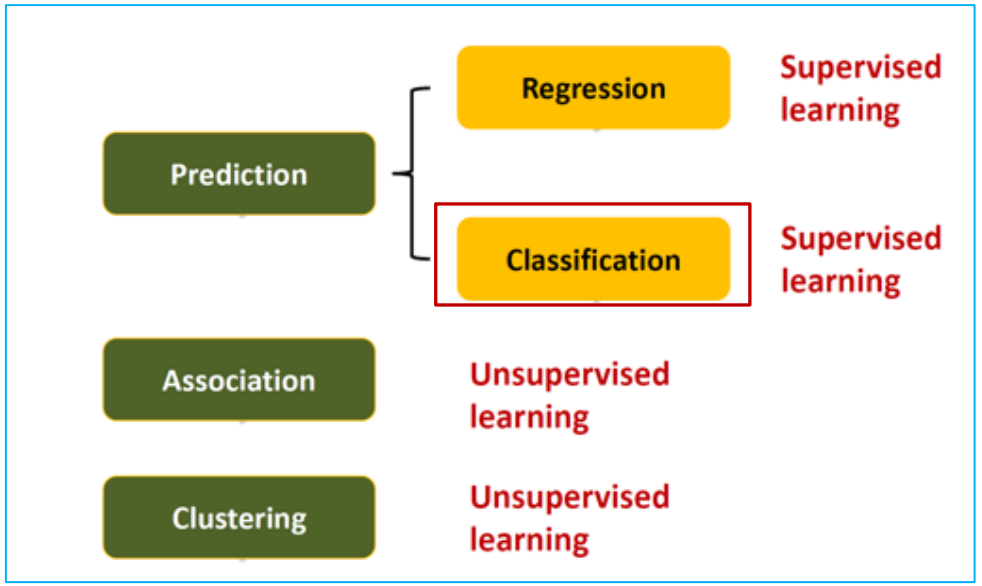
MM5425 商业分析

WEEK 5 LECTURE – DECISION TREE

WK5 内容

- 线性回归回顾
- 决策树Decision Tree
  - 商业案例: 流失Churn
  - 什么是决策树
  - 概率分类
  - 构建分类树



分类是通过添加标签将数据集划分为不同类别或组的过程。

商业中的流失

客户流失是指客户或订阅者停止与某个服务提供商进行业务往来的情况。

客户流失是企业的一个关键指标。：

- 降低成本：保留现有客户比获取新客户的成本更低。.
- 更高利润：NPS评分体系的创始人Fred Reichheld发现，如果你仅仅保留5%的客户，长期来看至少可以带来25%以上的利润增长。

通过监控流失率，组织试图制定提升客户保留率的策略。

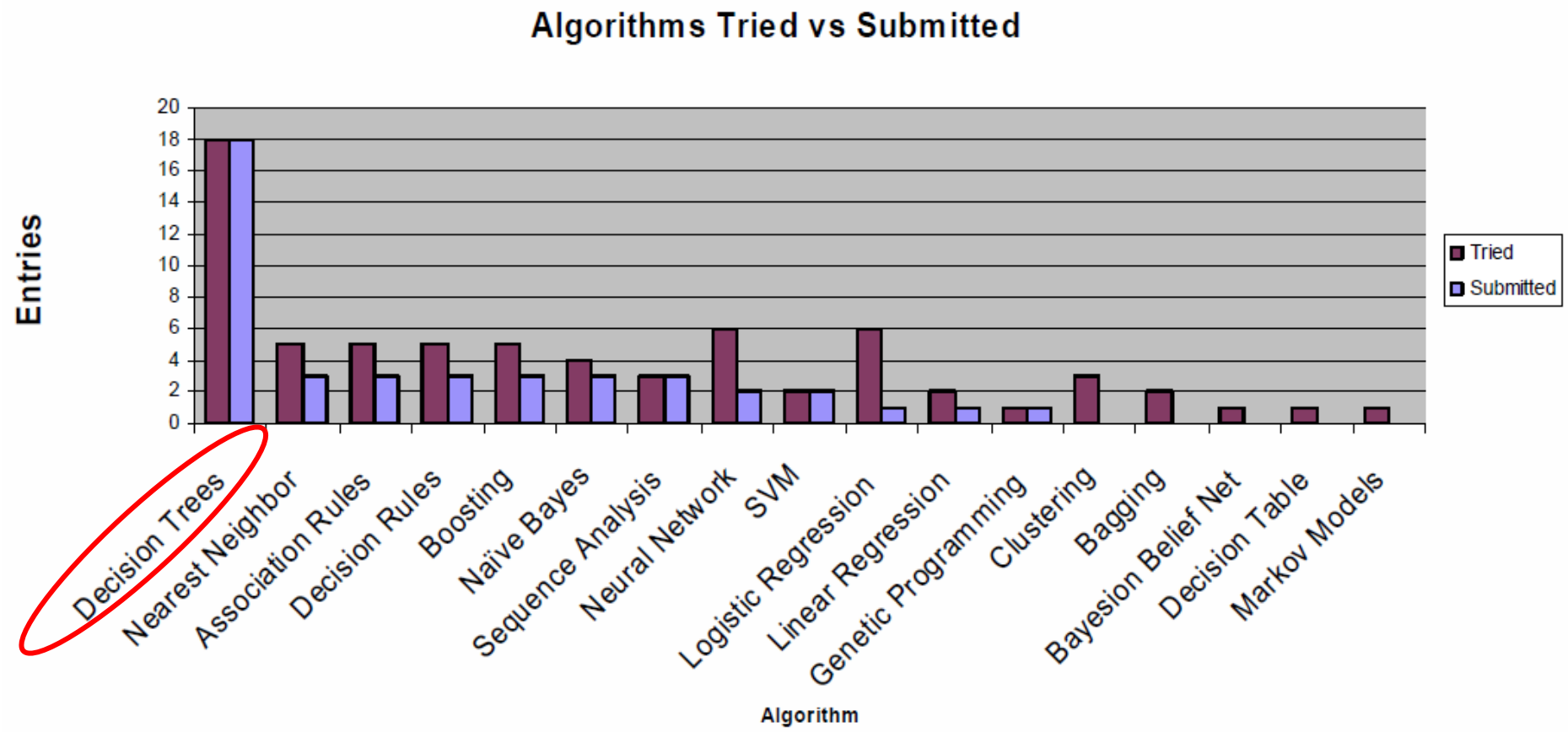
- 一种方法是针对合适的客户并采取干预措施以防止客户流失。.

https://blog.gramener.com/churn-analysis-customer-retention/
https://posthog.com/product-engineers/churn-rate-vs-retention-rate

决策树

NODES AND LEAVES

常用商业分析算法

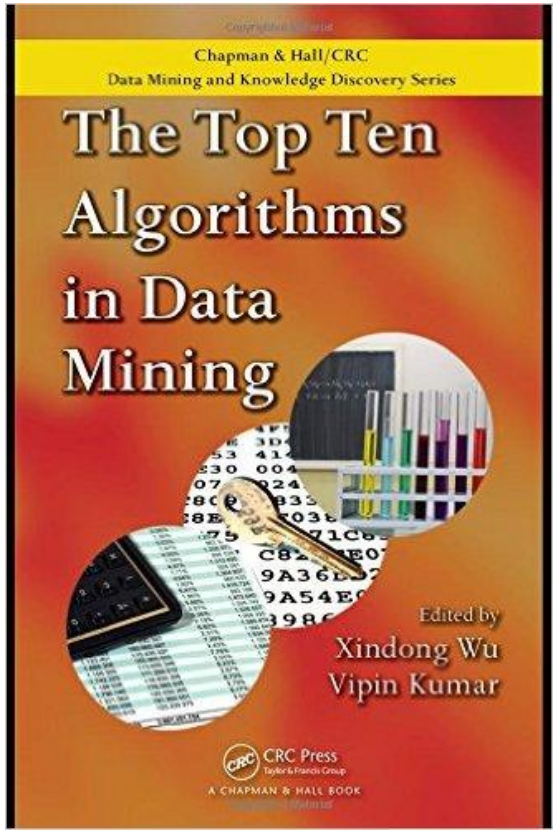

Algorithms Tried vs Submitted
Entries
20
18
16
14
12
10
8
6
4
2
0
Tried
Submitted
Decision Trees
Nearest Neighbor
Association Rules
Decision Rules
Boosting
Naïve Bayes
Sequence Analysis
Neural Network
SVM
Logistic Regression
Linear Regression
Genetic Programming
Clustering
Bagging
Bayesion Belief Net
Decision Table
Markov Models
Algorithm

为什么采用决策树?

决策树（或分类树）是最受欢迎的数据挖掘工具之一。

- 易于理解
- 易于实现
- 易于使用
- 计算成本低

决策树在模型可解释性方面具有优势，这对于模型评估以及向非业务分析专业人士传达结果非常重要。



分类树

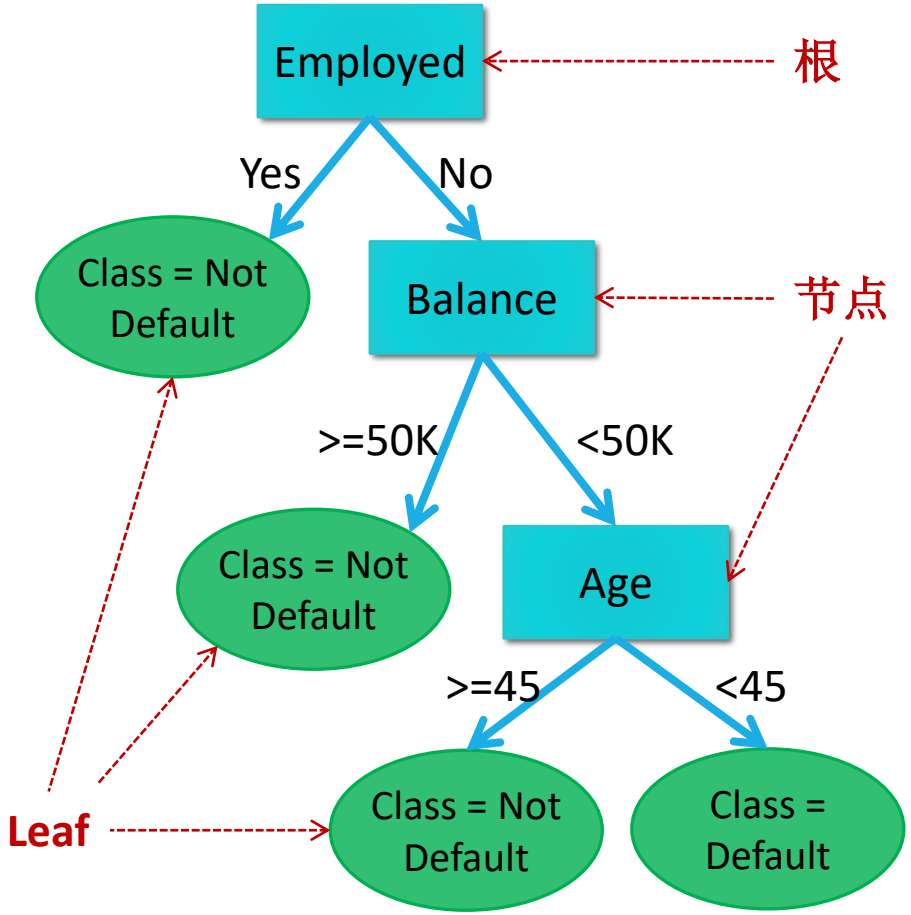
Employed	Balance	Age	Default
Yes	123,000	50	No
No	41,100	40	Yes
No	48,000	55	No
Yes	34,000	46	No
Yes	50,000	46	No
No	100,000	25	No

从数据样本中生成分类树

IF Employed=No and Balance < 50K and Age < 45
Then Default = 'yes'
Else Default = 'no'

分类树

分类树（倒置）
表示方法与术语



分类树如何用于分类?

要确定一个新样本的类别，例如：Mark，年龄40岁，已退休，账户余额38K。

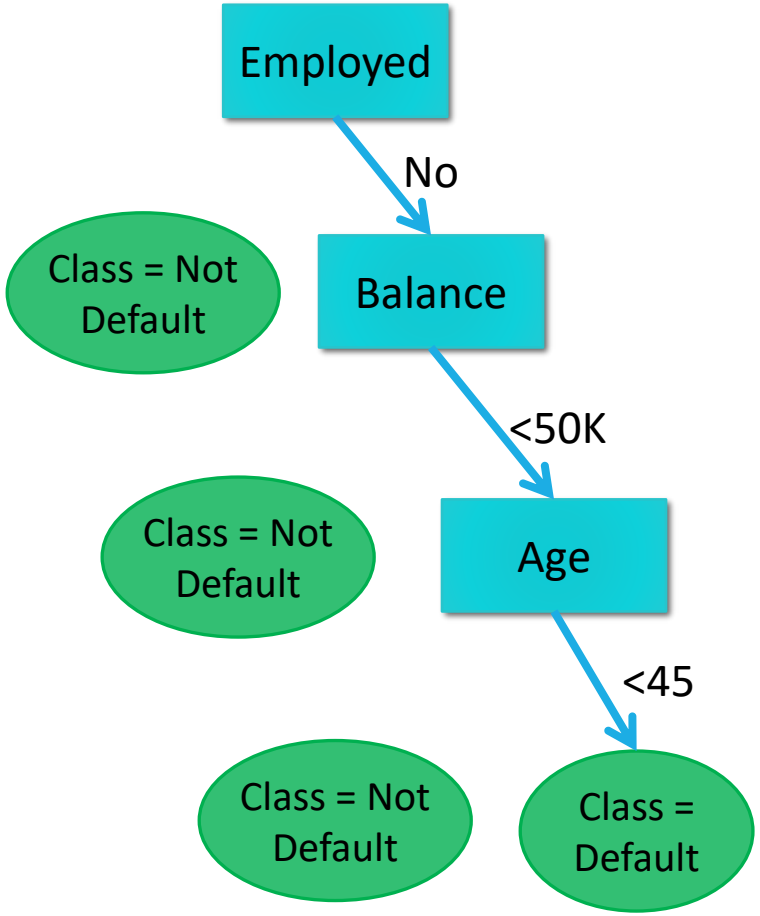
该样本会根据各属性的取值，依次沿着树结构向下分流：

每个节点：在每个节点上，都会对一个属性进行测试（如年龄、职业、余额等）。

选择分支：根据测试结果，选择相应的分支继续向下。

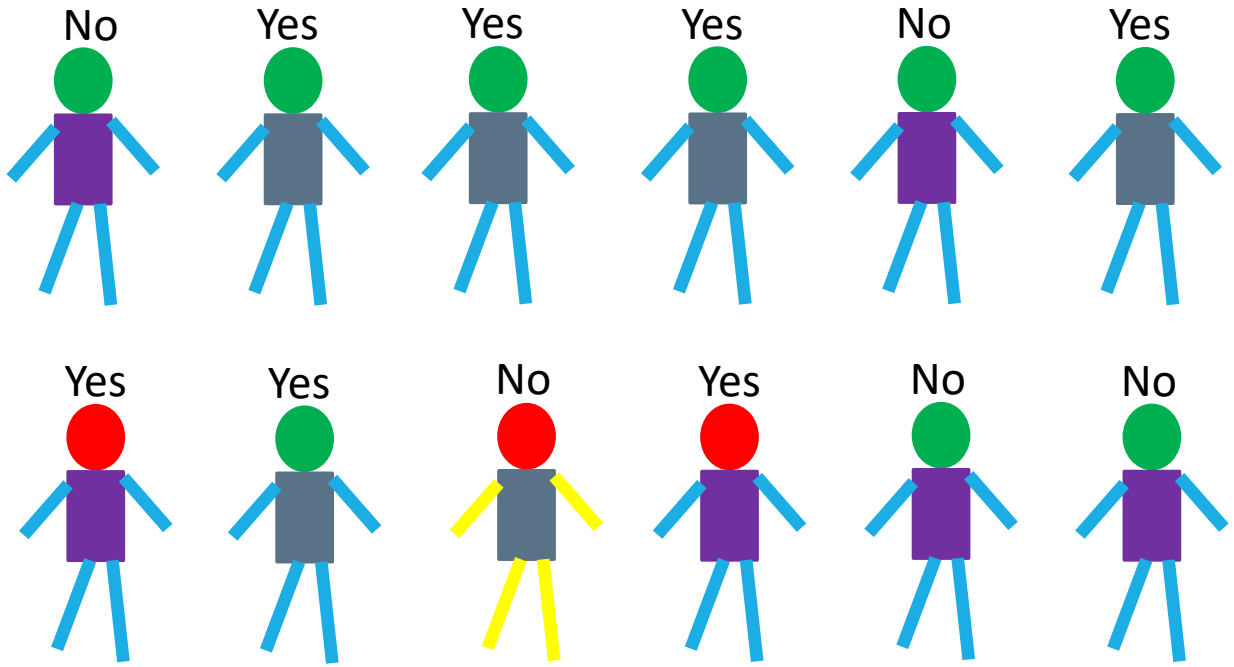
到达叶节点：当到达叶节点时，样本会被分配到某个类别，或者分配一个可能类别的概率分布。

这种方法可以系统地将新样本归类，或评估其属于各类别的概率。



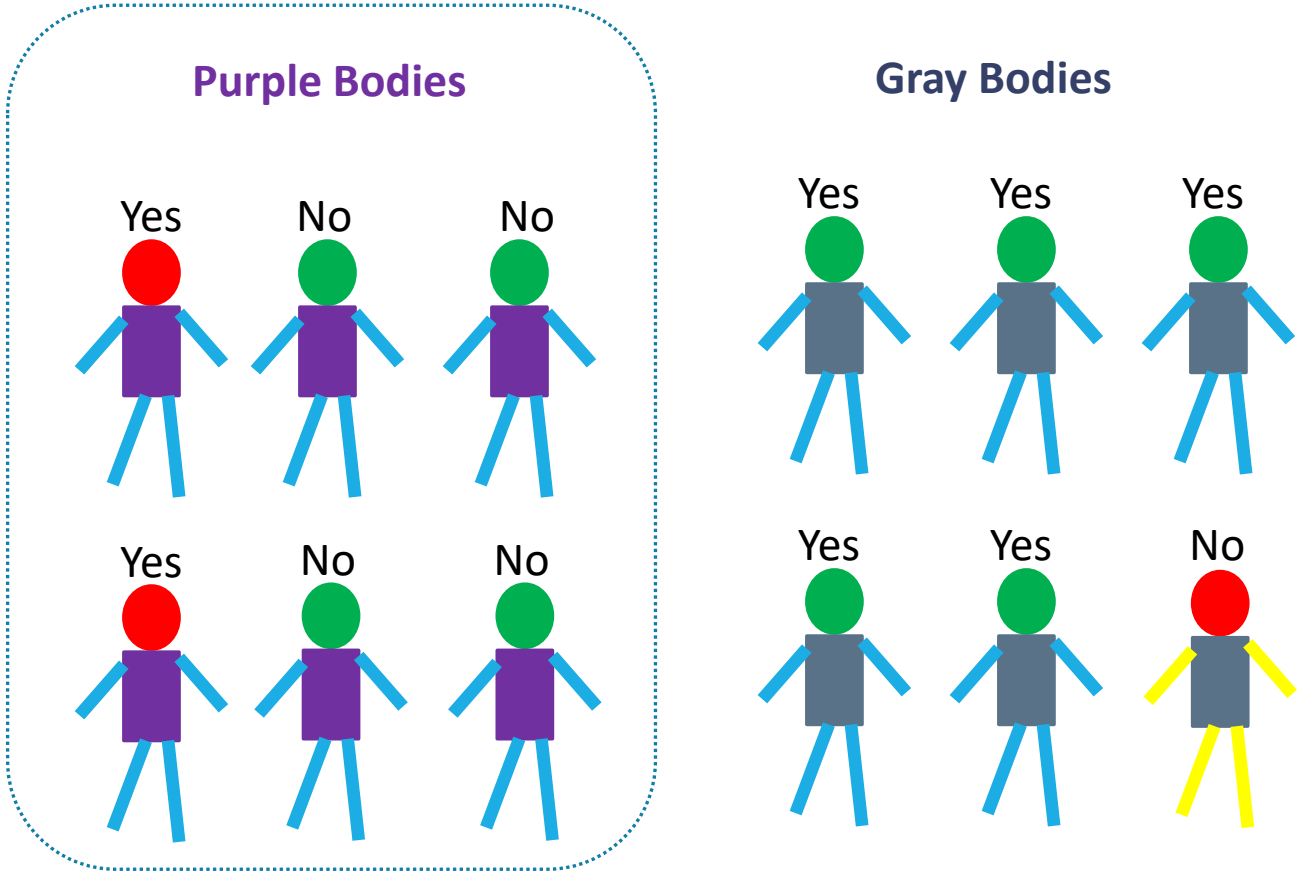
分类树归纳

目标：根据客户的属性，将客户划分为若干子群体，使每个子群体在类别上更加“纯净”（即每个群体中的大多数实例都属于同一类别）。



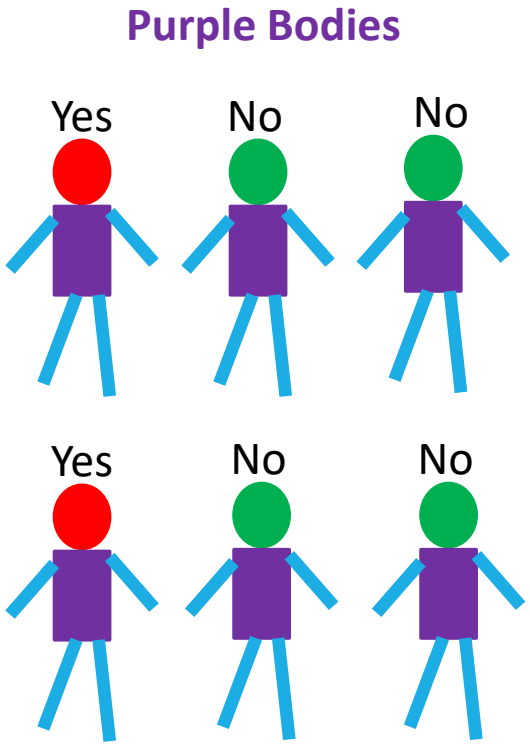
分类树归纳

划分为“更纯净”的群体

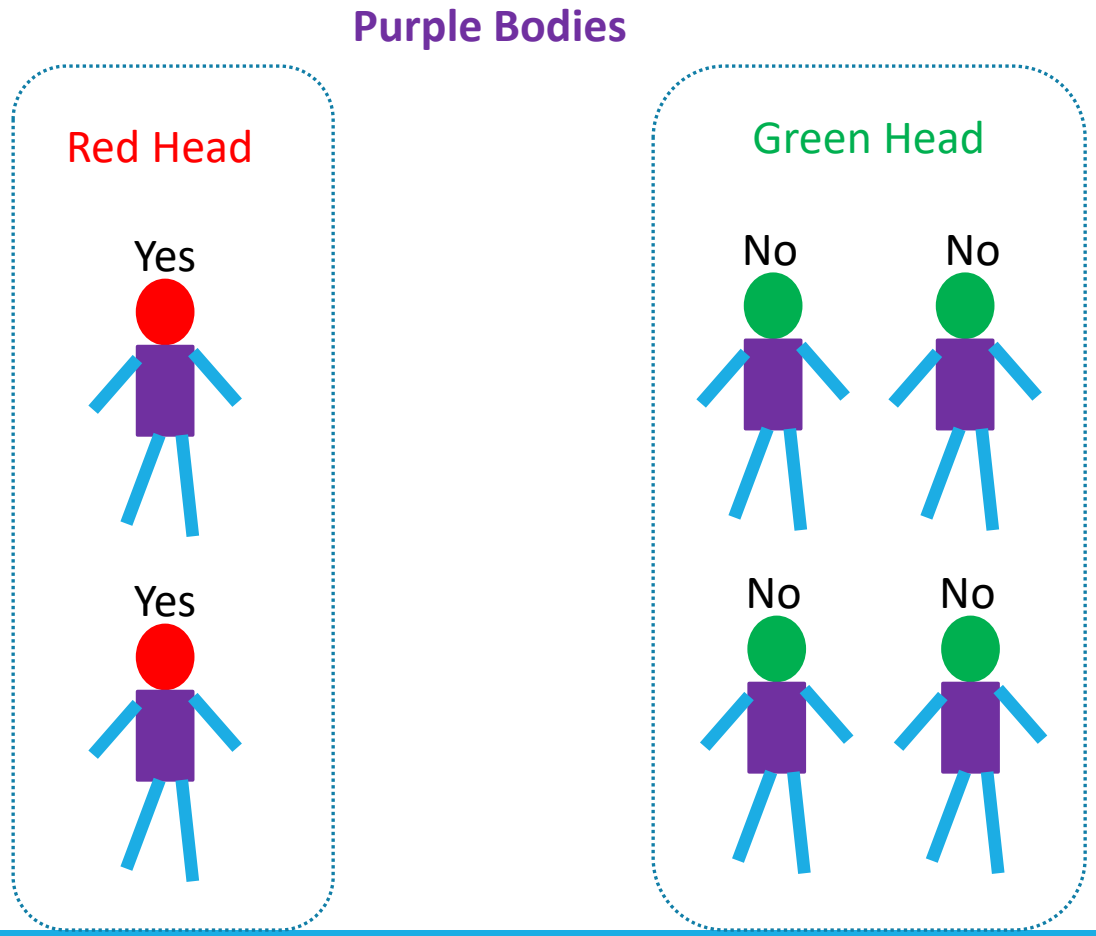


分类树归纳

递归地划分为“更纯净”的群体

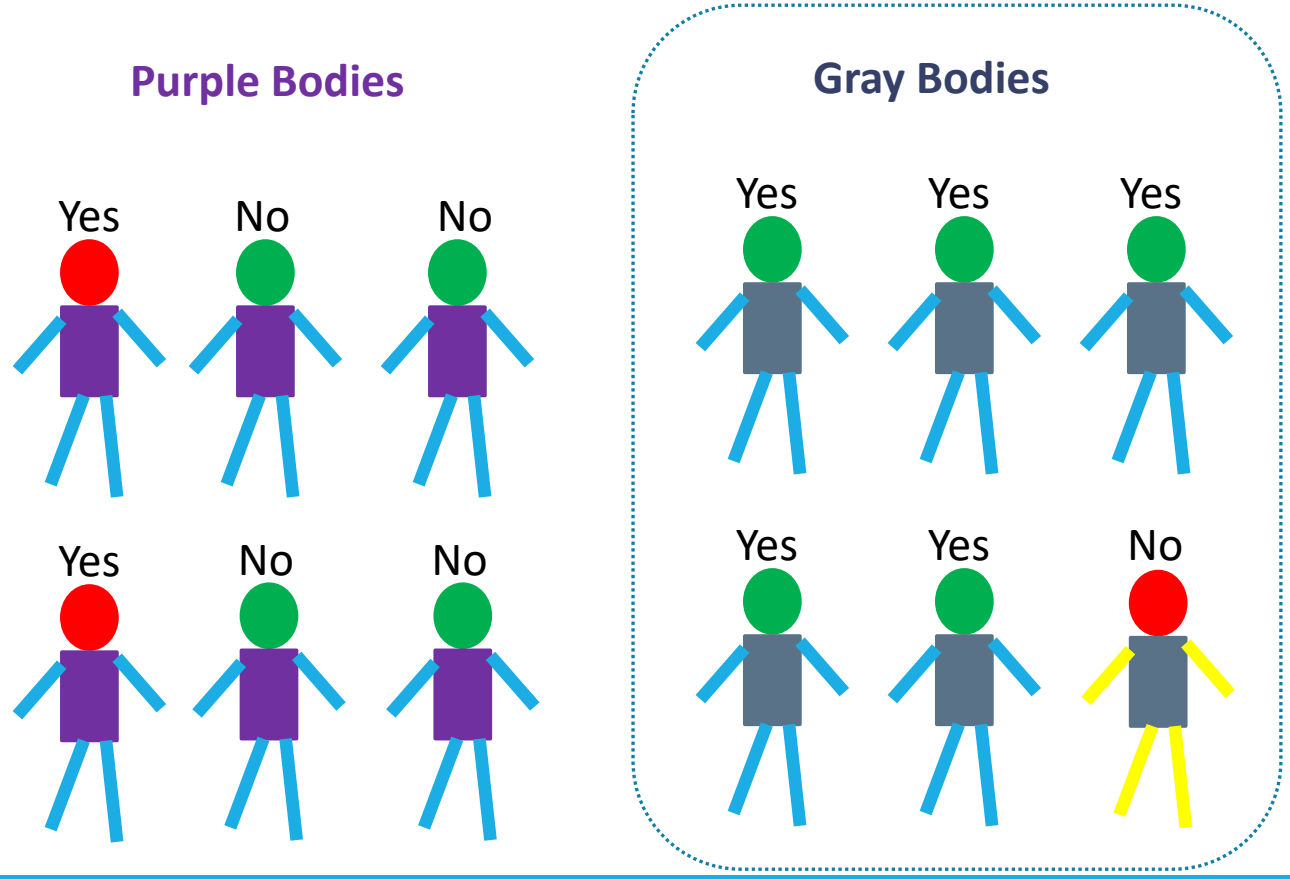


分类树归纳



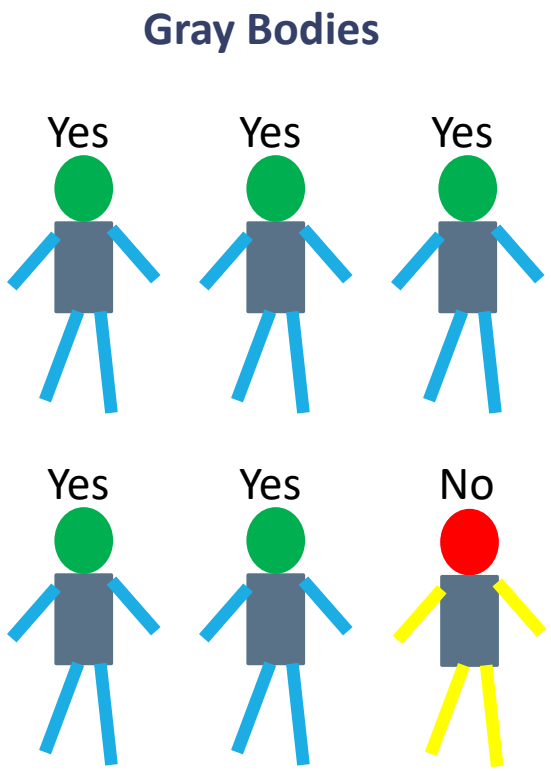
分类树归纳

划分为“更纯净”的群体

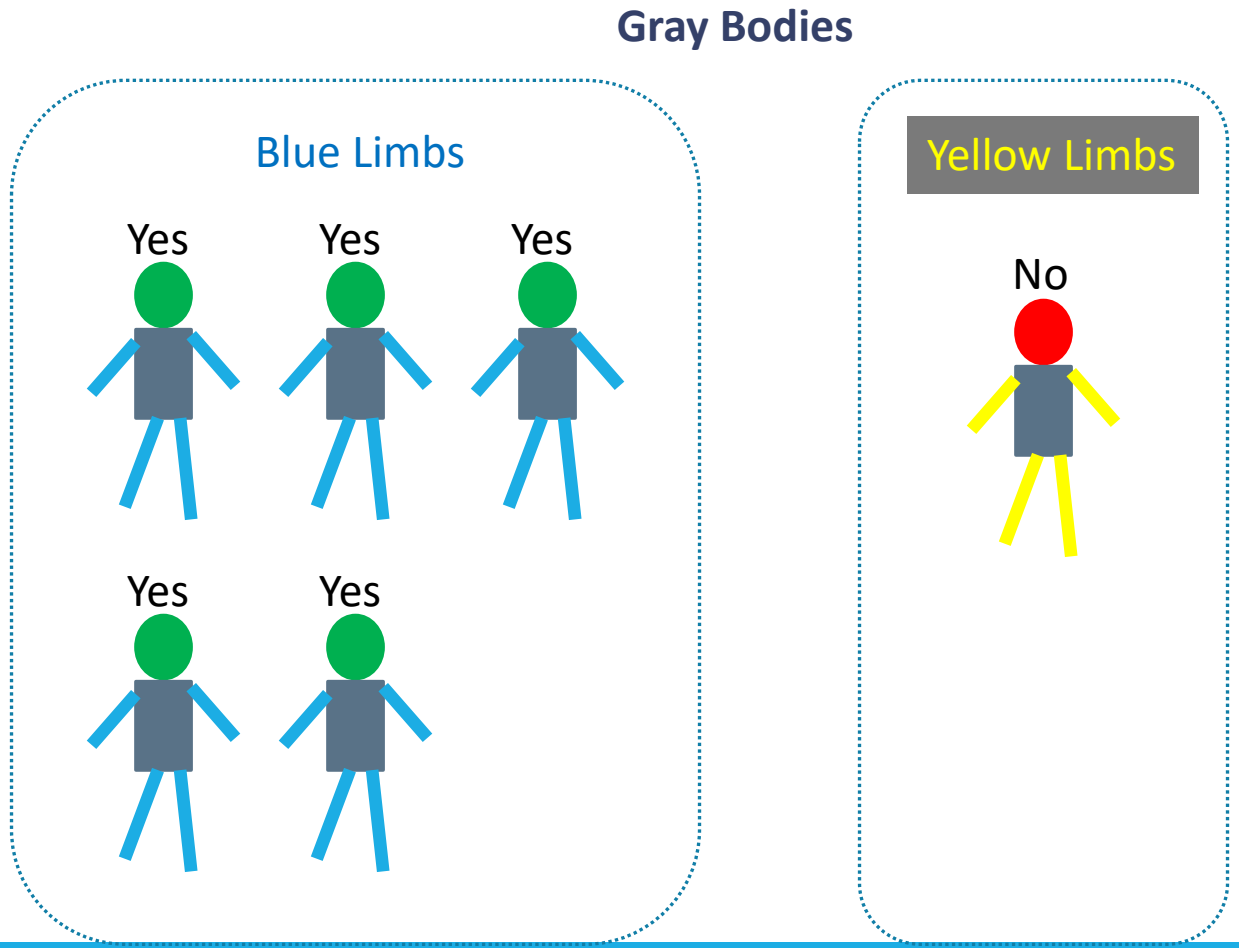


分类树归纳

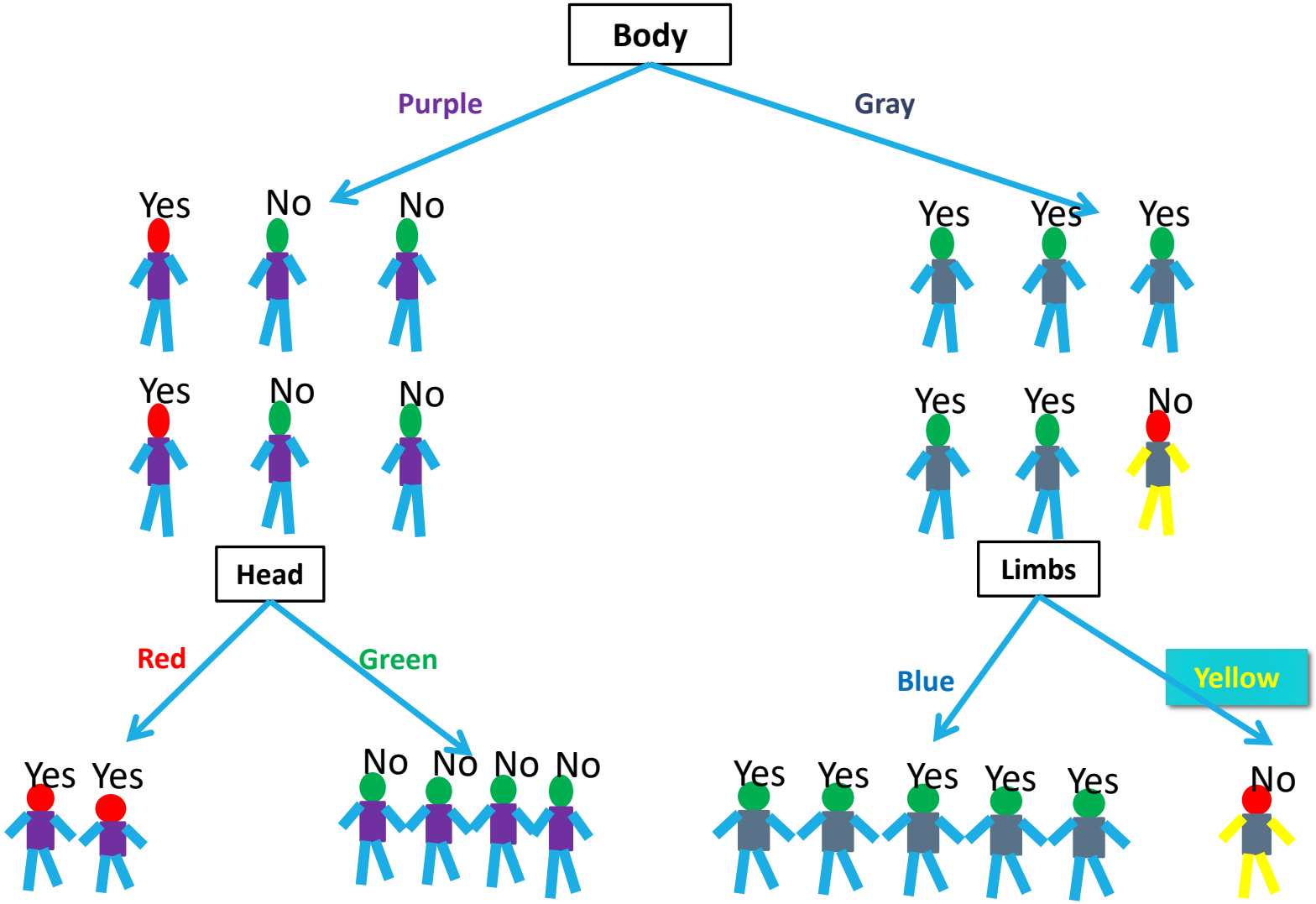
递归地划分为“更纯净”的群体



分类树归纳



分类树归纳


Body
Purple
Gray
Yes
No
No
Yes
Yes
Yes
Yes
No
No
Yes
Yes
No
Head
Limbs
Red
Green
Blue
Yellow
Yes
Yes
No
No
No
No
Yes
Yes
Yes
Yes
Yes
No

分类树归纳

通过递归地划分实例来构建树结构。

每一次划分，实例都会被分成“越来越纯净”的子群体。

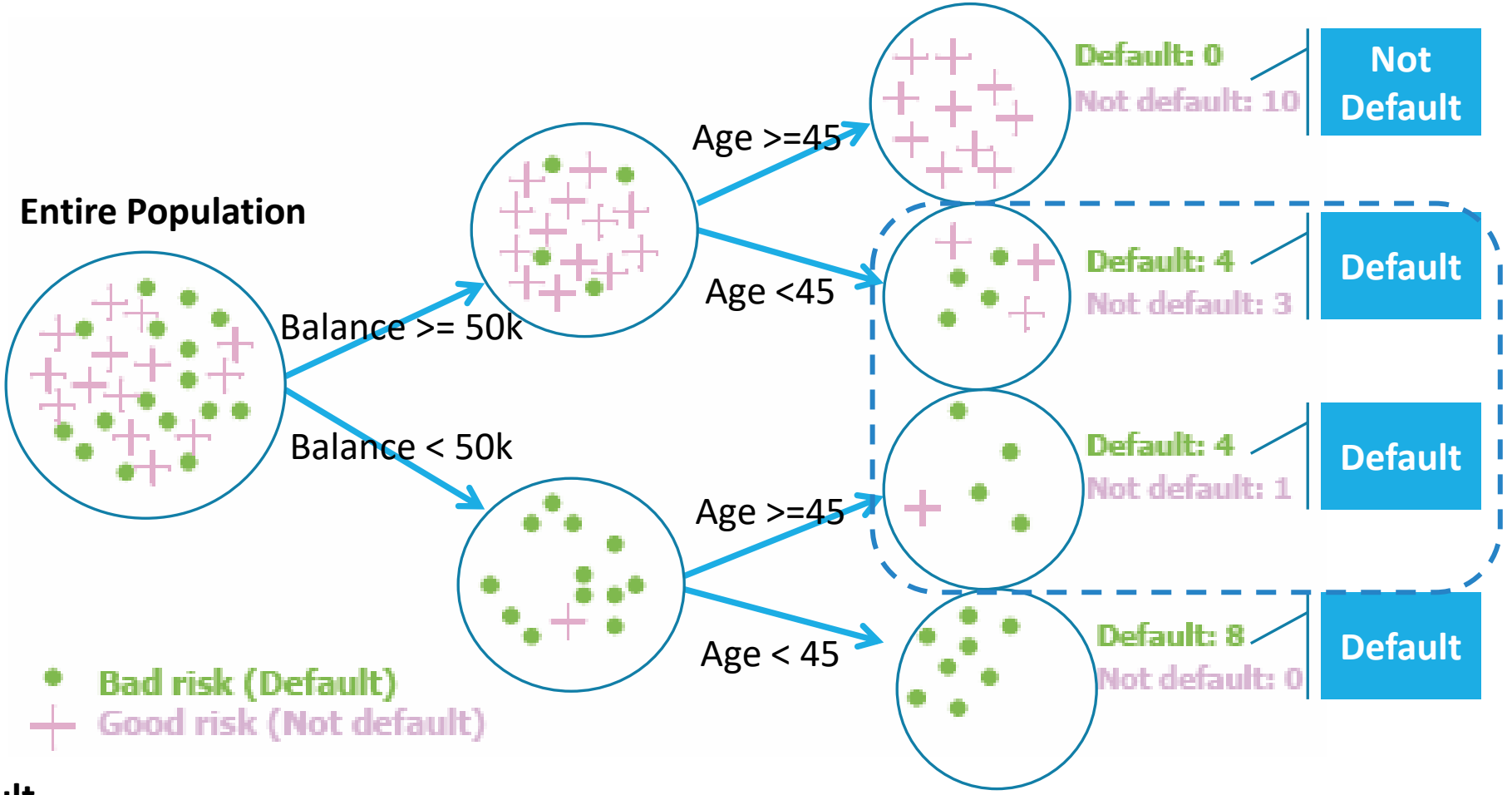
- ***Q1: 根据树归纳如何分配类别概率估计？（例如违约概率）***
- ***Q2: 如何自动选择用于划分数据的属性?***

类别概率估计

基于频率的估计

- *基本假设：对应于树叶节点的每个分段成员属于相应类别的概率相同*
- *如果一个叶节点包含n个正例和m个负例（二分类），则任何新实例为正例的概率估计为* $\frac{n}{n+m}$

练习：违约问题



16: Default

14: Not default

练习

基于你从以往违约数据中学习得到的决策树,

1) 一位新客户，45岁，账户余额为2万，正在申请你公司发行的信用卡.

请预测这位新客户是否会违约？你对你的预测有多大信心？？

2) 另一位女孩也在申请同样的信用卡，但我们唯一掌握的信息是她的账户余额为7万。.

你能预测她是否会违约吗？你对此有多确定？？

如何构建分类树?

MAXIMUM INFORMATION GAIN

构建分类树
分而治之

一棵分类树是通过递归地划分样本实例来构建的。

每次划分时，样本被分成“越来越纯净”的子组。

- ***Q2: 如何自动选择用于划分数据的属性?***

Answer: 在每个节点，选择能够获得最大信息增益的属性进行划分！

基本原则

目标Objectives

- 对于每一个分裂节点，选择能够将总体最有效地划分为更纯净子组的属性。
- 在其他条件相同的情况下，节点越少越好。.

不纯度度量：熵（Entropy)

划分标准：信息增益（基于熵）

- *最常用的方法*
- *衡量属性在区分不同实例时的信息量（即该属性有多大能力将实例区分开）*

不纯度度量——熵（Entropy)

熵Entropy = $\sum_i - p_i \log_2 p_i$

其中 \(p_i \) 是数据中第 \(i \) 类的比例。

熵衡量数据集的“混乱”程度。

来源于信息论（Shannon, 1948）。

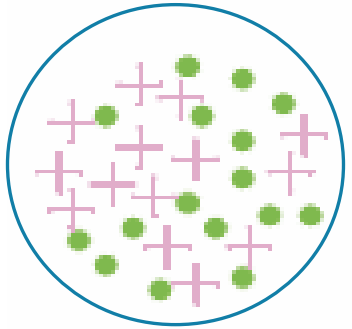
熵的取值范围从 0（最有序，最纯净）到 1（最混乱，最不纯）。

练习Exercise – 熵Entropy

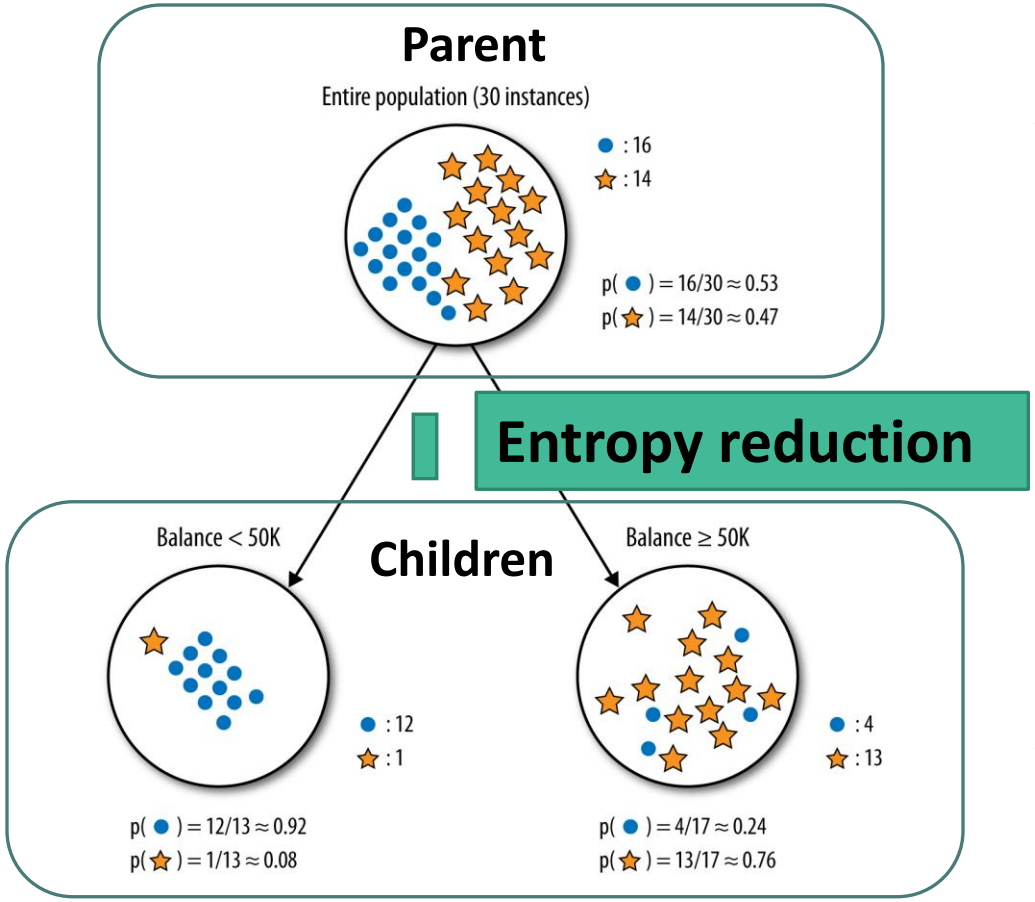
我们的初始样本包含14个“未违约”（Not Default）类别和16个“违约”（Default）类别的案例。

熵增Entropy (***entire population*** of examples) =

$$-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$$



信息增益



信息增益：

指由于引入新信息（如用某个属性进行划分）而导致的熵的变化（减少）。

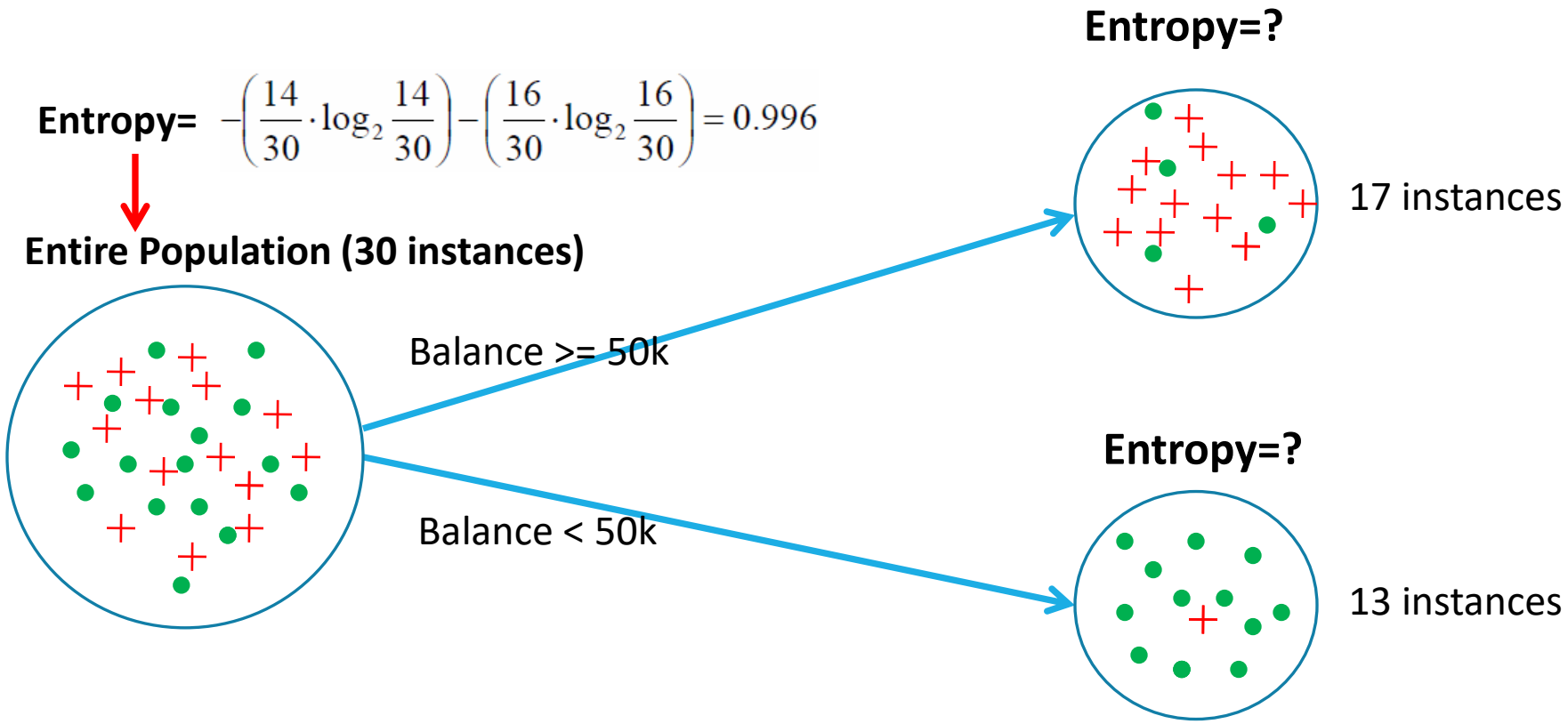
换句话说，信息增益衡量的是通过某个属性对数据集进行划分后，数据集的不确定性（熵）减少了多少。信息增益越大，说明该属性带来的“新信息”越多，划分效果越好。

子集（子节点）中的子组权重:

p(c1); p(c2); …

$$IG(parent, children) = entropy(parent) - [p(c_1) \times entropy(c_1) + p(c_2) \times entropy(c_2) + \cdots]$$

练习：信息增益（Information Gain）



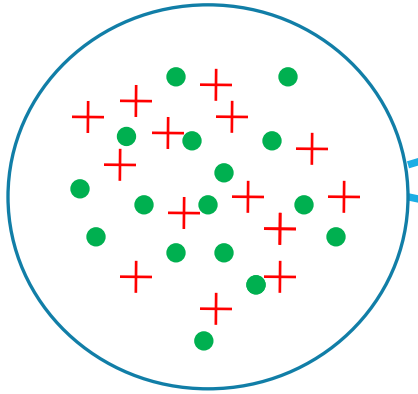
(Weighted) Average Entropy of Children = ?

Information Gain ?

练习：信息增益（Information Gain）

Entropy= $-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$

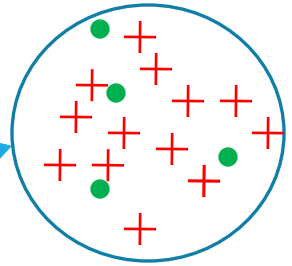
Entire Population (30 instances)



Balance >= 50k

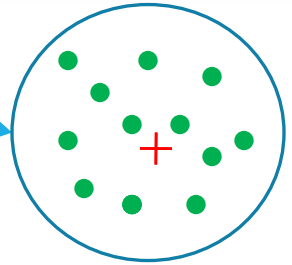
Balance < 50k

Entropy= $-\left(\frac{13}{17}\cdot\log_2\frac{13}{17}\right)-\left(\frac{4}{17}\cdot\log_2\frac{4}{17}\right)=0.787$



17 instances

Entropy= $-\left(\frac{1}{13}\cdot\log_2\frac{1}{13}\right)-\left(\frac{12}{13}\cdot\log_2\frac{12}{13}\right)=0.391$



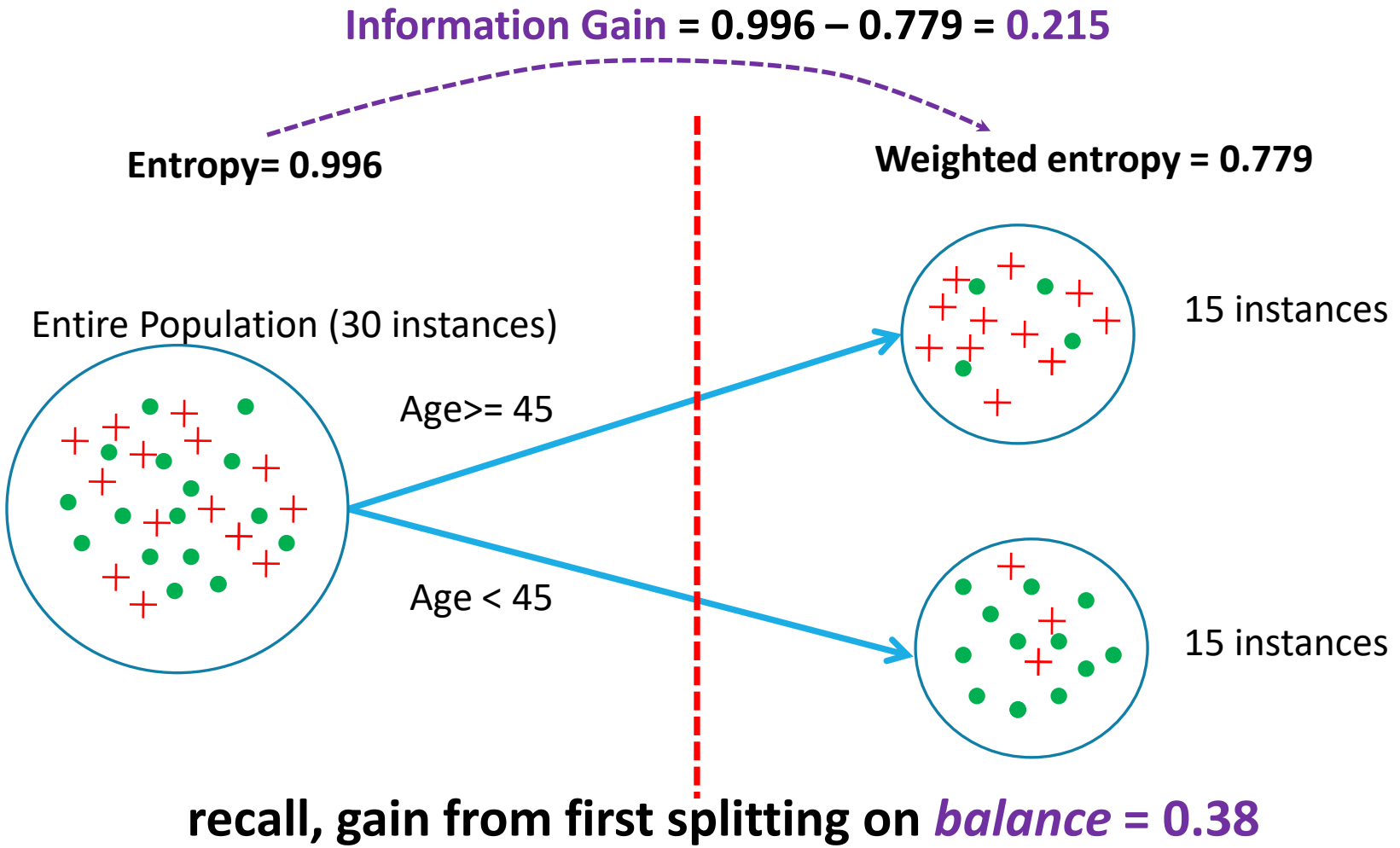
13 instances

(Weighted) Average Entropy of Children = $\left(\frac{17}{30}\cdot 0.787\right)+\left(\frac{13}{30}\cdot 0.391\right)=0.615$

Information Gain = 0.996 - 0.615 = 0.38

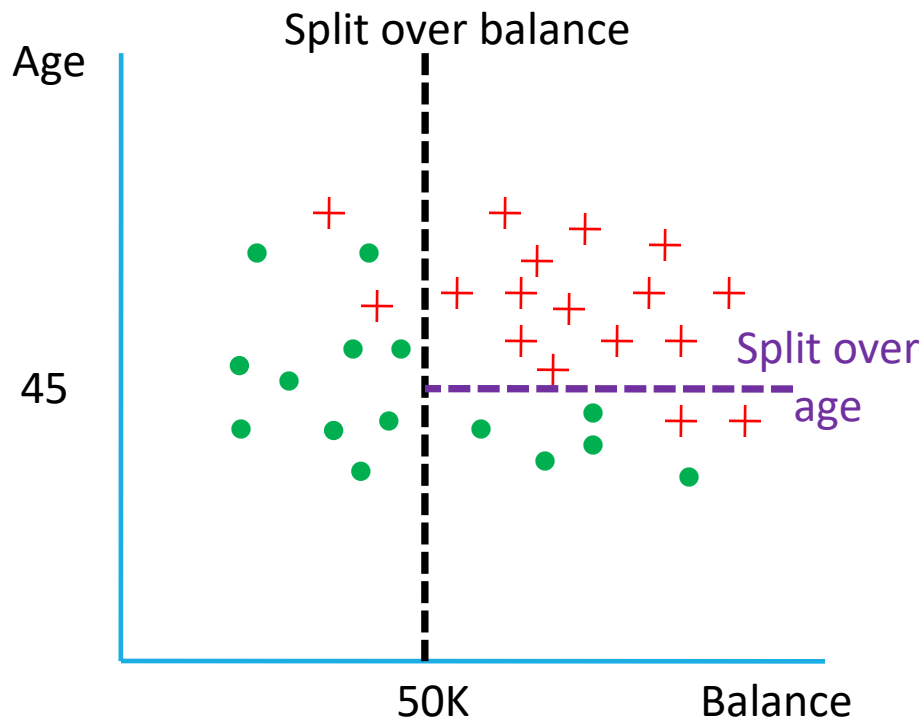
Q2: 如何自动选择用于划分数据的属性?

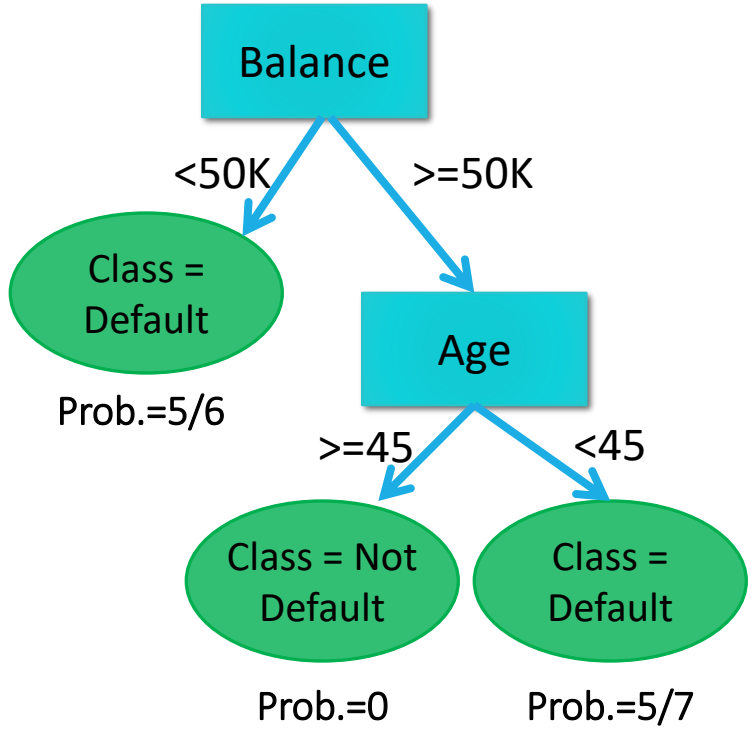
如果我们首先按照“年龄”进行划分，会怎样呢？？



几何解释

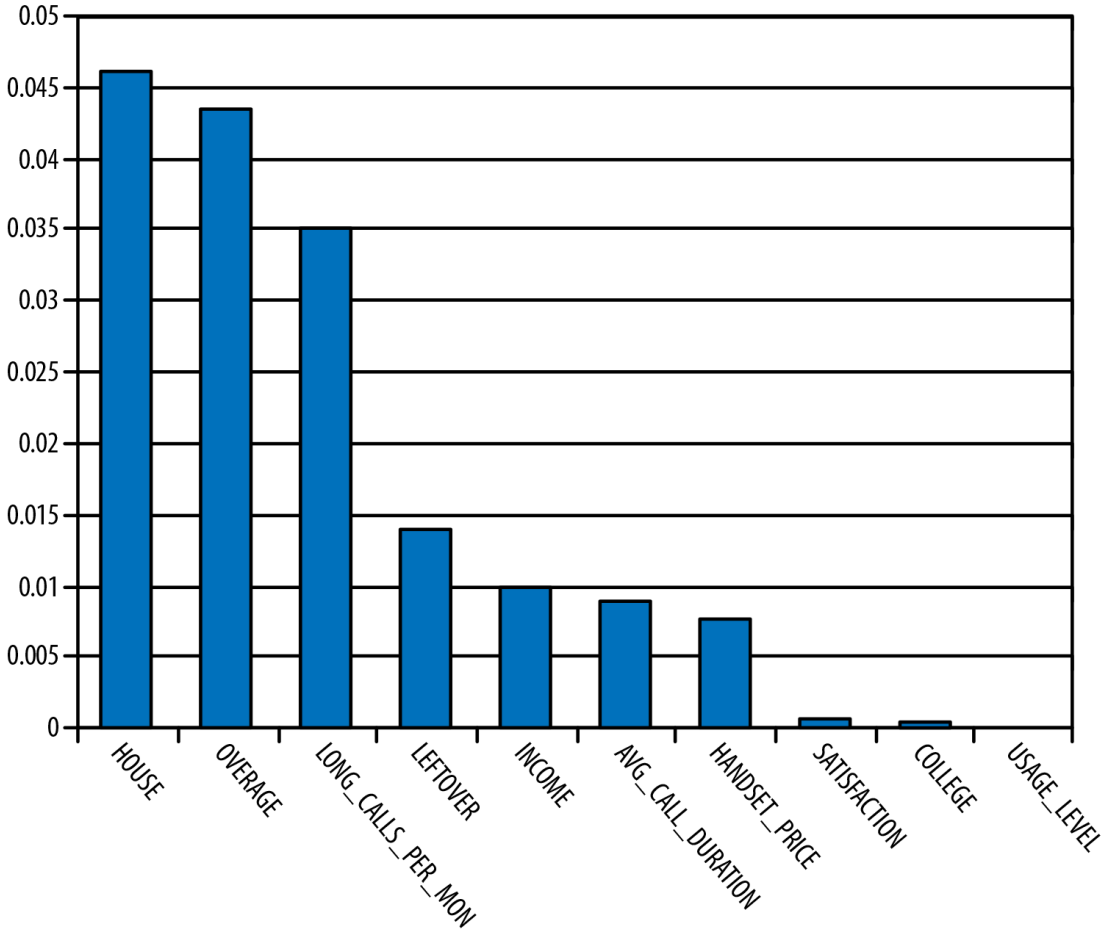
分类树用与坐标轴平行的决策边界对样本空间进行划分。



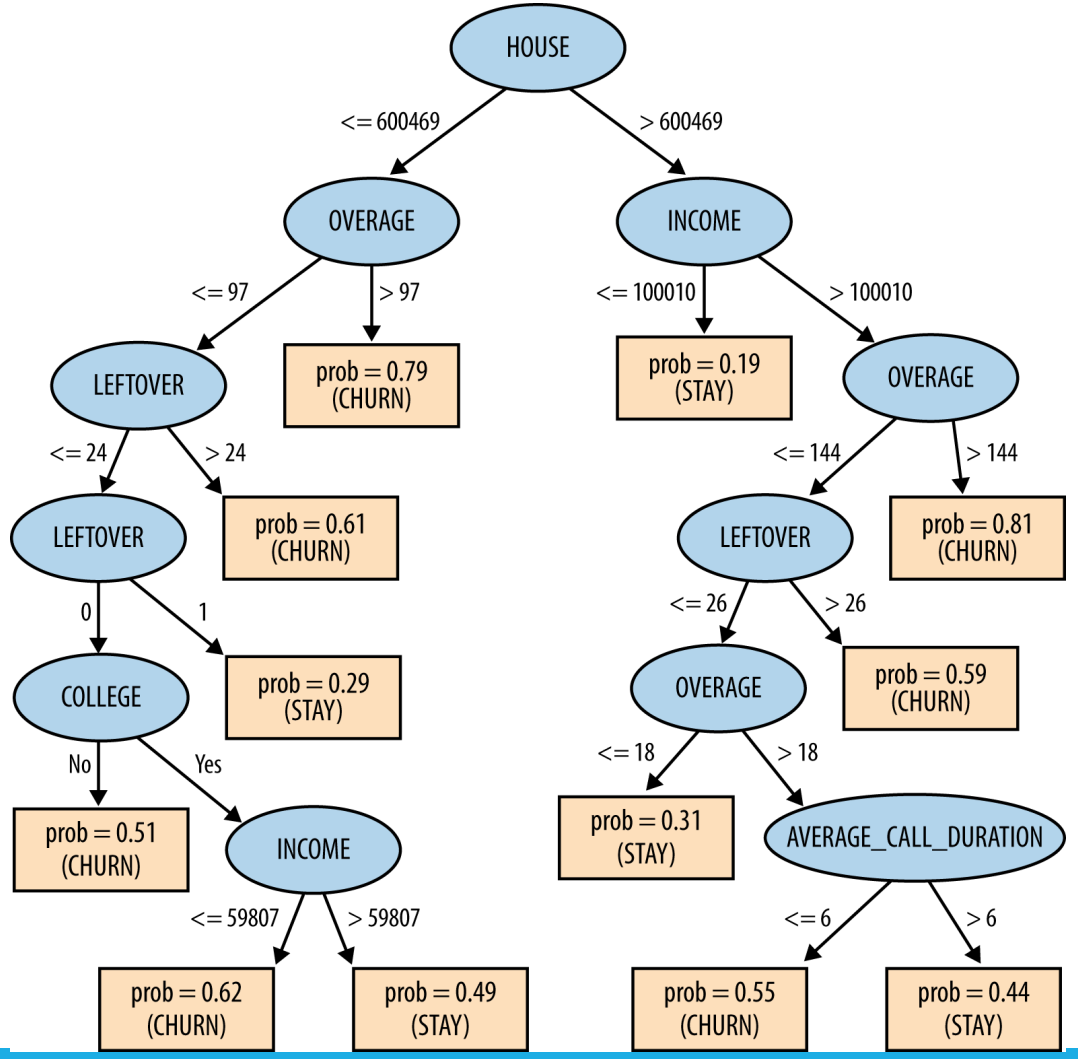


- **Bad risk (Default) – 15 cases**
- **Good risk (Not default) – 17 cases**

TelCo: 用决策树预测客户流失



Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.000	COLLEGE
10	0.000	USAGE_LEVEL



总结 Summary

一棵决策树是通过递归地划分样本构建的。

每一次划分，样本都会被分成“纯度越来越高”的子组。

如何选择用于划分数据的属性？

最大化信息增益！

Reference：
Business Statistics: A First Course, 8th edition, David M. Levine, YorkDavid F. Stephan, TechnologyKathryn A. Szabat, Pearson 2020
Business analytics: The science of data-driven decision making / U. Dinesh Kumar, New Delhi Wiley India, 2022