COMP9414: 人工智能第6a讲。学习

韦恩-沃布克

电邮：w.wobcke@unsw.edu.au

本讲座

- 机器学习
  - 方法学问题
- 监督学习
  - 决策树学习

学习的类型

- 监督学习
  - 代理人被告知输入及其目标输出的例子，必须学习从输入到输出的函数，以达到与训练实例相一致，并对新的实例进行归纳总结
- 强化学习
  代理人没有为每个输入提供目标输出，但会定期获得奖励，并且必须学习如何最大化(预期)回报随着时间的推移
- 无监督学习
  代理人只得到一系列的输入，并且必须在这些输入中找到有用的模式。

监督学习

- 文本分类
  - 贝叶斯分类
- 数据科学与伦理

- 给定一个训练集和一个测试集，每个训练集由一组项目组成，训练集的每个项目都有一组特征和一个目标输出
- 学习者必须学习一个能够预测任何给定项目的目标输出的模型（由其特征集来描述）。
- 学习者被赋予训练集中每个项目的输入特征和目标输出。
  - ⚠ 项目可以一次性提出（批处理）或按顺序提出（在线）。
  - ⚠ 项目可以随机或按时间顺序呈现（流）。
  - 学习者在定义模型时完全不能使用测试集
- 模型的评估是通过预测测试集中每个项目的输出的性能来进行的。

新南威尔士大学
©W.Wobcke等人，2019-2022年

方法与模式

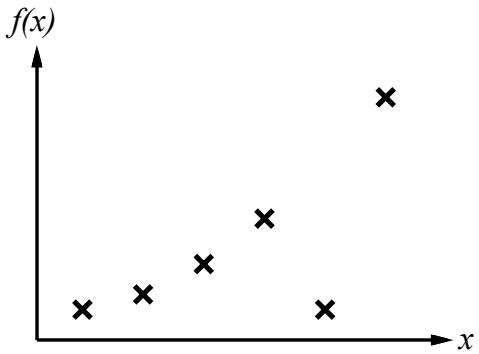
- 可以使用各种学习方法来生成模型
  - 决策树
  - 支持向量机
  - 神经网络/深度学习
- 通过在各种数据集上评价模型来评估方法
  - 标准基准数据集的可用性问题
  - 模型取决于问题的表述和参数
  - 终端用户可能只关心一个模型，而不是一个一般的方法
  - 大多数机器学习研究评估的是方法而不是模型

监督学习 - 方法学

- 特征 "工程" – 选择相关特征
- 选择输入特征和输出的表示方法
- 从原始数据中提取特征的预处理方法
- 选择学习方法来评估
- 选择训练制度（包括参数）。
- 评价
  - 选择现实的基线进行比较
  - 选择内部验证的类型，例如交叉验证。

曲线 拟合

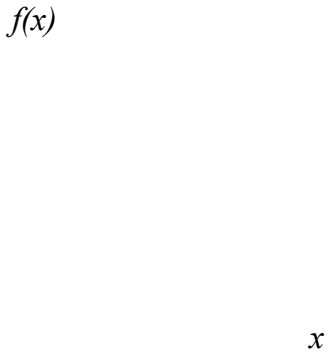
哪条曲线能对这个数据进行 "最佳拟合"？



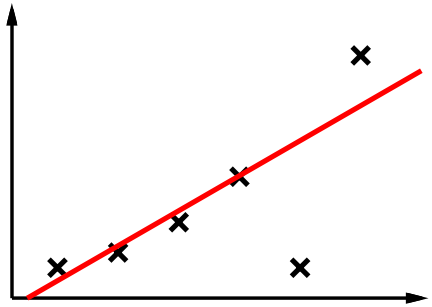
曲线拟合

- 用人类的专业知识和其他基准对结果进行理智检查

哪条曲线能对这个数据进行 "最佳拟合"？



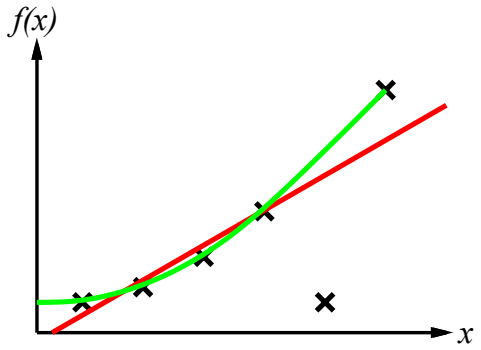
直线？



新南威尔士大学
©W.Wobcke等人，2019-2022年

曲线拟合

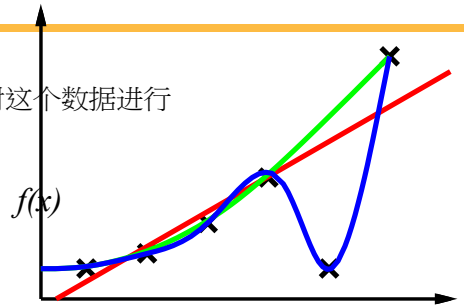
哪条曲线能对这个数据进行
"最佳拟合"？



抛物线？

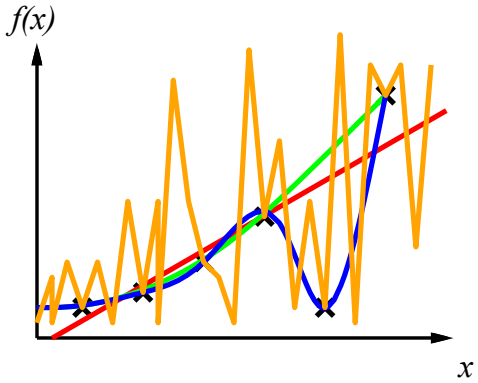
曲线拟合

哪条曲线能对这个数据进行
"最佳拟合"？



曲线拟合

哪条曲线能对这个数据进行 "最佳拟合"？

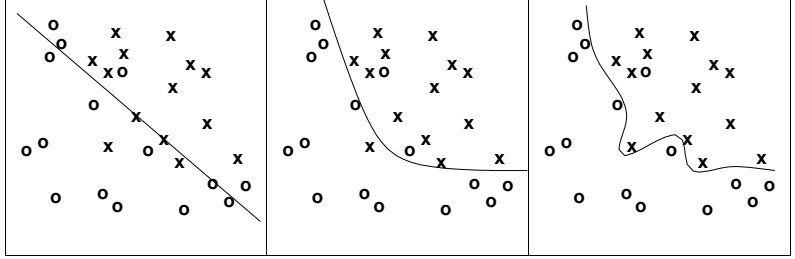


其他的东西？

奥卡姆的剃刀

"最可能的假设是与数据一致的最简单的假设"。

四阶多项式？

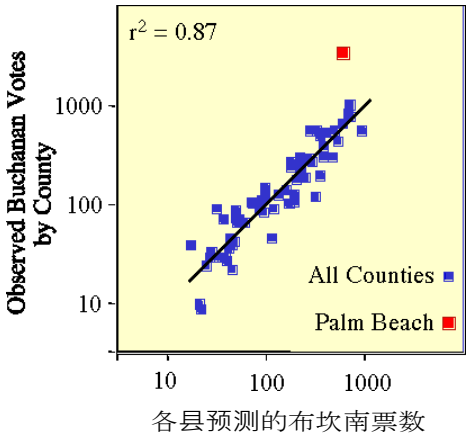


不足的　　良好的妥协
　　过度拟合

由于测量中可能存在噪声，在实践中需要在假设的简单性和它对数据的适合程度之间做出权衡。

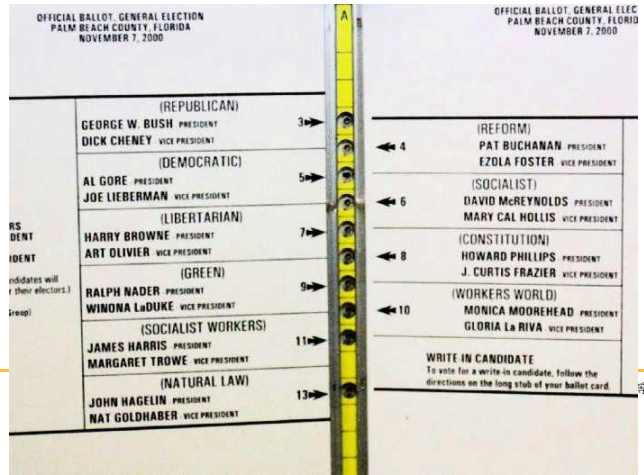
新南威尔士大学
©W.Wobcke等人，2019-2022年

离群索居者



什么时候可以删除离群值？

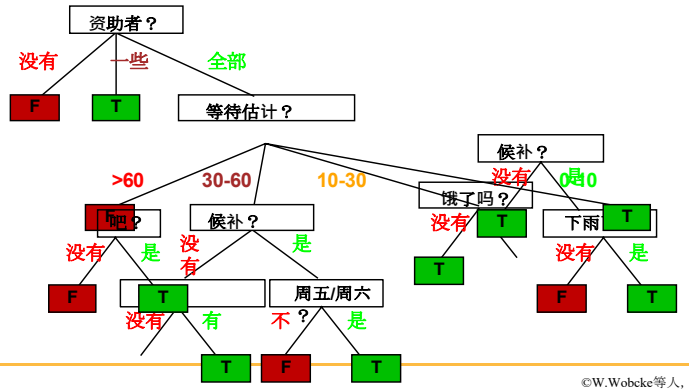
蝴蝶选票



餐厅训练数据

	符号	吧台	F/S	匈奴	裴斯泰洛齐	价格	雨	共和国	类型	遗产	等等？
X_1	T	F	F	T	一些	$$$	F	T	法国	0-10	T
X_2	T	F	F	T	全程	$	F	F	泰国	30-60	F
X_3	F	T	F	F	一些	$	F	F	汉堡	0-10	T
X_4	T	F	T	T	全程	$	F	F	泰国	10-30	T
X_5	T	F	T	F	全程	$$$	F	T	法国	> 60	F
X_6	F	T	F	T	一些	$$	T	T	意大利语	0-10	T
X_7	F	T	F	F	无	$	T	F	汉堡	0-10	F
X_8	F	F	F	T	一些	$$	T	T	泰国	0-10	T
X_9	F	T	T	F	全程	$	T	F	汉堡	> 60	F
X10	T	T	T	T	全程	$$$	F	T	意大利语	10-30	F
X11	F	F	F	F	无	$	F	F	泰国	0-10	F
X12	T	T	T	T	全程	$	F	F	汉堡	30-60	T

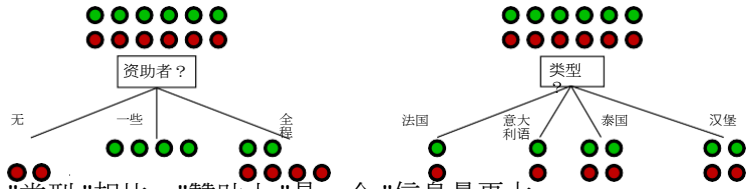
决策树



归纳

- 只要训练集不是不一致的，属性可以按任何顺序分割，以产生一棵能正确分类训练集中所有例子的树。
- 然而，我们需要的是一棵有可能泛化的树--对测试集中（未见过的）例子进行正确分类。
- 考虑到奥卡姆剃刀，更简单的假设是首选--"更简单"=更小的树。
- 如何选择属性以生产小树？

选择一个属性进行分割



与 "类型 "相比，"赞助人 "是一个 "信息量更大 "的属性，因为它将例子更接近于分成 "所有正面 "或 "所有负面 "的集合。

这种 "信息量 "的概念可以用数学的方式来量化。

"熵 "的概念

熵

- 熵是对 "随机性"（缺乏统一性）的一种衡量。
  与某个随机变量的先验分布有关的
  - 更高的熵意味着更多的随机性
  - "信息"（关于分布）降低了熵值
- 想法。基于信息增益的分裂
  - 基于属性的 "交流 "值的熵的损失
  - 与香农的信息理论有关
  - 测量信息增益的比特数

定义。如果n个属性值的先验概率为p_1，- - - ，p_n， 那么分布的熵为

$$H((p_1, - - , p_n)) = \sum_{i=1}^{n} -p_i \log 2\, p_i$$

熵和赫夫曼编码

熵是指通过一个（块）哈夫曼编码方案实现的每个符号的比特数

例1: $H((0.5, 0.5)) = 1$比特

为了用0和1来编码，一个由两个字母A和1组成的长信息频次相同的B，指定A=0，B=1

编码每个字母需要一个比特（二进制数字）。

提醒您。 记 $p = \frac{\log 10\, p}{}$或 $\frac{\log e\, p}{}$

录

通过最小化每一步的熵，可以建立一个解析树

$\log_{10} 2 \log_e 2$

新南威
尔士大
学
©W.Wobcke等人，2019-
2022年

熵和赫夫曼编码

例2：H ((0.5, 0.25, 0.25)) = 1.5比特

要对一个由字母A、B和C组成的信息进行编码，其中B和C出现的频率相同，但A出现的频率是其他两个字母的两倍，请指定A=0，B=10，C=11。

编码每个字母所需的平均比特数为1.5

如果这些字母以其他比例出现，它们需要被 "阻断 "在

为了有效地对它们进行编码
最有效的编码方案所需的平均比特数

是由熵给出的

$$H((p_1, \ldots, p_n)) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

熵

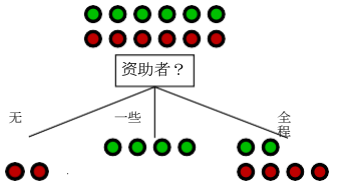
假设在一个节点上有p个正的和n个负的例子------。
→ $H((p/(p+n)), n/(p+n))$分类一个新例子所需的比特 -
对于12个餐厅的例子，$p = n = 6$，所以需要1比特。

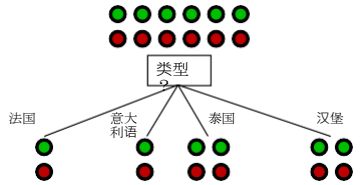
一个属性将实例E分割成子集E_i

，其中每个子集需要较少的信息来完成分类（减少熵）。

让E_i有p_i正面和n_i负面的例子 -
→ $H((p_i/(p_i + n_i)), n_i/(p_i + n_i))$
对一个新例子进行分类所需的位数

信息获取



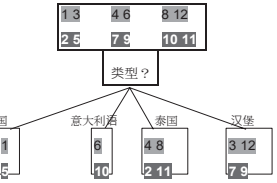


对于赞助人来说，熵值 $\frac{1}{6}(0) + \frac{1}{3}(0) + \frac{1}{2}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right] = \frac{1}{2}(0.459) = 0.459$

对于类型，熵 $= -\left(\frac{1}{6}(1) + \frac{1}{3}(1) + \frac{1}{3}(1)\right) = 1$

选择下一个属性

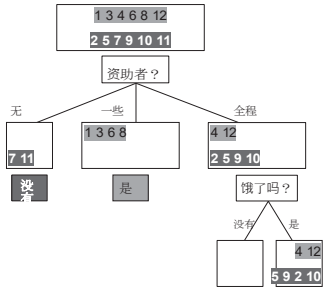
→ 所有分支的每个例子的预期比特数是



(a)

$$\sum_i \frac{p_i + n_i}{p + n} H((\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}))$$

对顾客来说，这是0.459比特，对类型来说，这（仍然）是1比特 ∴对顾客的分割

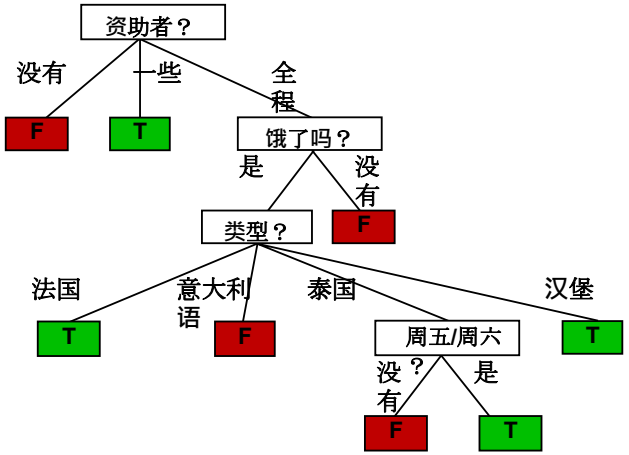


(b)

在对食客进行分割后，将节点Patrons=Full分割到Hungry上。

新南威尔士大学
©W.Wobcke等人，2019-2022年

诱导的决策树



拉普拉斯误差和修剪

按照奥卡姆剃刀，修剪对分类项目没有太大好处的分支（帮助归纳，避免过度拟合）。

对于一个叶子节点，所有的项目都将分配给该节点的大多数类别。使用拉普拉斯误差估计（未见过的）测试项目的错误率

$$E = 1 - \frac{n+1}{N+k}$$

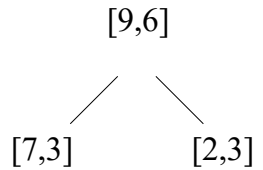
N = 节点上的（训练）项目总数
n = 多数类中的（训练）项目数

最小的错误修剪

这个节点的子女是否应该被修剪？左边的孩子有

班级频率[7,3]。

$$E = \frac{n+1}{N+k} = \frac{7+1}{10+2} = \frac{1}{0.333} =$$



右边的孩子有$E = 0.429$
父节点有$E = 0.412$
左边和右边孩子的平均数是

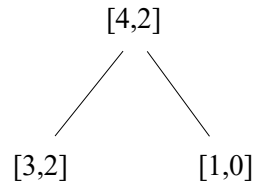
$$E = \frac{10}{15}(0.333) + \frac{5}{15}(0.429) = 0.365$$

因为$0.365 < 0.412$，所以不应该修剪儿童。

最小的错误修剪

这个节点的孩子是否应该被修剪？左边的孩子有班级频率[3,2]。

$$E = 1 - \frac{n+1}{N+k} = 1 - \frac{3+1}{5+2} = 0.429$$



右边的孩子有$E = 0.333$
父节点有$\bar{E} = 0.375$
左边和右边孩子的平均数是

$k =$ 班级的数量

如果子节点的平均拉普拉斯误差超过了父节点的误差。

去掉孩子

$$E = \frac{5}{6}(0.429) + \frac{1}{6}(0.333) = 0.413$$

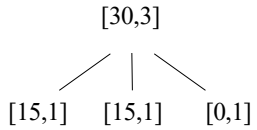
由于0.413>0.375，儿童应该被修剪掉。

新南威尔士大学
©W.Wobcke等人，2019-2022年

最小的错误修剪

这个节点的子节点是否应该被修剪？左边和中间的孩子有类的频率[15,1]。

$$E = 1 - \frac{n+1}{N+k} = 1 - \frac{15+1}{16+2} = 0.111$$



右边的孩子有$E = 0.333$

父节点有$E = \frac{4}{35} = 0.114$

左边、中间和右边孩子的平均数是

$$E = \frac{16}{33}(0.111) + \frac{16}{33}(0.111) + \frac{1}{33}(0.333) = 0.118$$

由于0.118>0.114，儿童应该被修剪掉。

摘要

- 监督学习
  - 训练集和测试集
  - 根据输入特征预测目标值
- 奥卡姆的剃刀
  - 简单性和准确性之间的权衡
- 决策树
  - 通过建立一个较小的树来进行概括（使用熵）。
  - 根据拉普拉斯误差修剪节点
  - 修剪决策树的其他方法