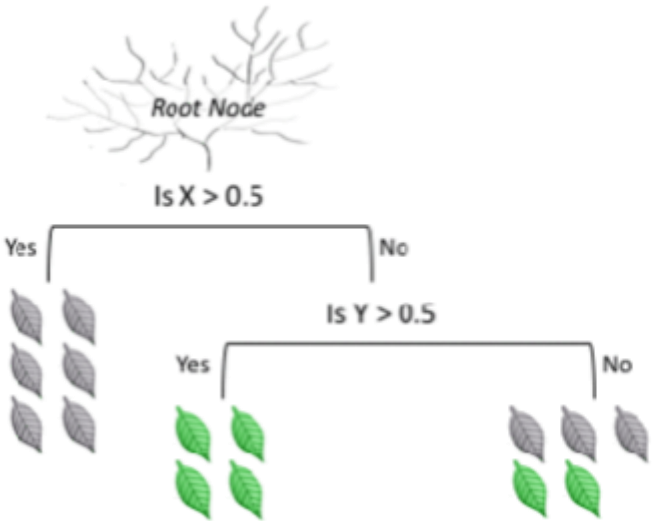
另一种空间划分的方法

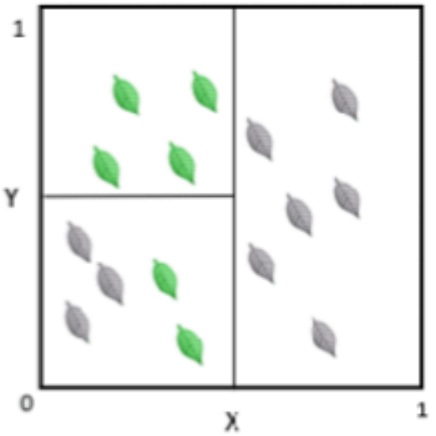




https://algobeans.com/

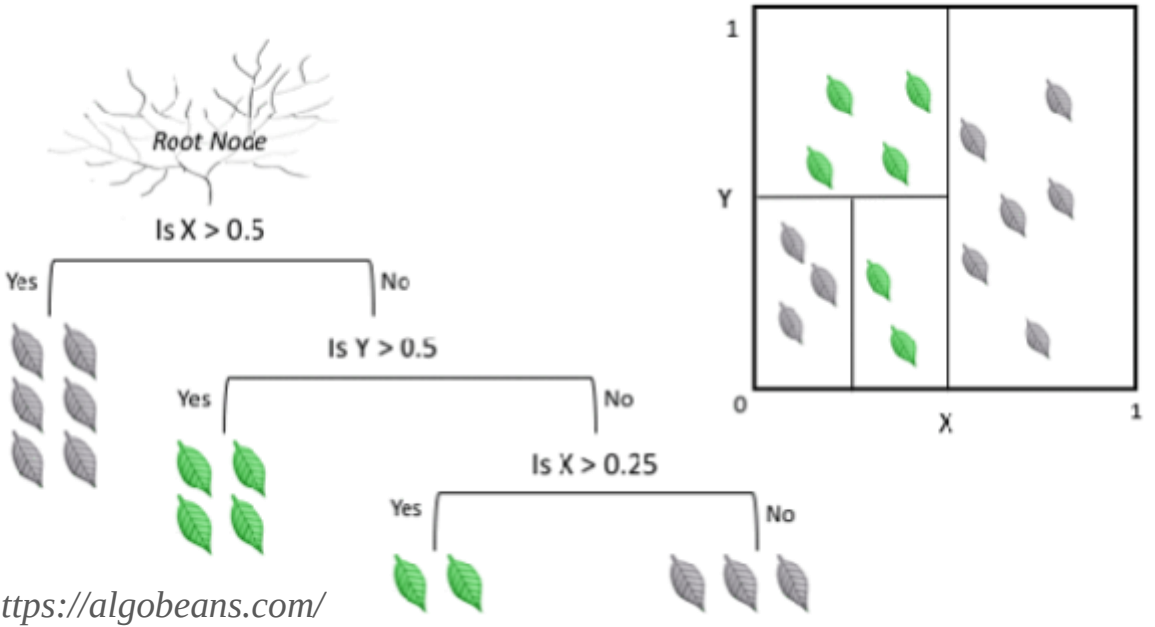
另一种空间划分的方法





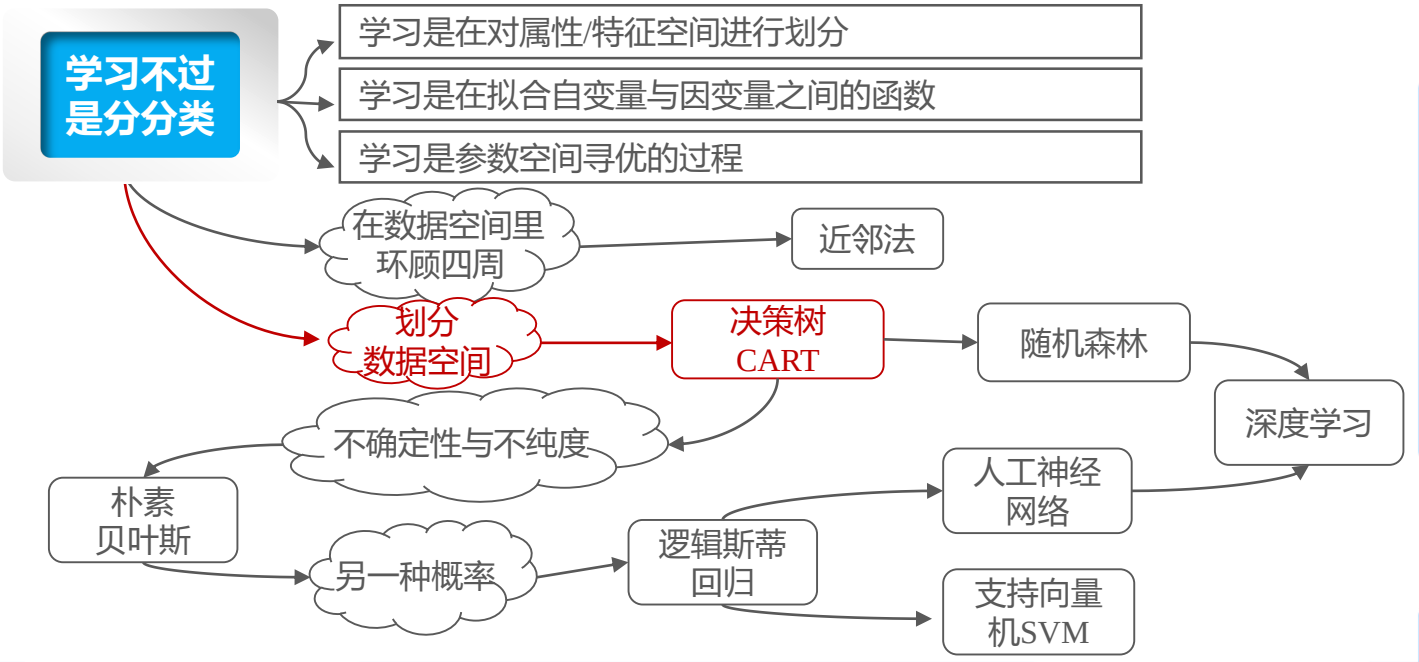
https://algobeans.com/

另一种空间划分的方法



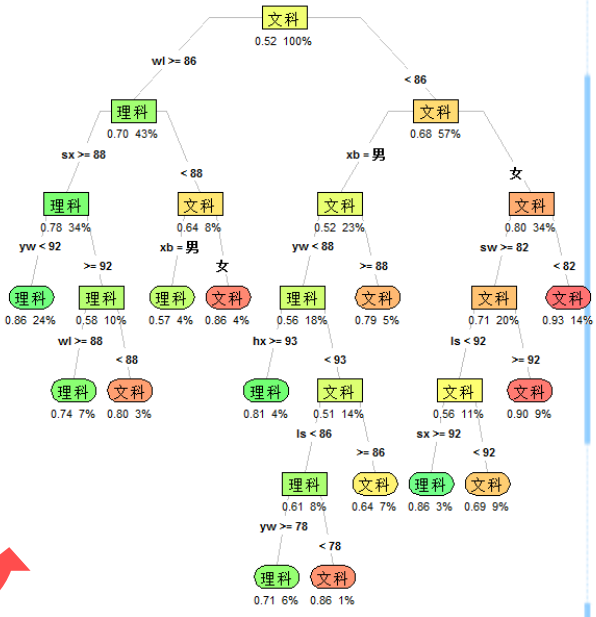
https://algobeans.com/

算法模型


学习不过是分分类
学习是在对属性/特征空间进行划分
学习是在拟合自变量与因变量之间的函数
学习是参数空间寻优的过程
在数据空间里环顾四周
近邻法
划分数据空间
决策树CART
随机森林
深度学习
不确定性与不纯度
朴素贝叶斯
另一种概率
逻辑斯蒂回归
人工神经网络
支持向量机SVM

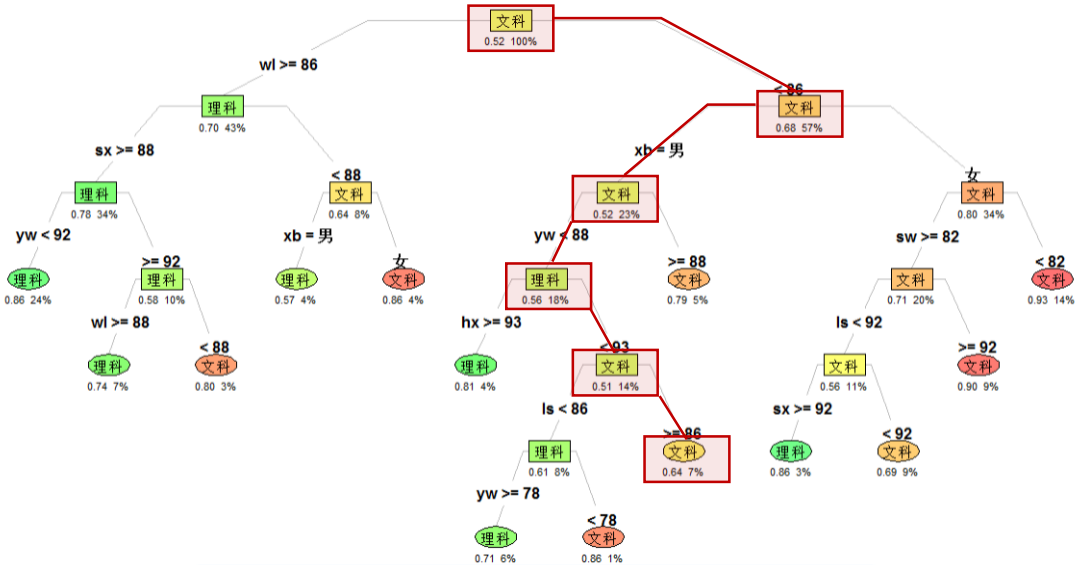
数据框背后的规律：一张表、一棵树

xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
男	85	81	92	95	86	90	81	92	76	文科
男	89	96	82	93	97	96	74	100	73	文科
女	91	85	88	89	61	92	71	82	79	理科
女	94	82	96	97	97	98	95	94	88	文科
女	87	73	93	95	85	94	61	90	85	文科
男	88	95	87	93	96	92	77	92	90	理科
男	87	90	90	96	99	100	90	96	95	文科
女	80	84	91	91	85	100	73	90	90	文科
女	92	95	94	96	92	94	94	100	97	理科



数据框背后的规律：一张表、一棵树

2	3				5		1	4		6
xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
男	81	89	87	97	94	96	81	88	83	?



根据特征进行决策

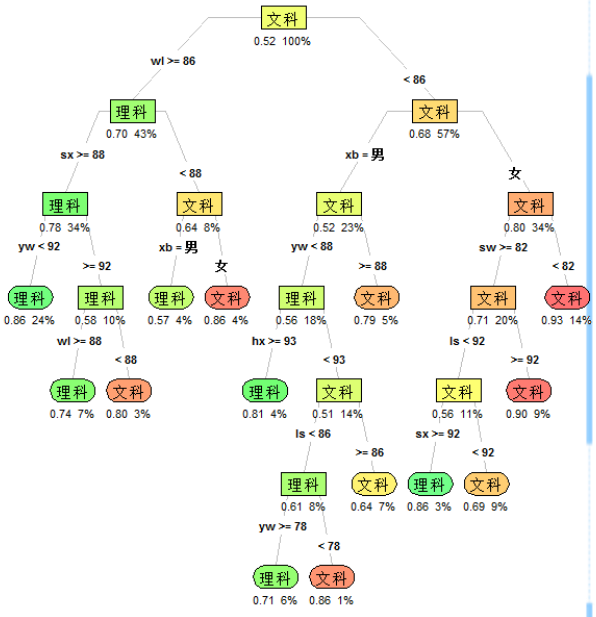
xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
男	85	81	92	95	86	90	81	92	76	文科
男	89	96	82	93	97	96	74	100	73	文科
女	91	85	88	89	61	92	71	82	79	理科
女	94	82	96	97	97	98	95	94	88	文科
女	87	73	93	95	85	94	61	90	85	文科
男	88	95	87	93	96	92	77	92	90	理科
男	87	90	90	96	99	100	90	96	95	文科
女	80	84	91	91	85	100	73	90	90	文科
女	92	95	94	96	92	94	94	100	97	理科

```
IF:
 wl < 86
 & xb = 男
 & yw < 88
 & hx < 93
 & ls >= 86
THEN:
 wlfk = 文科
.....
```

全局最优是一个NP问题：不同特征组合进行判断，面临的将是组合爆炸

根据特征进行决策

xb	yw	sx	wy	zz	ls	dl	wl	hx	sw	wlfk
男	85	81	92	95	86	90	81	92	76	文科
男	89	96	82	93	97	96	74	100	73	文科
女	91	85	88	89	61	92	71	82	79	理科
女	94	82	96	97	97	98	95	94	88	文科
女	87	73	93	95	85	94	61	90	85	文科
男	88	95	87	93	96	92	77	92	90	理科
男	87	90	90	96	99	100	90	96	95	文科
女	80	84	91	91	85	100	73	90	90	文科
女	92	95	94	96	92	94	94	100	97	理科



决策树：局部最优、步步为赢

根据特征进行决策

序号	性别	语文	文理分科
1	女	好	理科
2	男	不好	理科
3	女	好	文科
4	男	好	理科
5	女	不好	文科
6	男	好	文科
7	女	好	文科
8	男	不好	理科
9	男	好	理科

不同特征有不同的判别能力：以性别和语文为例

根据特征进行决策

序号	性别	语文	文理分科
2	男	不好	理科
4	男	好	理科
6	男	好	文科
8	男	不好	理科
9	男	好	理科
1	女	好	理科
3	女	好	文科
5	女	不好	文科
7	女	好	文科

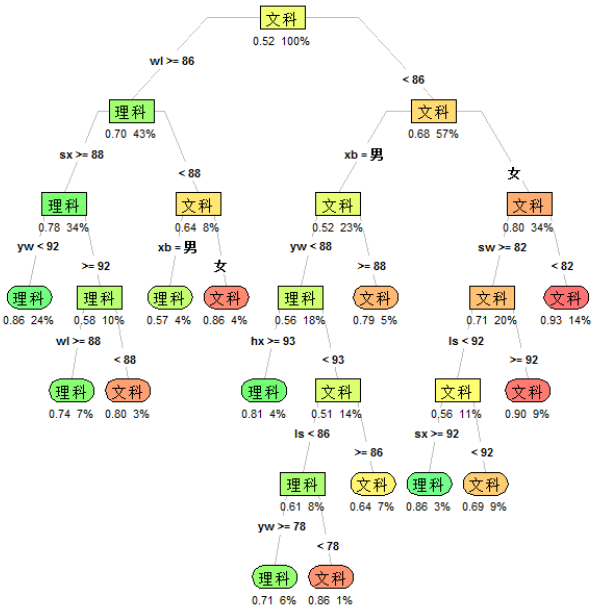
序号	性别	语文	文理分科
2	男	不好	理科
8	男	不好	理科
5	女	不好	文科
4	男	好	理科
6	男	好	文科
9	男	好	理科
1	女	好	理科
3	女	好	文科
7	女	好	文科

不同特征有不同的判别能力：以性别和语文为例

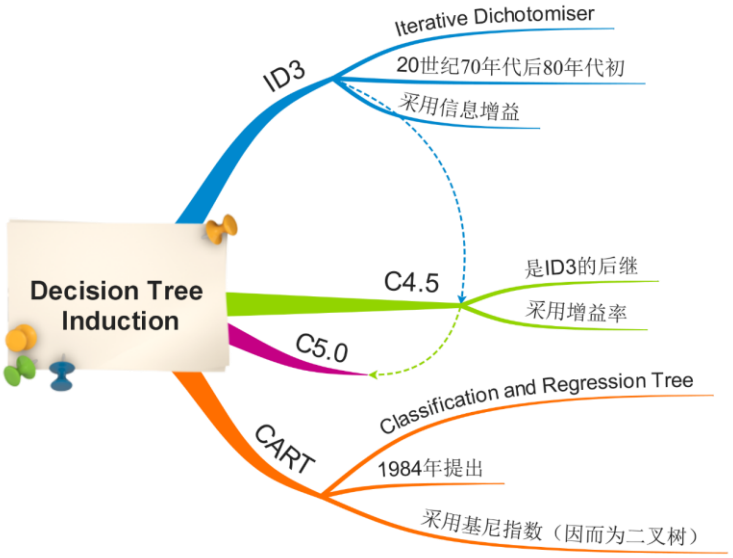
决策树归纳法

决策树归纳算法TreeGrowth(D, F)

```
if stopping_cond(D, F) = TRUE then
    leaf = createNode()
    leaf.label = Classify(D)
    return leaf
else 将边(root → child)标记为v
    root = createNode()
    root.test_cond = find_best_split(D, F)
    令V = {v|v是root.test_cond的一个可能的输出}
    for 每个v ∈ V do
        D_v = {d|root.test_cond(d) = v且d ∈ D}
        child = TreeGrowth(D_v, F)
        将child作为root的派生节点添加到树中
    end for
end if
return root
```



决策树归纳法



常用的不纯度指标：

$$\mathbf{Entropy}(t) = -\sum_{\mathrm{i}=1}^{c} \mathrm{p}(\mathrm{i}|t) log_2 \mathrm{p}(\mathrm{i}|t)$$

$$\mathbf{Gini}(t) = 1 - \sum_{\mathrm{i}=1}^{c} [\mathrm{p}(\mathrm{i}|t)]^2$$

$$\mathbf{Classification\ error}(t) = 1 - \max_i [\mathrm{p}(\mathrm{i}|t)]$$

增益Δ——不纯度之差：

$$\Delta = \mathrm{I}(parent) - \sum_{\mathrm{j}=1}^{k} \frac{\mathrm{N}(v_j)}{N} \mathrm{I}(v_j)$$

最小化子节点不纯度加权平均值

序号	性别	语文	文理分科
2	男	不好	理科
4	男	好	理科
6	男	好	文科
8	男	不好	理科
9	男	好	理科
1	女	好	理科
3	女	好	文科
5	女	不好	文科
7	女	好	文科

$$\sum_{j=1}^{k}\frac{N(v_j)}{N}I(v_j)=\frac{5}{9}\times\frac{1}{5}+\frac{4}{9}\times\frac{1}{4}=0.22$$

序号	性别	语文	文理分科
2	男	不好	理科
8	男	不好	理科
5	女	不好	文科
4	男	好	理科
6	男	好	文科
9	男	好	理科
1	女	好	理科
3	女	好	文科
7	女	好	文科

$$\sum_{j=1}^{k}\frac{N(v_j)}{N}I(v_j)=\frac{3}{9}\times\frac{1}{3}+\frac{6}{9}\times\frac{3}{6}=0.44$$

决策树归纳法

决策树的生长：递归划分、不断生长；局部最优、步步为赢

如果数据无需再分（如当前数据记录类标签同为一类）、或是无法再分（如所有记录的属性值相同、或是数据记录数太少），则建立叶子结点，并按照少数服从多数（有生于无）的原则，给叶子结点打上标签；

否则，寻找一个属性，根据该属性的不同取值情况，把数据分成纯度较大的两个（或几个）子集。对于这些子集，递归执行以上操作，开枝散叶。

R语言实现：Task Views

Recursive Partitioning

Package rpart is recommended for computing CART-like trees.

The C50 package can fit C5.0 classification trees, rule-based models, and boosted versions of these.

Extensible tools for visualizing binary trees and node distributions of the response are available in package party as well.

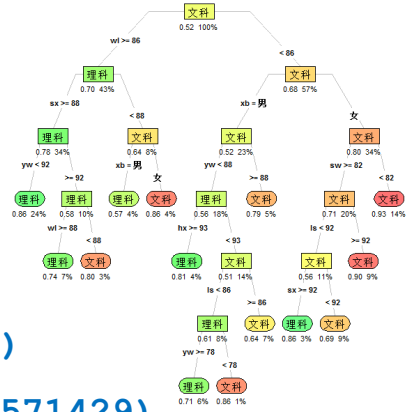
An adaptation of rpart for multivariate responses is available in package mvpart.

Graphical tools for the visualization of trees are available in package maptree.

R语言实现

```
#rpart.plot包会自动加载rpart包
library(rpart.plot)
imodel <- rpart(wlfk~.,
                data = cjb[train_set_idx,])
imodel
```

R语言实现




```
#> n= 542
#> node), split, n, loss, yval, (yprob)
#> * denotes terminal node
#> 1) root 542 266 文科 (0.4907749 0.5092251)
#>   2) sx>=85.5 322 115 理科 (0.6428571 0.3571429)
#>     4) wl>=86.5 178  36 理科 (0.7977528 0.2022472)
#>       8) ls< 95.5 126  16 理科 (0.8730159 0.1269841) *
#>       9) ls>=95.5 52  20 理科 (0.6153846 0.3846154)
#>         18) sw>=92.5 36   8 理科 (0.7777778 0.2222222) *
#>         19) sw< 92.5 16   4 文科 (0.2500000 0.7500000) *
```

R语言实现

```
#训练集上的拟合效果
predicted_train <-
  predict(imodel,
          newdata = cjb[train_set_idx,],
          type = "class")
Metrics::ce(cjb$wlfk[train_set_idx],
            predicted_train)
#> [1] 0.1959335
```

R语言实现

```
#当然，我们更关注的是测试误差
predicted_test <-
  predict(imodel,
          newdata = cjb[-train_set_idx, ],
          type = "class")
Metrics::ce(cjb$wlfk[-train_set_idx],
            predicted_test)
#> [1] 0.2575107
```

过拟合举例：瓶子的二值分类








谢谢聆听

Thank you

教师个人联系方式

艾新波

手机：13641159546

QQ： 23127789

微信：13641159546

E-mail: 13641159546@126.com

axb@bupt.edu.cn

地址：北京邮电大学科研楼917室

课程网址：https://github.com/byaxb/RDataAnalytics