

树学习

高 阳，李文斌

<http://cs.nju.edu.cn/rl>

2023年09月26日

大纲

符号学习

变型空间

归纳偏置

ID3决策树算法

其他树算法

大纲

符号学习

变型空间

归纳偏置

ID3决策树算法

其他树算法

推理的角度

演绎（正向）推理： 已知 $P \rightarrow Q$ ，P为真，则Q为真

规则：南京大学软件学院的学生都是天之骄子；（大前提）

情况：小明来自南京大学软件学院；（小前提）

结果：所以，小明是天之骄子。（结论）

推理的角度

演绎（正向）推理：已知 $P \rightarrow Q$ ，P为真，则Q为真

反绎（溯因/反向）推理：已知 $P \rightarrow Q$ ，Q为真，则P为真

规则：南京大学软件学院的学生都是天之骄子；

情况：小明是天之骄子；

结果：所以，小明来自南京大学软件学院。

推理的角度

演绎（正向）推理：已知 $P \rightarrow Q$ ， P 为真，则 Q 为真

反绎（溯因/反向）推理：已知 $P \rightarrow Q$ ， Q 为真，则 P 为真

归纳推理：已知前件为真，后件未必为真

情况：小明来自南京大学软件学院；

结果：小明是天之骄子；

规则：南京大学软件学院的学生都是天之骄子。

推理的角度

演绎（正向）推理：已知 $P \rightarrow Q$ ， P 为真，则 Q 为真

反绎（溯因/反向）推理：已知 $P \rightarrow Q$ ， Q 为真，则 P 为真

归纳推理：已知前件为真，后件未必为真

符号(概念)学习是一类归纳推理

概念学习 (Concept Learning)

定义：给定样例集合，以及每个样例是否属于某个概念，自动地推断出该概念的一般定义。

样例	天气	空气温度	湿度	风速	水温	预测	享受运动
Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes



样例数

概念学习 (Concept Learning)

定义：给定样例集合，以及每个样例是否属于某个概念，自动地推断出该概念的一般定义。

样例	天气	空气温度	湿度	风速	水温	预测	享受运动
Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes



属性1:
可取值 + Cloudy

概念学习 (Concept Learning)

定义：给定样例集合，以及每个样例是否属于某个概念，自动地推断出该概念的一般定义。

样例	天气	空气温度	湿度	风速	水温	预测	享受运动
Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes



目标概念

概念学习任务

实例集合X：上例中用六个属性表示

目标概念c：定义在实例集上的布尔函数 $c:X \rightarrow \{0,1\}$

训练样例：正例($c(x)=1$)，反例($c(x)=0$)

假设集H：每个假设h表示X上定义的布尔函数 $h:X \rightarrow \{0,1\}$

概念学习：寻找一个假设h，使对于X中的所有x，

$$h(x)=c(x)$$

实例空间和假设数

最一般的假设： $\langle ?, ?, ?, ?, ?, ? \rangle$

最特殊的假设： $\langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle$

实例空间： $3*2*2*2*2*2=96$

假设空间： $5*4*4*4*4*4=5120$ (语法不同)

$1+4*3*3*3*3*3=973$ (语义不同)

□ 归纳学习假设

任一假设如果在足够大的训练样例集合中能很好地逼近目标概念函数，它也能在未见实例中很好地逼近目标概念

作为搜索的概念学习

- 当假设的表示确定后，也就确定了概念学习算法所有假设的空间
- 搜索的目标是为了寻找最好地拟合训练样例的假设
- 搜索(泛化)的操作
 - ✓ 用逻辑变量替换常量
 - ✓ 合取表达式去掉部分条件
 - ✓ 对表达式增加析取项
 - ✓ 用属性的超类来替换属性

假设的一般到特殊序

$h_1 = \langle \text{Sunny}, ?, ?, \text{Strong}, ?, ? \rangle$

$h_2 = \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle$



h_2 包含的实例数多于 h_1

□ 更泛化 (more general than or equal to)

令 h_j 和 h_k 是定义在 X 上的布尔函数, 若 $h_j \geq_g h_k$

当且仅当, $(\forall x \in X) [(h_k(x)=1) \rightarrow (h_j(x)=1)]$

□ 严格泛化 $h_j >_g h_k$

□ 更特化 $h_j \leq_s h_k$

Specific

$h = \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle$

$h = \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle$

$h = \langle \text{?, Warm, ?, Strong, ?, ?} \rangle$

$h = \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle$

$h = \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle$

$h = \langle \text{?, Warm, ?, ?, ?, ?} \rangle$

Generalize

Find-S:寻找极大特殊假设

1. 将h初始化为H中最特殊的假设

2. 对每个正例x

- 对h的每个属性约束 a_i

如果x满足 a_i

那么不做任何处理

否则将h中 a_i 替换为x满足的另一个最一般的约束

3. 输出假设h

$\langle S, W, N, S, W, S \rangle Y$

$\langle S, W, H, S, W, S \rangle Y$

$\langle R, C, H, S, W, S \rangle N$

$\langle S, W, H, S, C, C \rangle Y$

$h = \langle \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset \rangle \longrightarrow h = \langle S, W, N, S, W, S \rangle$

$\longrightarrow h = \langle S, W, ?, S, W, S \rangle \longrightarrow h = \langle S, W, ?, S, ?, ? \rangle$

Find-S:算法特点

- 对以属性合取式表示的假设空间，输出与正例一致的最特殊的假设

□ 思考

- ✓ 找到了正确的目标概念了吗？
- ✓ 为什么用最特殊的假设？
- ✓ 训练样例是否一致？
- ✓ 如果存在多个极大特殊假设，如何处理？

列表消除算法: List-Then-Eliminate

1. 变型空间 Version Space \leftarrow 假设空间 \mathcal{H} 中所有假设的列表
 2. 对每个样例 $\langle x, c(x) \rangle$
 - 从变型空间中移除 $h(x) \neq c(x)$ 的假设 h
 3. 输出 Version Space 中的假设列表
-



要列出所有假设，在实际中往往不可能

变型空间

□ 一致 (Consistent)

一个假设 h 与训练样例集合 D 一致

当且仅当, $Consistent(h, D) \equiv (\forall \langle x, c(x) \rangle \in D) \boxed{h(x)=c(x)}$

包含反例



□ 变型 (版本) 空间 (version space)

关于假设空间 \mathcal{H} 和训练样例集合 D 的变型空间, 是 \mathcal{H} 中与训练样例 D 一致的所有假设构成的子集

$$VS_{\mathcal{H}, D} \equiv \{h \in \mathcal{H} \mid Consistent(h, D)\}$$

注: 反例对于超泛化的抑制作用

变型空间

□ 极大泛化 (maximally general)

\mathcal{H} 中与训练样例集合 D 一致的极大一般成员的集合

$$G \equiv \{g \in \mathcal{H} \mid \text{Consistent}(g, D) \wedge (\neg \exists g' \in \mathcal{H} [(g' >_g g) \wedge \text{Consistent}(g', D)])\}$$

□ 极大特化 (maximally specific)

\mathcal{H} 中与训练样例集合 D 一致的极大特殊成员的集合

$$S \equiv \{s \in \mathcal{H} \mid \text{Consistent}(s, D) \wedge (\neg \exists s' \in \mathcal{H} [(s >_s s') \wedge \text{Consistent}(s', D)])\}$$

变型空间表示定理

□ 表示定理

令 X 为任意的实例集合， \mathcal{H} 为 X 上定义的布尔函数集合。令 $c: X \rightarrow [0,1]$ 为 X 上定义的任一目标概念，并令 D 为任意训练样例的集合 $\{\langle x, c(x) \rangle\}$ 。对所有的 X, \mathcal{H}, c, D 以及良好定义的 S 和 G ：

$$VS_{H,D} \equiv \{h \in \mathcal{H} \mid (\exists s \in S)(\exists g \in G)[g >_g h >_s s]\}$$

G : 极大泛化集合

S : 极大特化集合

正例和反例的作用

正例用于S泛化，搜索S集合

反例用于G特化，缩小G集合

候选消除算法: Candidate-Eliminate

将G集合初始化为H中最一般假设


$$G_0 = \{ \langle ? , ? , ? , ? , ? , ? \rangle \}$$

将S集合初始化为H中最特殊假设


$$S_0 = \{ \langle \emptyset , \emptyset , \emptyset , \emptyset , \emptyset , \emptyset \rangle \}$$

对每个训练样例d,

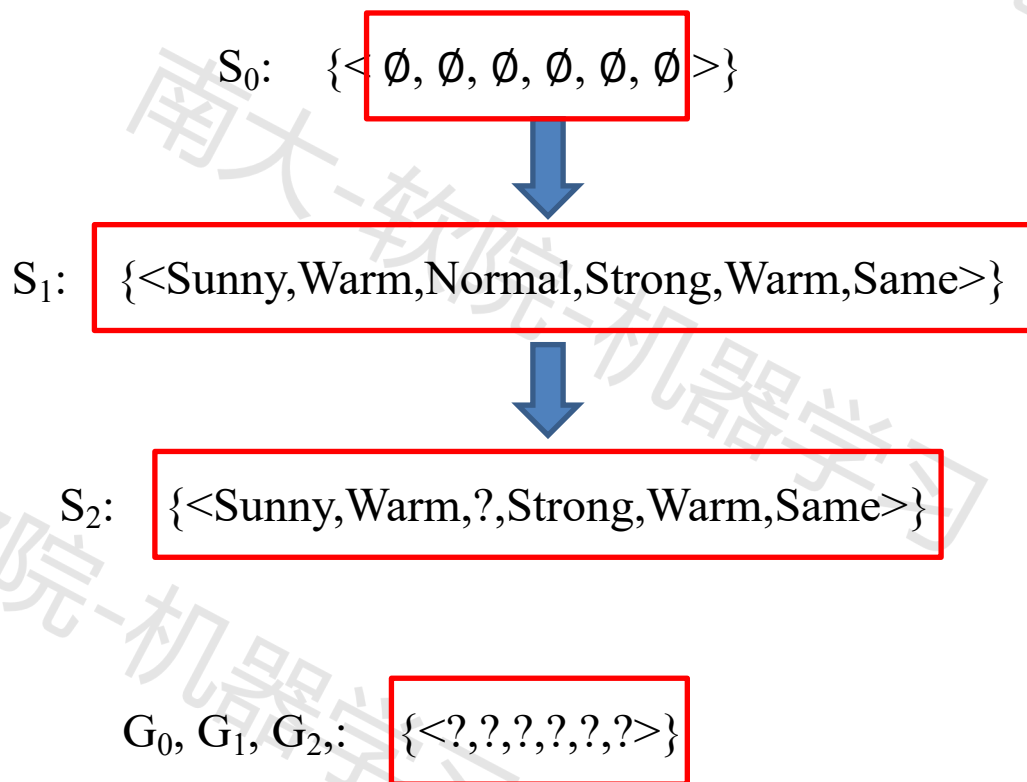
如果d是正例

- 从G中移去所有和d不一致的假设
 - 对S中每一个与d不一致的假设s
 - 从S中移除s
 - 把s的所有极小泛化假设h加入到S中
 - 且h满足与D一致, 而且G中的某个成员比h更一般
 - 从S中移去所有这样的假设:
它比S中另一假设更一般
- 

如果d是反例

- 从S中移去所有和d不一致的假设
 - 对G中每一个与d不一致的假设g
 - 从G中移除g
 - 把g的所有极小特化假设h加入到G中
 - 且h满足与D一致, 而且S中的某个成员比h更特殊
 - 从G中移去所有这样的假设:
它比G中另一假设更特殊
- 

候选消除算法-步骤1



训练样例:

1. $\langle \text{Sunny}, \text{Warm}, \text{Normal}, \text{Strong}, \text{Warm}, \text{Same} \rangle$, Enjoy Sport=Yes
2. $\langle \text{Sunny}, \text{Warm}, \text{High}, \text{Strong}, \text{Warm}, \text{Same} \rangle$, Enjoy Sport=Yes

$\langle S, W, N, S, W, S \rangle Y$
 $\langle S, W, H, S, W, S \rangle Y$
 $\langle R, C, H, S, W, S \rangle N$
 $\langle S, W, H, S, C, C \rangle Y$

候选消除算法-步骤2

S_2, S_3 : {<Sunny, Warm, ?, Strong, Warm, Same>}

G_3 : {<Sunny, ?, ?, ?, ?, ?> <?, Warm, ?, ?, ?, ?> <?, ?, ?, ?, ?, Same>}

G_2 : {?, ?, ?, ?, ?, ?}

训练样例:

3. <Rain, Cold, High, Strong, Warm, Change>, Enjoy Sport=No

<S, W, N, S, W, S> Y

<S, W, H, S, W, S> Y

<R, C, H, S, W, S> N

<S, W, H, S, C, C> Y

候选消除算法-步骤3

$S_3: \{ \langle \text{Sunny, Warm, ?, Strong, Warm, Same} \rangle \}$



$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$

$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$



$G_3: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \langle \text{?, ?, ?, ?, ?, Same} \rangle \}$

训练样例:

4. $\langle \text{Sunny, Warm, High, Strong, Cool, Change} \rangle$, Enjoy Sport=Yes

$\langle \text{S, W, N, S, W, S} \rangle \text{ Y}$
 $\langle \text{S, W, H, S, W, S} \rangle \text{ Y}$
 $\langle \text{R, C, H, S, W, S} \rangle \text{ N}$
 $\langle \text{S, W, H, S, C, C} \rangle \text{ Y}$

候选消除算法-步骤4

$S_4: \{ \langle \text{Sunny}, \text{Warm}, ?, \text{Strong}, ?, ? \rangle \}$



$\{ \langle \text{Sunny}, ?, ?, \text{Strong}, ?, ? \rangle \langle \text{Sunny}, \text{Warm}, ?, ?, ?, ? \rangle \langle ?, \text{Warm}, ?, \text{Strong}, ?, ? \rangle \}$



$G_4: \{ \langle \text{Sunny}, ?, ?, ?, ?, ? \rangle \langle ?, \text{Warm}, ?, ?, ?, ? \rangle \}$

讨论

□ 能收敛到正确的目标概念假设吗？

- ✓ 训练样例集没有错误
- ✓ H 中确实包含描述目标概念的正确假设（请思考....）

□ 需要什么样的训练样例？

- ✓ 例：<Sunny, Warm, Normal, Light, Warm, Same>
- ✓ 新样例将变型空间缩小一半 $\rightarrow \log_2|VS|$

□ 未完全学习的概念的应用？

- ✓ 预测分类
- ✓ 投票分类
- ✓ 确定最优的查询

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	Yes
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	No
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	Yes
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	?

$S_4: \{ \langle \text{Sunny, Warm, ?, Strong, ?, ?} \rangle \}$



$\{ \langle \text{Sunny, ?, ?, Strong, ?, ?} \rangle \langle \text{Sunny, Warm, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$



$G_4: \{ \langle \text{Sunny, ?, ?, ?, ?, ?} \rangle \langle \text{?, Warm, ?, ?, ?, ?} \rangle \}$

Instance	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	?
2	Rainy	Cold	Normal	Light	Warm	Same	?
3	Sunny	Warm	Normal	Light	Warm	Same	?
4	Sunny	Cold	Normal	Strong	Warm	Same	No

大纲

符号学习

变型空间

归纳偏置

ID3决策树算法

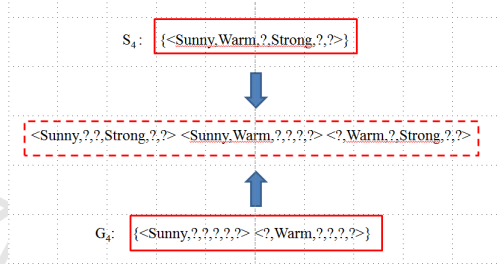
其他树算法

归纳偏置

- 目标概念假设不在假设空间怎么办？
- 能设计包含所有假设的空间吗？
- 假设空间大小对未见实例的泛化能力有什么影响？
- 假设空间大小对所需训练样例数量有什么影响？

以上问题是归纳推理/机器学习的根本问题！

新的样例



Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Cool	Change	Yes
2	Cloudy	Warm	Normal	Strong	Cool	Change	Yes
3	Rainy	Warm	Normal	Strong	Cool	Change	No

$S_5: \{ \langle \text{?, Warm, ?, Strong, ?, ?} \rangle \}$

问题在于：原假设空间是由合取式(有偏)表示，
而真实空间是由析取式表示！

构造无偏的学习器

□ 幂集

- ✓ 集合 X 所有子集的集合
- ✓ 前例: $|X|=96$, $2^{|X|}=2^{96}\approx 10^{28}$
- ✓ 有偏: $973 \ll 10^{28}$

□ 无偏的表示

- ✓ 上例: $\langle \text{Sunny}, ?, ?, ?, ?, ? \rangle \vee \langle \text{Cloudy}, ?, ?, ?, ?, ? \rangle$

□ 无偏学习的泛化

- ✓ 给定3个正例 x_1, x_2, x_3 , 2个反例 x_4, x_5
- ✓ $S: \{x_1 \vee x_2 \vee x_3\}$, $G: \{\neg(x_4 \vee x_5)\}$

无偏学习的无用性

□ 无偏学习的泛化

- ✓ 给定3个正例 x_1, x_2, x_3 , 2个反例 x_4, x_5
- ✓ $S: \{x_1 \vee x_2 \vee x_3\}$, $G: \{\neg(x_4 \vee x_5)\}$

□ 目标概念的学习

- ✓ 必须提供 X 中的所有实例作为训练样例
- ✓ 无法进行泛化
- ✓ 变型空间和候选消除算法失效

因此，归纳学习必须给定某种形式的预先假定(归纳偏置)！

归纳偏置(inductive bias)

□ 核心

- ✓ 学习器从训练样例中泛化并推断新实例分类过程中所采用的策略

□ 精确定义

□ 给定任意训练数据 $D_c = \{x, c(x)\}$, 目标概念 c , 学习算法 L

□ 推断新实例 x_i

□ 则归纳推理过程为:

$$(D_c \wedge x_i) \rightarrow L(x_i, D_c)$$



无法判定

$$(B \wedge D_c \wedge x_i) \vdash L(x_i, D_c)$$

学习器的归纳偏置为附加的前提集合 B ,通过 B ,则归纳推理可由演绎推理派生!

不同的归纳偏置

□ 有偏程度不同的三种归纳学习算法

- ✓ 机械式学习器(Rote-Learner)
- ✓ 候选消除算法
- ✓ FIND-S

□ 有偏性

- ✓ 无归纳偏置
- ✓ $\{c \in H\}$
- ✓ $\{c \in H\} + \text{“任何实例，除非可由其他先验推出，否则为反例”}$

有偏性越强，则学习器的归纳能力越强！

南大-软院-机器学习

南大-软院-机器学习

如何学习具有析取表示的假设空间呢？

南大-软院-机器学习

大纲

符号学习

变型空间

归纳偏置

ID3决策树算法

其他树算法

决策树学习

□ 决策树学习

- ✓ 实例：“属性-值”对表示，应用最广的归纳推理算法之一
- ✓ 目标函数具有离散的输出值
- ✓ 很好的健壮性(样例可以包含错误，也可以处理缺少属性值的实例)
- ✓ 能够学习析取表达式

□ 算法

- ✓ ID3, Assistant, C4.5
- ✓ 搜索一个完整表示的假设空间，表示为多个If-then规则

□ 归纳偏置

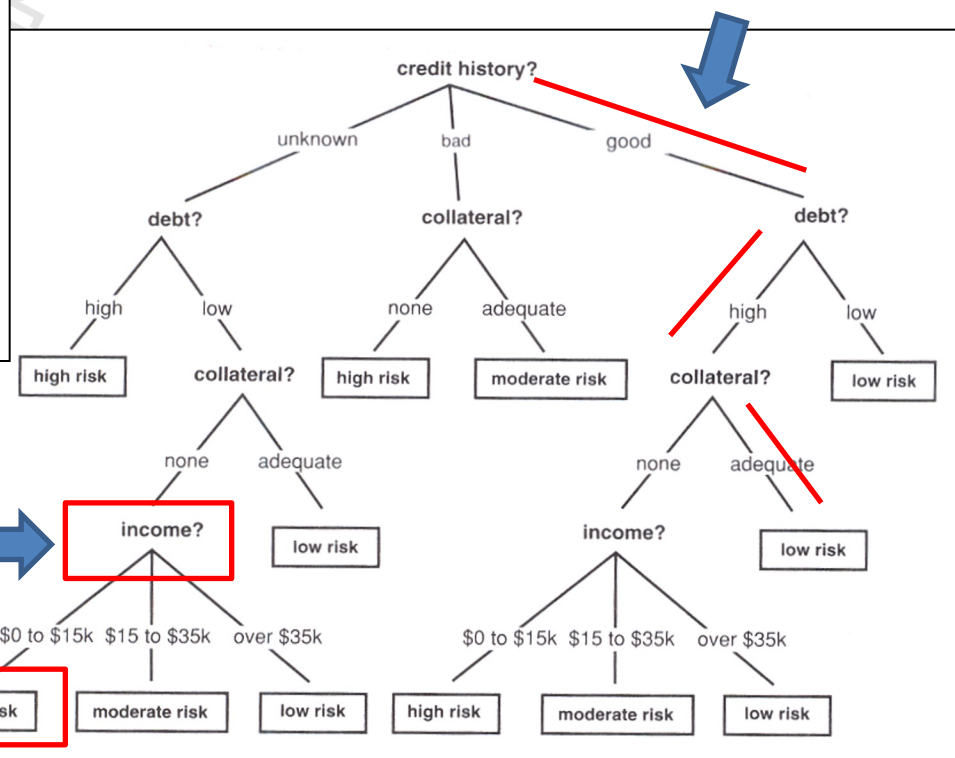
- ✓ 优先选择较小的树

风险 信用历史 债务 抵押物 收入

RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
high	bad	high	none	\$0 to \$15k
high	unknown	high	none	\$15 to \$35k
moderate	unknown	low	none	\$15 to \$35k
high	unknown	low	none	\$0 to \$15k
low	unknown	low	none	over \$35k
low	unknown	low	adequate	over \$35k
high	bad	low	none	\$0 to \$15k
moderate	bad	low	adequate	over \$35k
low	good	low	none	over \$35k
low	good	high	adequate	over \$35k
high	good	high	none	\$0 to \$15k
moderate	good	high	none	\$15 to \$35k
low	good	high	none	over \$35k
high	bad	high	none	\$15 to \$35k

决策树表示

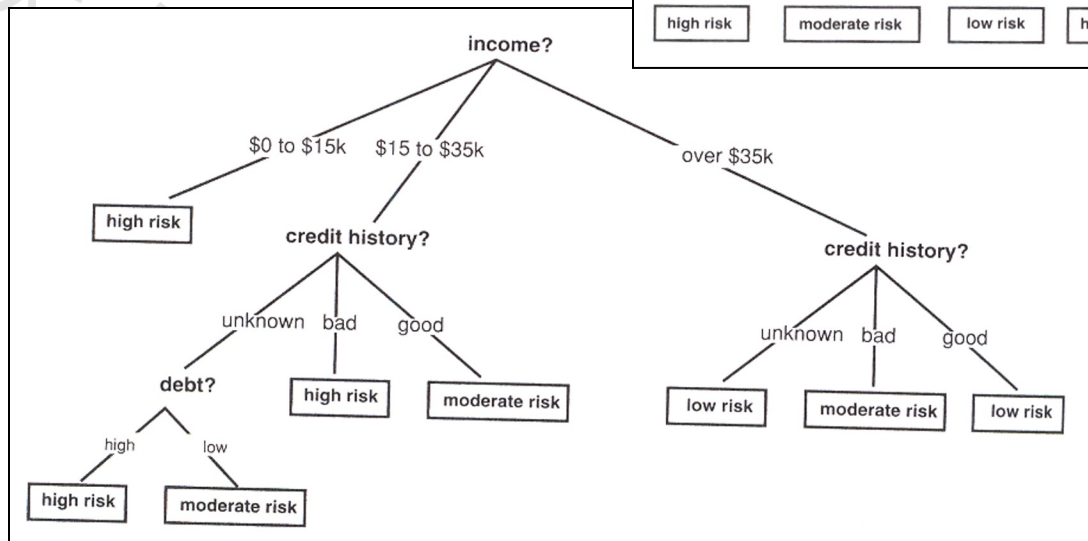
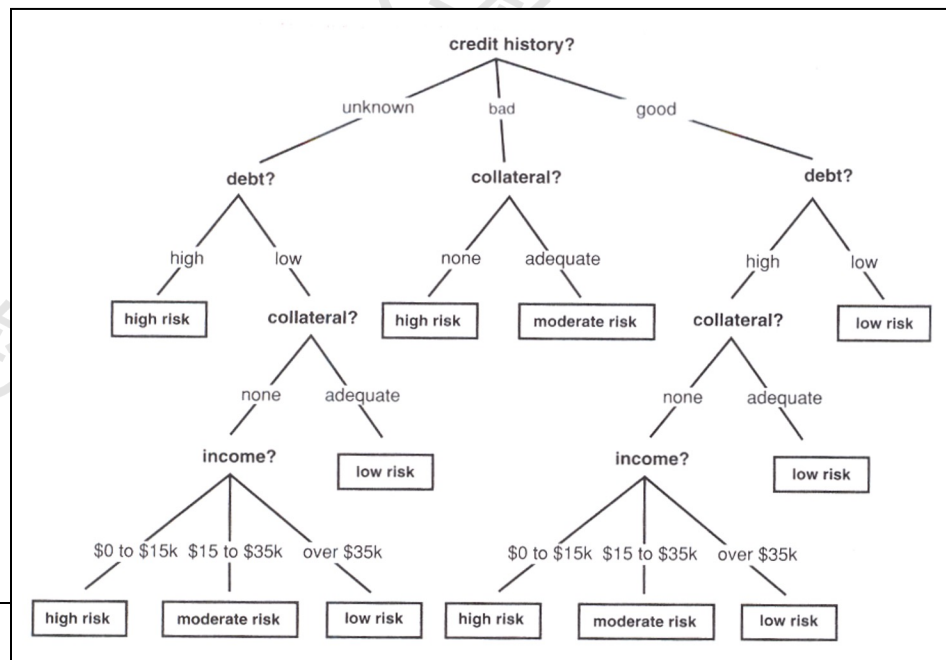
If-then规则



测试结点

叶结点

哪一个决策树能最大可能的对未知实例进行正确分类？



问题设置

□ 问题设置

- ✓ 可能的实例集 X
- ✓ 未知的目标函数 $f: X \rightarrow Y$
- ✓ 假设函数集 $H = \{h | h: X \rightarrow Y\}$

□ 输入

- ✓ 未知目标函数 f 的训练实例 $\{ \langle x_i, y_i \rangle \}$

□ 输出

- ✓ 最佳近似 f 的假设 $h \in H$

□ 算法框架

- ✓ 处理基本情况
- ✓ 寻找最好的分类属性 A_{best}
- ✓ 用 A_{best} 建立一个结点划分样例
- ✓ 递归处理每一个划分，作为其子结点/子树

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

打网球



目标属性

RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
high	bad	high	none	\$0 to \$15k
high	unknown	high	none	\$15 to \$35k
moderate	unknown	low	none	\$15 to \$35k
high	unknown	low	none	\$0 to \$15k
low	unknown	low	none	over \$35k
low	unknown	low	adequate	over \$35k
high	bad	low	none	\$0 to \$15k
moderate	bad	low	adequate	over \$35k
low	good	low	none	over \$35k
low	good	high	adequate	over \$35k
high	good	high	none	\$0 to \$15k
moderate	good	high	none	\$15 to \$35k
low	good	high	none	over \$35k
high	bad	high	none	\$15 to \$35k

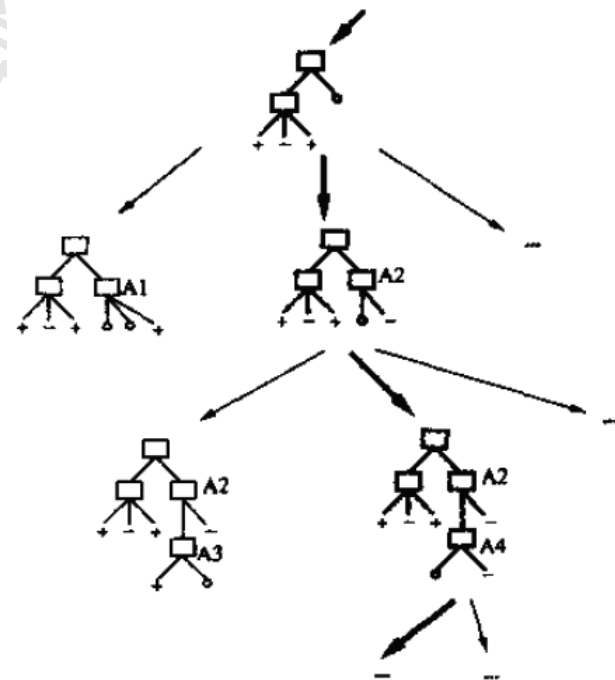


信用卡风险

目标属性

决策树学习的假设空间搜索

- 从一个假设空间中**搜索一个正确拟合训练样例**的假设。
- 搜索的假设空间就是**可能的决策树的集合**。
- 从简单到复杂的**爬山算法遍历假设空间**。从空的树开始，然后逐步考虑更加复杂的假设。引导爬山搜索的评估函数是信息增益度量。



用于学习布尔函数的ID3算法

ID3(Examples, Attributes)

1. 创建树的Root结点
2. 如果所有 Examples 的目标属性均为“正”，那么返回 label= “+” 的单结点树Root
3. 如果所有 Examples 的目标属性均为“反”，那么返回 label= “-” 的单结点树Root
4. 如果Attributes为“空”，那么返回单结点树Root，label设置为Examples中最普遍的目标属性值

.....




基本情况

用于学习布尔函数的ID3算法

ID3(Examples, Attributes)

5. 否则

- $A \leftarrow$ Attributes中分类Examples能力最好的属性
- Root的决策属性 $\leftarrow A$
- 对于A的每个可能值 v_i
 - 令Examples $_{v_i}$ 为Examples中满足A属性值为 v_i 的子集
 - 如果Examples $_{v_i}$ 为空  **特殊情况**
 - 在这个分支下加一个叶子结点，结点的label设置为Examples中最普遍的目标属性值
 - 否则，在这个分支下加一个子树ID3(Examples $_{v_i}$, Attributes- $\{A\}$)

6. 结束，返回树Root

如何选择最佳属性

□ 衡量给定的属性区分训练样例的能力

✓ 信息增益 information gain

□ 信息的度量

✓ 熵 entropy

✓ 刻画了样例集的纯度(purity)

□ 假设 X 是一个有限取值的离散随机变量

概率分布: $P(X = x_i) = p_i, i = 1, 2, \dots, n$

随机变量 X 的熵: $\text{Entropy}(X) = -\sum_{i=1}^n p_i \log p_i$

如何选择最佳属性

- 衡量给定的属性区分训练样例的能力

 - ✓ 信息增益 information gain

- 信息的度量

 - ✓ 熵 entropy

 - ✓ 刻画了样例集的纯度(purity)

- 目标属性为布尔值的样例集S的熵

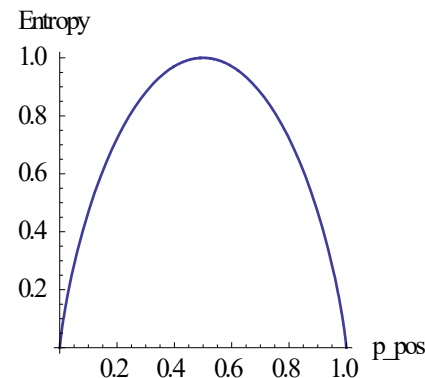
$$\text{Entropy}(S) = -p_+ \log p_+ - p_- \log p_-$$

例子

- 假设 S 是一个关于某概念的有14 个样例的集合，它包括9 个正例和5 个反例（我们用记号 $[9_+, 5_-]$ 来概括这样的数据样例）。那么 S 相对于这个布尔分类的熵（Entropy）为

$$\text{Entropy}([9_+, 5_-]) = -\frac{9}{14} \log \frac{9}{14} - \frac{5}{14} \log \frac{5}{14} = 0.940$$

- 如果 S 的所有成员属于同一类， $-1\log_2 1 + 0\log_2 0 = 0$



信息增益

□ 熵的一般定义

$$\text{Entropy}(S) = - \sum_{i=1}^n p_i \log p_i$$

□ 信息增益

- ✓ 使用属性分割样例，导致的期望熵降低

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

- ✓ 其中 $\text{Values}(A)$ 是属性 A 所有可能值的集合
- ✓ S_v 是 S 中属性 A 的值为 v 的子集(也就是, $S_v = \{s \in S | A(s) = v\}$)
- ✓ 第一项就是原来集合 S 的熵, 第二项是用 A 分类 S 后熵的**期望值**

ID3算法特点

- ❑ 假设空间：包含所有的决策树
- ❑ 遍历过程：仅维持单一的当前假设
 - ✓ 不同于变型空间候选消除算法（维持满足训练样例的所有假设）
 - ✓ 如何评判其他候选假设？
- ❑ 回溯：不进行回溯
 - ✓ 局部最优
- ❑ 基于统计
 - ✓ 对错误样例不敏感
 - ✓ 不适用于增量处理



C4.5等改进算法

决策树学习中的归纳偏置

近似：优先选择较短的树



优先选择信息增益高的属性更接近根结点的树

搜索策略决定了归纳偏置



某种假设优于其他假设(preference)

优先偏置(搜索偏置) vs 限定偏置(语言偏置)

奥卡姆剃刀原理

奥卡姆剃刀：如果对于同一现象有两种不同的假说，应该采取比较简单的那一种。



奥卡姆剃刀：优先选择拟合数据的最简单假设。



不是简单的选择最简化的假设，而是推理所依据的是使可证伪的假设的数目更少

Occam's razor (1324)公理：

“如果少做就能完成，多做即是徒劳……如无必要，勿增实体。”

It is vain to do with more what can be done with less. ... Entities should not be multiplied beyond necessity.

大纲

符号学习

变型空间

归纳偏置

ID3决策树算法

其他树算法

C4.5算法

□ 属性选择指标

- ✓ 信息增益准则对可取值数目较多的属性有所偏好

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

- ✓ 信息增益比：信息增益/该属性的熵

$$\text{GainRate}(S, A) = \frac{\text{Gain}(S, A)}{\text{Entropy}_A(S)}$$

$$\text{Entropy}_A(S) = - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|}$$

C4.5算法

□ 属性选择指标

- ✓ 信息增益准则对可取值数目较多的属性有所偏好

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

- ✓ 信息增益比：信息增益/该属性的熵

$$\text{GainRate}(S, A) = \frac{\text{Gain}(S, A)}{\text{Entropy}_A(S)}$$

$$\text{Entropy}_A(S) = - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|}$$

- ✓ 避免对可取值数目较少的属性有所偏好，采用以下启发式
 - 先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益比最高的

CART算法

□ 属性选择指标(分类)

- ✓ 信息增益和增益比准则均需要计算对数
- ✓ Gini指数：模型的纯度，越小越好，越小纯度越高
- ✓ 只需要计算比例
- K 个类，样本点属于第 k 类的概率为 p_k

基尼指数：
$$Gini(p) = \sum_{k=1}^K (1 - p_k)p_k = 1 - \sum_{k=1}^K p_k^2$$

CART算法

□ 属性选择指标(分类)

- ✓ 信息增益和增益比准则均需要计算对数
- ✓ Gini指数：模型的纯度，越小越好，越小纯度越高
- ✓ 只需要计算比例

- K 个类，样本点属于第 k 类的概率为 p_k

基尼指数：
$$Gini(p) = \sum_{k=1}^K (1 - p_k)p_k = 1 - \sum_{k=1}^K p_k^2$$

- 二分类问题

基尼指数：
$$Gini(p) = 2p(1 - p)$$

CART算法

□ 属性选择指标(分类)

- ✓ 信息增益和增益比准则均需要计算对数
- ✓ Gini指数：模型的纯度，越小越好，越小纯度越高
- ✓ 只需要计算比例

- K 个类，样本点属于第 k 类的概率为 p_k

基尼指数：
$$Gini(p) = \sum_{k=1}^K (1 - p_k)p_k = 1 - \sum_{k=1}^K p_k^2$$

- 样本集合 D

基尼指数：
$$Gini(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2$$

C_k : 样本集合 D 中属于第 k 类的样本子集

CART算法

□ 属性选择指标(分类)

✓ 信息增益和增益比准则均需要计算对数

✓ Gini指数：模型的纯度，越小越好，越小纯度越高

✓ 只需要计算比例

- 样本集合 D

基尼指数：
$$Gini(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2$$

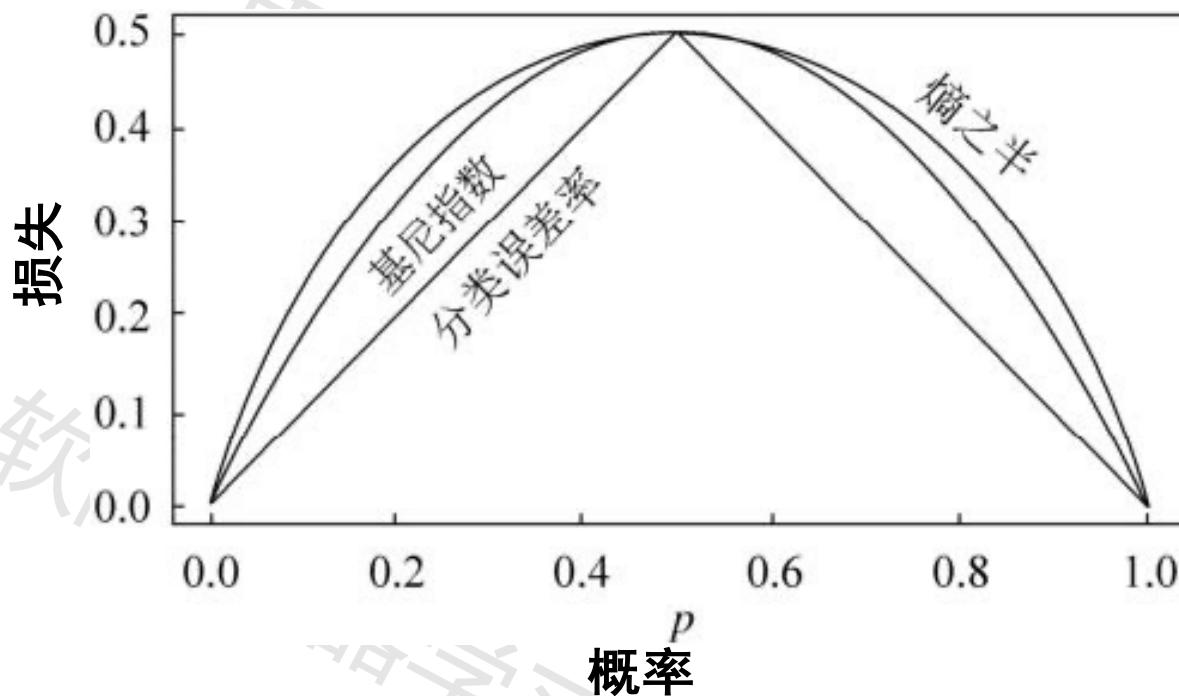
C_k : 样本集合 D 中属于第 k 类的样本子集

- 属性 A 对样本集合划分下的基尼指数

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

基尼系数 VS 信息增益

- 二分类问题



- 基尼系数和熵正相关，并且很接近，可近似代表分类误差率
- 基尼系数和熵均可表示一个集合的混乱程度，并作为叶子结点的损失

CART算法

□ 属性选择指标(回归)

- ✓ 采用方差和度量
- ✓ 度量目标是对于划分特征A，对应划分点s两边的数据集 D_1 和 D_2 ，求出使 D_1 和 D_2 各自集合的均方差最小，同时 D_1 和 D_2 的均方差之和最小。

$$\underbrace{\min}_{A,s} \left[\underbrace{\min}_{c_1} \sum_{x_i \in D_1(A,s)} (y_i - c_1)^2 + \underbrace{\min}_{c_2} \sum_{x_i \in D_2(A,s)} (y_i - c_2)^2 \right]$$

回归树输出不是类别，采用叶子结点的均值或者中位数来预测输出结果

连续值处理

□ 连续的特征离散化

- ✓ C4.5基于信息增益比离散化，CART是基于基尼系数离散化。
- ✓ m 个样本的连续特征 A 有 m 个，从小到大排列 a_1, a_2, \dots, a_m ，则CART取相邻两样本值的平均数做划分点，一共取 $m-1$ 个，其中第 i 个划分点 T_i 表示为： $T_i = (a_i + a_{i+1})/2$ 。
- ✓ 分别计算以这 $m-1$ 个点作为二元分类点时的基尼系数。
- ✓ 选择基尼系数最小的点为该连续特征的二元离散分类点。
- ✓ 如取到的基尼系数最小的点为 a_t ，则小于 a_t 的值为类别1；大于 a_t 的值为类别2，这样就做到了连续特征的离散化。

离散值处理

□ CART分类树算法

- ✓ 对离散值的处理，采用不停的二分离散特征。

- ✓ 多叉树

 - ✓ 在ID3、C4.5，特征A被选中，如果它有3个取值A1,A2,A3，则建立三叉子树。

- ✓ 二叉树

 - ✓ CART将特征A分成{A1}和{A2,A3}、{A2}和{A1,A3}、{A3}和{A1,A2}三种情况，找到基尼系数最小的组合，比如{A2}和{A1,A3}，然后建立二叉树结点。

 - ✓ 由于并没有把特征A的取值完全分开，后面还有机会对子结点继续选择特征A划分A1和A3。

代表性树算法对比

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝
ID3	分类	多叉树	信息增益	不支持	不支持	不支持
C4.5	分类	多叉树	信息增益比	支持	支持	支持
CART	分类/回归	二叉树	基尼系数 均方差	支持	支持	支持

剪枝处理

□ 后剪枝

- ✓ 从完全生长的决策树的底端剪去一些子树，使决策树变小（模型简单），从而增强泛化能力。
- ✓ 首先从生成算法产生的决策树 T_0 底端开始不断剪枝，直到 T_0 的根结点，形成一个子序列 T_0, T_1, \dots, T_n
- ✓ 然后通过交叉验证在独立的验证集上对子树序列进行测试，从中选择最优子树。

剪枝处理

□ 最小化子树的损失函数

✓ $C_a(T) = C(T) + a|T|$

✓ T 为任意子树， $C(T)$ 为对数据的预测误差(如基尼系数)， $|T|$ 为子树叶结点个数。超参 $a \geq 0$ ，权衡训练数据的拟合程度与模型的复杂度。

- ✓ a 比较大，则最优子树 T_a 偏小；
- ✓ a 比较小，则最优子树 T_a 偏大；
- ✓ $a = 0$ ，则最优子树等于未剪枝的 T_0 ；
- ✓ $a \rightarrow \infty$ ，则最优子树为根结点数。

树学习算法优点

- 简单直观，生成的决策树很直观。相比于神经网络之类的黑盒分类模型，决策树在逻辑上可以很好解释，可解释性强。
- 基本不需要预处理，不需要提前归一化和处理缺失值。既可以处理离散值也可以处理连续值。很多算法只是专注于离散值或者连续值。
- 可以处理多维度输出的分类问题。
- 使用决策树预测的代价是 $O(\log_2 m)$ 。 m 为样本数。
- 可以交叉验证的剪枝来选择模型，从而提高泛化能力。
- 对于异常点的容错能力好，健壮性高。

树学习算法缺点

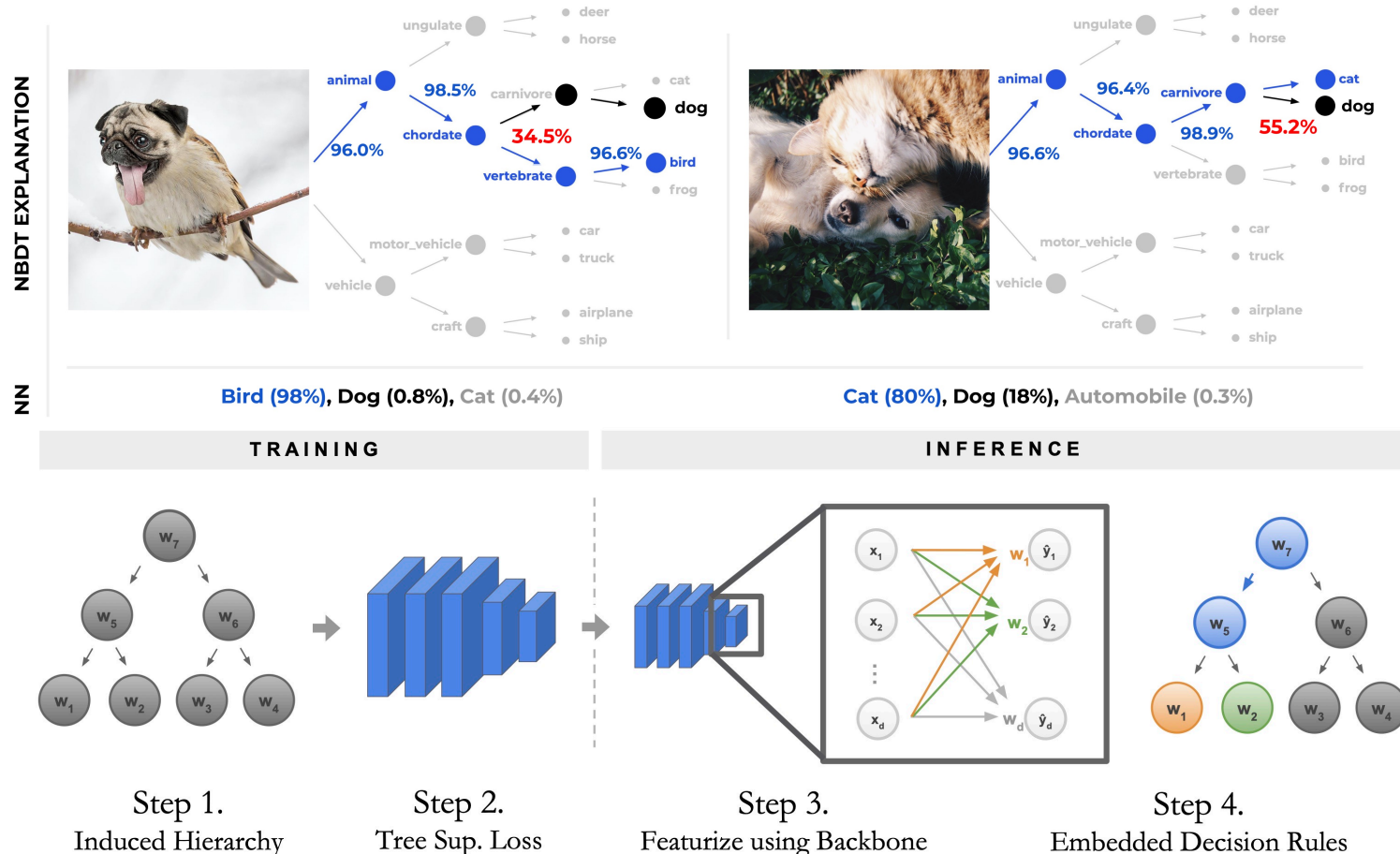
- ❑ 树算法非常容易过拟合，导致泛化能力不强。可以通过设置结点最少样本数量和限制决策树深度来改进。
- ❑ 决策树会因为样本发生一点的改动，导致树结构的剧烈改变。可以通过集成学习之类的方法解决。
- ❑ 寻找最优的决策树是一个NP难题，通过启发式方法，容易陷入局部最优。可以通过集成学习的方法来改善。
- ❑ 比较复杂的关系(如异或)，决策树很难学习。一般这种关系只能用其他学习方法(如神经网络)来解决。
- ❑ 如果某些特征的样本比例过大，生成决策树容易偏向于这些特征。可以通过调节样本权重来改善。

思考和讨论

1. 假设的一般到特殊的偏序结构？
2. FIND-S与候选消除算法的区别？
3. 符号(概念)学习如何处理噪声数据呢？
4. 不同算法的归纳偏置是什么？
5. 如何理解奥卡姆剃刀原则？
6. 学习CART算法的细节。

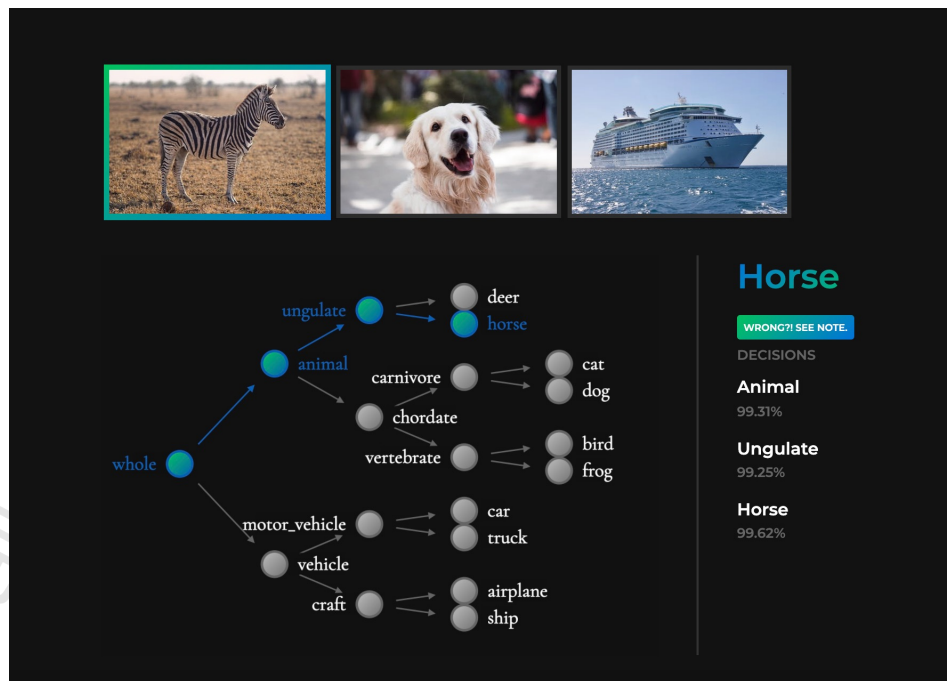
决策树的延伸

- 深度学习时代的决策树？



决策树的延伸

- 深度学习时代的决策树？



- BAIR 博客地址: <https://bair.berkeley.edu/blog/2020/04/23/decisions/>
- 论文地址: <https://arxiv.org/abs/2004.00221>
- 开源项目地址: <https://github.com/alvinwan/neu>

谢谢！