大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

- 分类的主要算法：

决策树（C4.5、CART等）、 **KNN算法**、SVM算法、贝叶斯算法、BP神经网络等

- 决策树学习是应用最广的归纳推理算法之一

4、经典的机器学习方法

- 4.1 分类算法原理
- 4.2 决策树算法
- 4.3 K-近邻分类算法（KNN算法）
- 4.4 K-均值聚类算法（K-means算法）
- 4.5 Apriori关联规则算法

决策树方法示例

广告册派送示例

将广告投递给那些对广告册感兴趣从而购买自行车的会员

分类模型的作用就是识别出什么样的会员可能购买自行车。

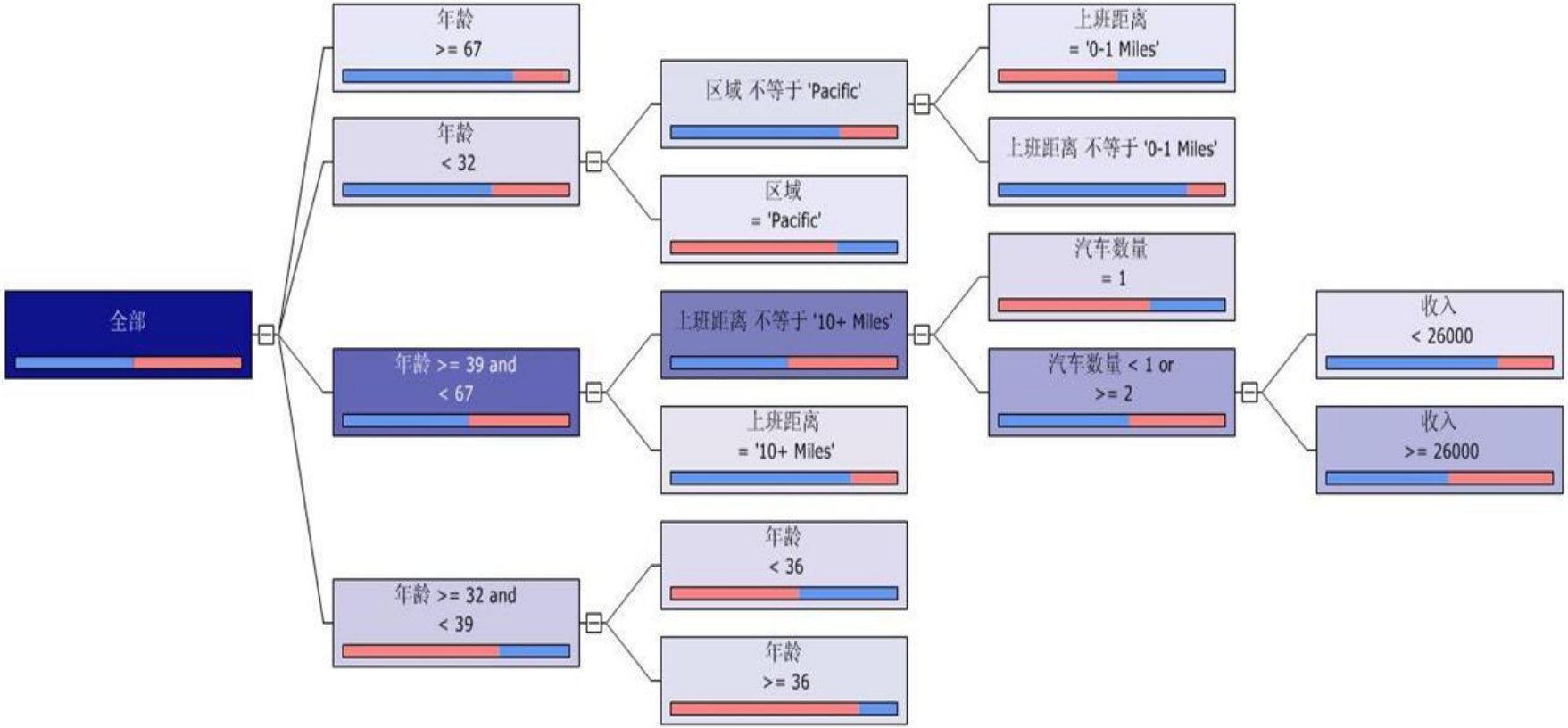
决策树方法示例

自行车厂商首先从所有会员中抽取了1000个会员，向这些会员投递广告册，然后记录这些收到广告的会员是否购买了自行车。
在分类模型中，每个会员作为一个样本事例，居民的婚姻状况、性别、年龄等特征作为输入列，所需预测的分类是客户是否购买了自行车。

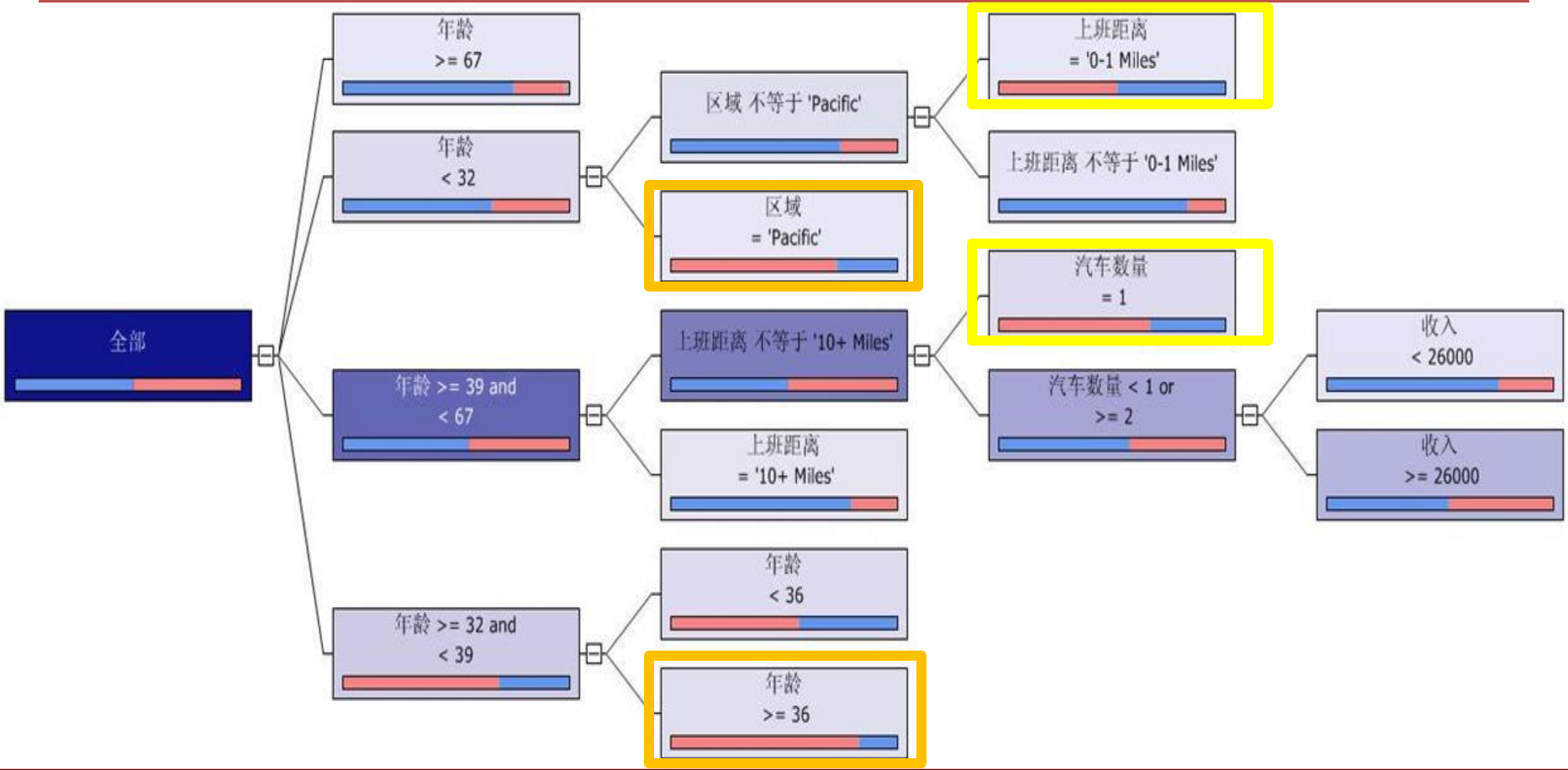
会员编号	婚姻状况	性别	收入	孩子数	教育背景	职业	是否有房	汽车数	上班距离	区域	年龄	是否购买自行车
12496	Married	Female	40000	1	Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42	No
14177	Married	Male	80000	5	Partial College	Professional	No	2	2-5 Miles	Europe	60	No
24381	Single	Male	70000	0	Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41	Yes
25597	Single	Male	30000	0	Bachelors	Clerical	No	0	0-1 Miles	Europe	36	Yes
……												

决策树方法示例

使用1000个会员事例训练模型后得到的决策树分类如下：



决策树方法示例


全部
年龄
>= 67
年龄
< 32
年龄 >= 39 and
< 67
年龄 >= 32 and
< 39
区域 不等于 'Pacific'
区域
= 'Pacific'
上班距离 不等于 '10+ Miles'
上班距离
= '10+ Miles'
年龄
< 36
年龄
>= 36
上班距离
= '0-1 Miles'
上班距离 不等于 '0-1 Miles'
汽车数量
= 1
汽车数量 < 1 or
>= 2
收入
< 26000
收入
>= 26000

决策树方法示例

- 获得的规则-分类模型

1. 年龄小于32岁，居住在太平洋地区的会员有72.75%的概率购买自行车；
2. 年龄在32和39岁之间的会员有68.42%的概率购买自行车；
3. 年龄在39和67岁之间，上班距离不大于10公里，只有1辆汽车的会员有66.08%的概率购买自行车；
4. 年龄小于32岁，不住在太平洋地区，上班距离在1公里范围内的会员有51.92%的概率购买自行车；

在得到了分类模型后，将其他的会员在分类模型中查找就可预测会员购买自行车的概率有多大。

随后自行车厂商就可以有选择性的投递广告册。

决策树表示法

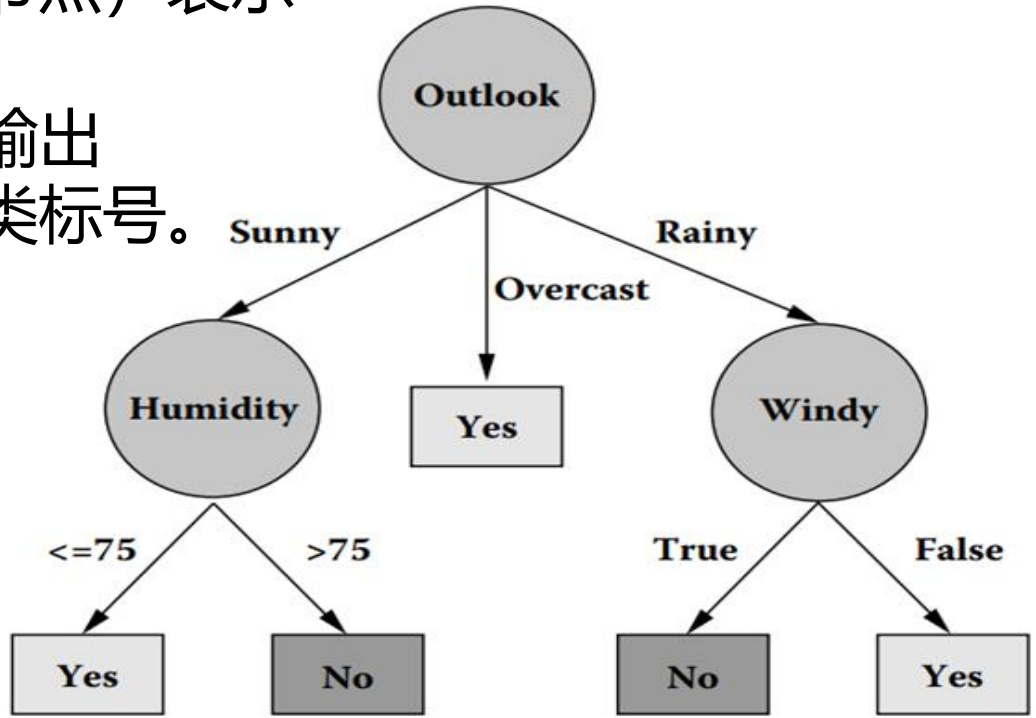
决策树示例：寻找天气情况与是否去打高尔夫球之间的关系

数据集如图所示，它表示的是天气情况与去不去打高尔夫球之间的关系。

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

决策树表示法

- 决策树是一种类似流程图的树结构：
  - 每个内部节点（非树叶节点）表示在一个属性上的测试，
  - 每个分枝代表一个测试输出
  - 每个树叶节点存放一个类标号。



决策树方法的适用问题

- 适用问题的特征
  - 实例由“属性-值”对表示
  - 目标函数具有离散的输出值
  - 可能需要析取的描述
  - 训练数据可以包含错误
  - 训练数据可以包含缺少属性值的实例
- 问题举例
  - 医学中的应用（如根据疾病分类患者、疾病分析与预测）
  - 根据起因分类设备故障（故障诊断）
  - 根据拖欠支付的可能性分类贷款申请
- 分类问题
  - 核心任务是把样例分类到各可能的离散值对应的类别