Peking University

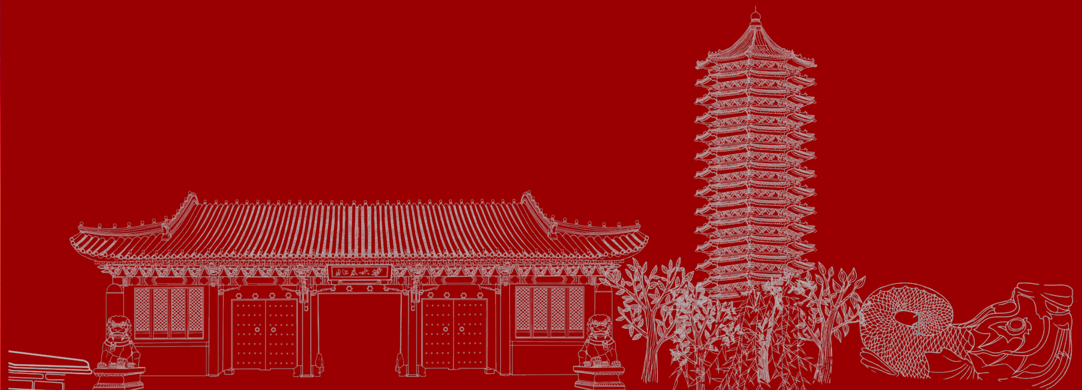
北京大学
PEKING UNIVERSITY

人工智能引论

11. 决策树与随机森林

授课教师 连宙辉

2022年4月3日



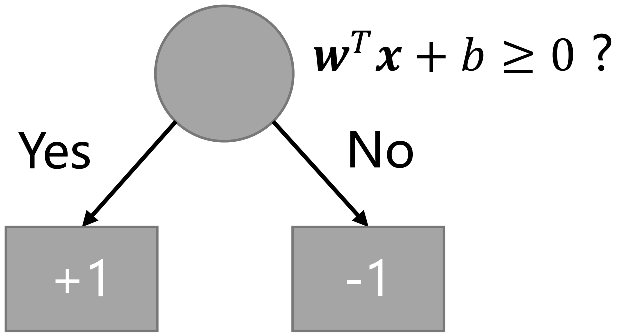
目录

- **决策树**
  - 什么是决策树
  - 划分准则：信息增益、增益率、基尼系数
  - 连续属性
- **回归树**
  - 连续标签
- **随机森林**

回顾

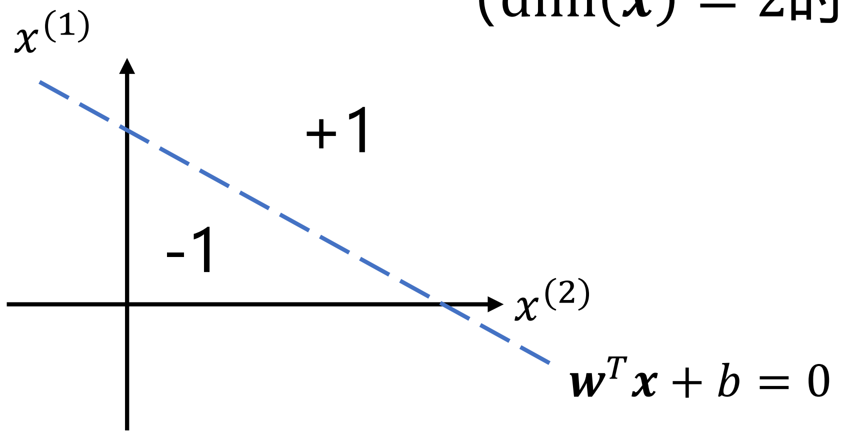
- 线性模型：$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$
  - 在分类任务中，分类逻辑为：$\begin{cases} y = +1 & if\ f(\boldsymbol{x}) > 0 \\ y = -1 & if\ f(\boldsymbol{x}) \leq 0 \end{cases}$

以“树”的形式表示该逻辑



以在特征空间中的分类边界表示该逻辑

($\dim(\boldsymbol{x}) = 2$时)



线性模型仅能给出简单（单层）的“树”，或简单的线性分隔超平面——

是否可以设计出具有比 $\text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$ 逻辑更复杂的分类模型?

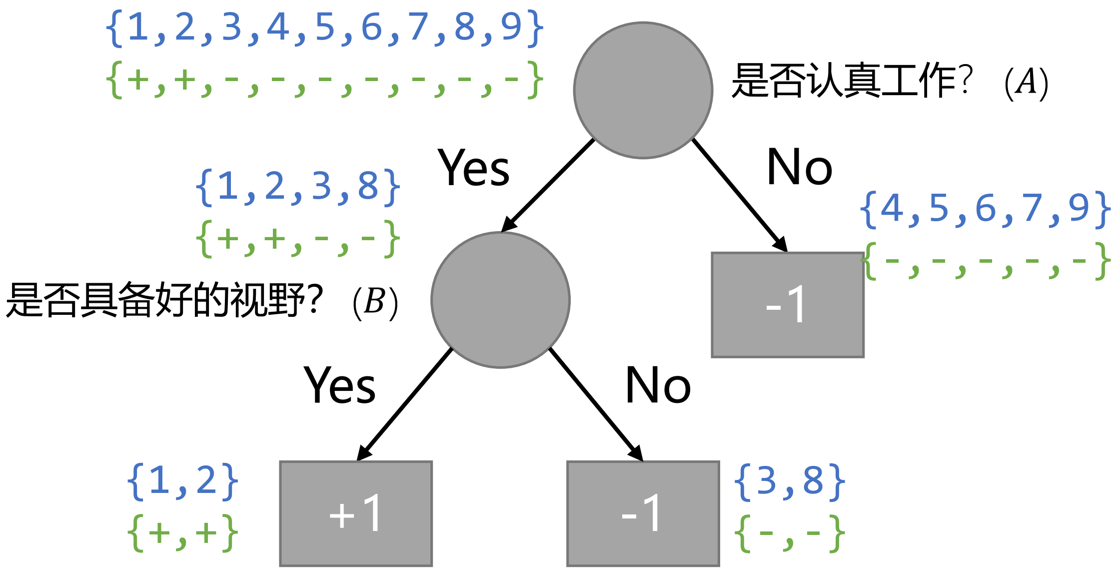
决策树

北京大学 PEKING UNIVERSITY

- 决策树是一种常见的机器学习方法
- 定义：决策树是一种树形结构，包含一个根节点、若干个内部节点和若干个叶节点。其中：
  - 根节点包含样本全集
  - 每个非叶节点对应于一个（特征）属性测试；其包含的样本集合根据属性测试的结果被划分到子节点中
  - 每个叶节点对应于决策结果（取多数类）

决策树

- Example：
  - 属性集：{A: 是否认真工作 (±1)，B: 是否具备好的视野 (±1)，C: 是否喜欢吃香蕉 (±1)}
  - 标签：是否是一个好的科学家 ($y \in \{\pm 1\}$)
  - 一个可能的决策树：



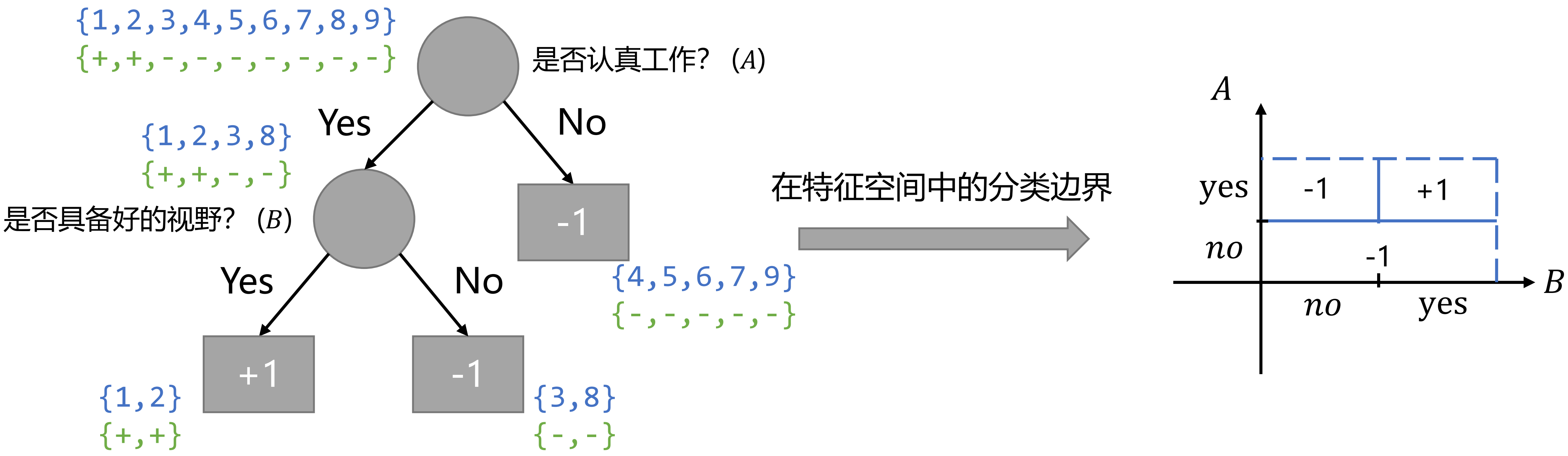
序号	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

表：训练数据

注：
{1,2...}表示data point序号
{+-}表示data point标签

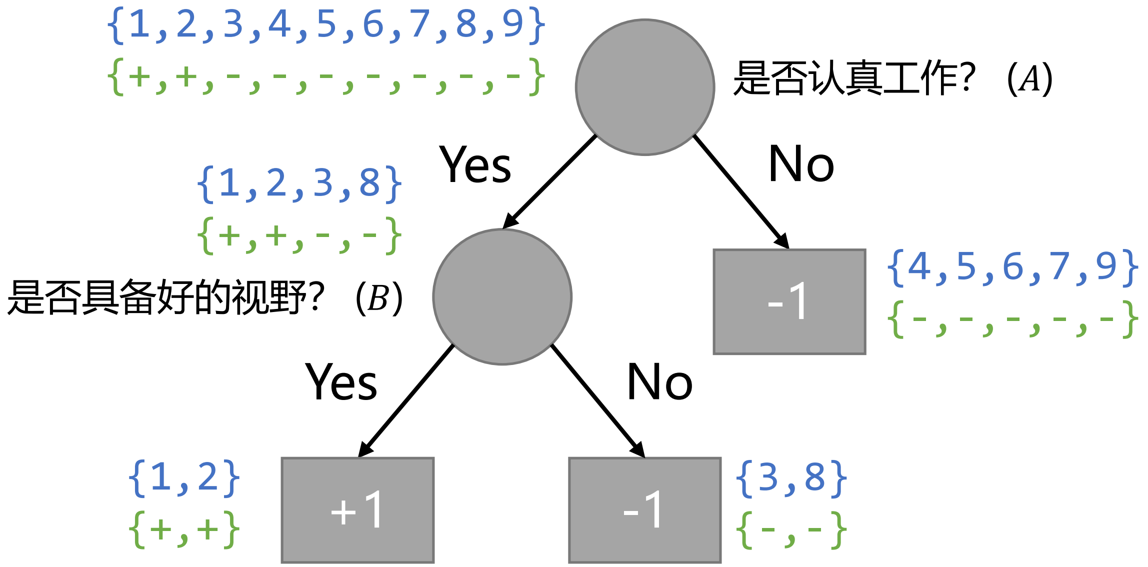
决策树

- Example：


{1,2,3,4,5,6,7,8,9}
{+,+,-,-,-,-,-,-,-}
是否认真工作？（A）
Yes
No
{1,2,3,8}
{+,+,-,-}
是否具备好的视野？（B）
-1
{4,5,6,7,9}
{-,-,-,-,-}
Yes
No
+1
-1
{1,2}
{+,+}
{3,8}
{-,-}
在特征空间中的分类边界
A
yes
no
-1
+1
-1
B
no
yes

决策树

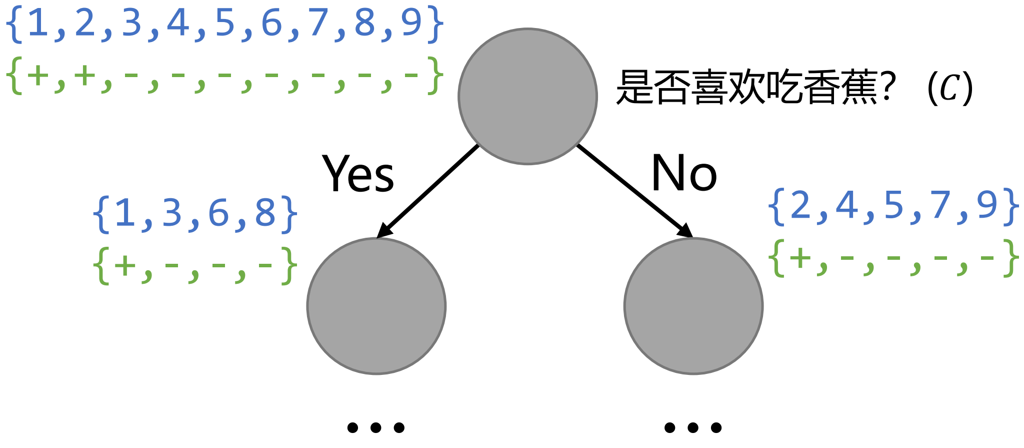
- Example：
  - 在该分类树中，A与B均能很好地将训练样本不断划分为标签“纯度(purity)”更高的子集。



  - 理想情况下，每个叶节点只包含同类别的训练样本

决策树

- Example：
  - 另一个可能的决策树：



序号	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

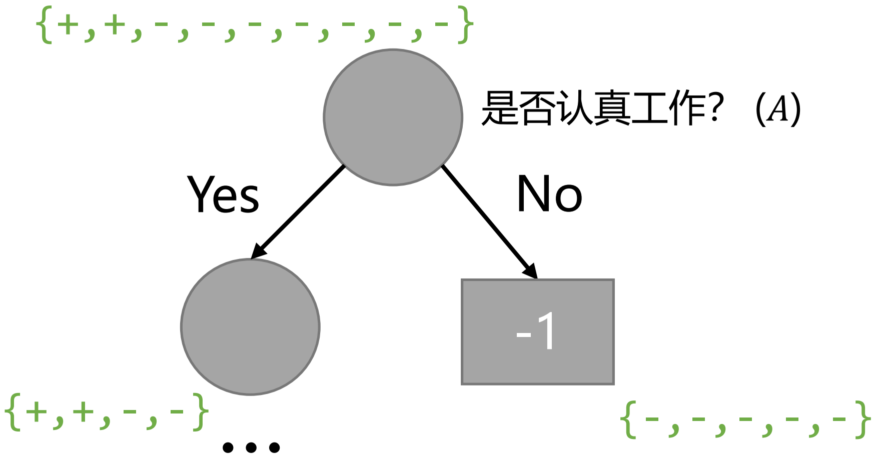
- 在该分类树中，C划分前后，样本集合中的标签“纯度”并没有显著提升
- C属性本身（是否喜欢吃香蕉）与预测目标（是否是一个好的科学家）没有直接联系

如何选择出最优划分属性？

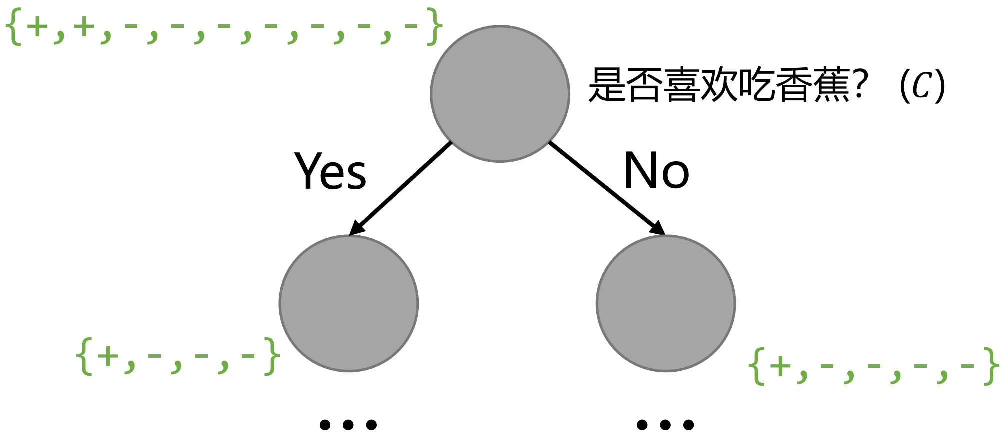
最优划分属性

- 如何选择出最优划分属性?
  - 一般而言，随着划分的不断进行，我们希望决策树的分支节点所包含的样本子集“纯度”不断增高
  - 若我们可以找到某一个属性M，根据属性M的不同取值对当前节点包含的样本集合进行划分后，每个子集都尽量只包含同类的标签
  - 则该属性M是一个较好的划分属性
  - **训练**：递归地使用当前最好的属性对训练集进行划分，直到纯度达到要求

序号	认真工作A	视野B	喜欢香蕉C	好科学家y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×



较好的划分



较差的划分

纯度度量与划分准则

- 如何度量一个集合的纯度?
  - 信息熵 (Information entropy)
  - 基尼指数 (Gini index)

- 如何选择一个好的划分准则?
  - 信息增益 (Information gain)
  - 增益率 (Gain ratio)
  - (属性的) 基尼指数 (Gini index)

熵的基本知识

北京大学 PEKING UNIVERSITY

- 熵是热力学和信息论中的一个重要概念，用于描述系统的混乱程度或者不确定性。在信息论中，熵被定义为一个随机变量的不确定性
  - 注：以下均为信息论中定义的熵
- 对于离散型随机变量 X，服从概率分布 $p(x)$，其信息熵定义为：

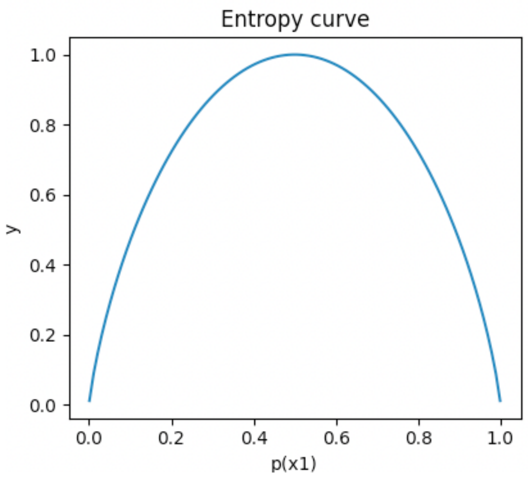
$$H(X) = -\sum_{x} p(x)\log_2 p(x) = \sum_{x} p(x)\log_2 \frac{1}{p(x)}$$

- 定义理解
  - 对于某一事件 $X = x$，其发生的概率 $p(x)$ 越小，则包含的“信息量”越大
  - 当 $X = x$ 为一确定事件，即 $p(x) = 1$时，其所包含的“信息量”最小（=0）
  - 因此可以用$\log_2 \frac{1}{p(x)}$来描述该事件所包含的“信息量”
  - 信息熵$H(X)$为系统内不同事件(X取不同值)的“信息量”的期望值 $\mathbb{E}_x[\log_2 \frac{1}{p(x)}]$

熵的基本知识

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

- 熵的一些基本性质
  - $H(X) \geq 0$恒成立
  - 当且仅当$p(x)$为一个“确定分布”时（即存在一个确定事件概率为1，其他事件概率为0），$H(X) = 0$
  - 由于$\log_2 x$为凹函数，由琴生不等式可以得到： $H(X) = \mathbb{E}_x\left[\log_2\left(\frac{1}{p(x)}\right)\right] \leq \log_2\left(\mathbb{E}_x\left[\frac{1}{p(x)}\right]\right) = \log_2 n$ (n为可能的事件个数)
    - 等号成立条件为 $\forall x$，$p(x) = \frac{1}{n}$
    - 即，所有事件等概率时，不确定性最大，此时达到最大熵$\log_2 n$
    - 当X仅有两种可能的取值x_1, x_2时，$H(X)$随着$p(x_1)$的变化曲线
    - 例如抛硬币，当正反两面等可能时不确定性最大



熵的基本知识

- 联合熵 (Joint Entropy)
  - 在离散情况下，一个联合概率分布$p(x,y)$的联合熵定义为:

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log_2 p(x,y)$$

  - 联合熵描述一个联合概率分布的不确定程度。

- 条件熵 (Conditional Entropy)
  - 条件熵描述在已知随机变量X的条件下随机变量Y的不确定性。
  - 在离散情况下，且在已知随机变量X时，随机变量Y的条件熵为:

$$H(Y\mid X) = \sum_{x} p(x)H(Y\mid X=x) = -\sum_{x}\sum_{y} p(x,y)\log_2 p(y\mid x)$$

熵的基本知识

- 条件熵 (Conditional Entropy)

$$H(Y \mid X) = -\sum_{x}\sum_{y} p(x,y)\log_2 p(y \mid x)$$

- 推导过程：

$$\begin{aligned}
H(Y \mid X) &= \sum_{x} p(x) H(Y \mid X = x) \\
&= -\sum_{x} p(x) \sum_{y} p(y \mid x)\log_2 p(y \mid x) \\
&= -\sum_{x}\sum_{y} p(x,y)\log_2 p(y \mid x)
\end{aligned}$$

熵的基本知识

- 条件熵与联合熵具有关系：

$$H(X,Y)=H(Y)+H(X\mid Y)=H(X)+H(Y\mid X)=H(Y,X)$$

- 证明

$$\begin{aligned}
H(Y)+H(X\mid Y) &= -\sum_y p(y)\log_2 p(y)-\sum_x\sum_y p(x,y)\log_2 p(x\mid y)\\
&= -\sum_y\sum_x p(x,y)\log_2 p(y)-\sum_x\sum_y p(x,y)\log_2 p(x\mid y)\\
&= -\sum_x\sum_y (p(x,y)\log_2 p(y)+p(x,y)\log_2 (\tfrac{p(x,y)}{p(y)}))\\
&= -\sum_x\sum_y p(x,y)\log_2 p(x,y)=H(X,Y)
\end{aligned}$$

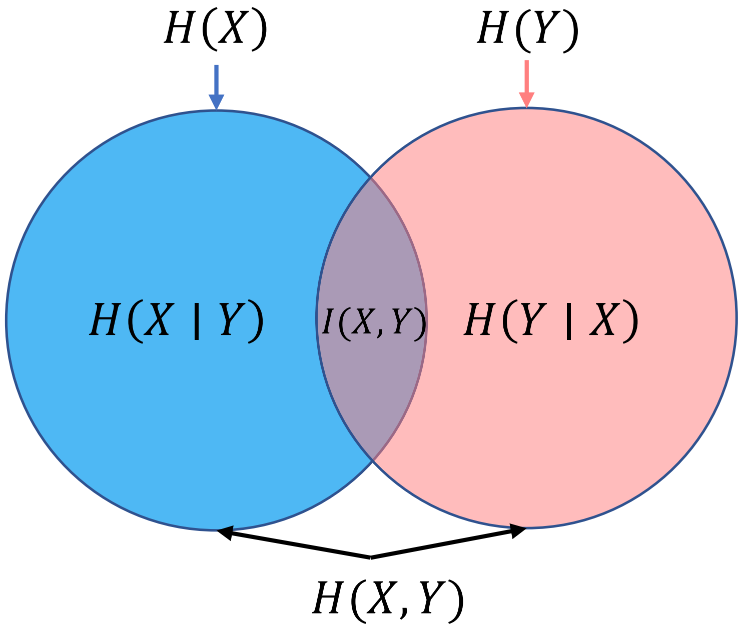
其中，$\sum_x p(x,y)=p(y)$（边缘概率定义），$p(x\mid y)=\frac{p(x,y)}{p(y)}$（条件概率定义）

其他同理

熵的基本知识

- 互信息 (Mutual Information)
  - 在离散情况下，两个随机变量X和Y的互信息为：
$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
  - 结合上式，有：
$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
  - 由此可见，互信息描述Y中“包含”X（或X中“包含”Y）的信息量
  - 结合定义及前式，有若干结论：
    - $I(X,Y) = I(Y,X)$
    - $I(X,X) = H(X)$
    - X,Y相关性越强，$I(X,Y)$越大
    - X,Y相互独立时 ($p(x,y) = p(x)p(y)$)，$I(X,Y) = 0$

几类熵的关系



划分准则1：信息增益

- 用信息熵度量样本集合D的标签纯度：

$$H(D)=-\sum_{k\in[K]}\frac{|D_k|}{|\mathrm{D}|}\log\frac{|D_k|}{|\mathrm{D}|}$$

  - $|\cdot|$表示集合元素数目，D_k表示集合D中标签为k的子集，$[K]$为标签集（K分类）
  - $H(D)$即为经验分布下标签的信息熵
  - $H(D)$越低，样本集合D的纯度越高，当D中样本均属某一类k时，$H(D)=0$

{1,2,3,4,5,6,7,8,9}
{+,+,-,-,-,-,-,-,-}

$\Rightarrow$ $H(D)=??$

划分准则1：信息增益

- 用信息熵度量样本集合D的纯度： $H(D) = -\sum_{k\in[K]} \frac{|D_k|}{|D|} \log \frac{|D_k|}{|D|}$
- 属性A对集合D的信息增益：

$$g(D,A) = H(D) - \sum_{i\in[m]} \frac{|D^{A=a_i}|}{|D|} H(D^{A=a_i})$$

  - 其中，属性A的可能取值为$\{a_i \mid i \in [m]\}$，$D^{A=a_i}$表示集合D中样本属性A取a_i的子集。
  - 若记随机变量Y为标签、随机变量X为属性A，则：
    - 第二项为标签相对于属性的条件熵： $\sum_{i\in[m]} \frac{|D^{A=a_i}|}{|D|} H(D^{A=a_i}) = H(Y \mid X)$
    - 信息增益为标签与属性之间的互信息：$g(D,A) = I(Y,X)$
  - 一般情况下，信息增益越大，则使用该属性进行划分所得到的“纯度提升”越大

划分准则1：信息增益

- 划分准则：选择信息增益最大的属性进行划分
- 对于该例，在根节点选择属性时：

$$D=\{\mathrm{data}_1 \sim \mathrm{data}_9\}$$

$$H(D)=(-\frac{2}{9}\log\frac{2}{9})+(-\frac{7}{9}\log\frac{7}{9})=0.764$$

$$g(D,A)=0.764-(\frac{4}{9}\left(-\frac{2}{4}\log\frac{2}{4}-\frac{2}{4}\log\frac{2}{4}\right)+\frac{5}{9}\left(-\frac{5}{5}\log\frac{5}{5}-0\right))$$

$$=0.764-(0.444+0)$$

$$=\mathbf{0.320}$$

$$g(D,B)=0.225$$

$$g(D,C)=0.0026$$

序号	认真工作 *A*	视野 *B*	喜欢香蕉 *C*	好科学家 *y*
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

- 因此选择属性A作为划分属性。

划分准则1：信息增益

- 思考：若将“序号”（ I ）也列为属性之一……
  - 划分后的平均信息熵为0（每一个I的取值仅有一个样本），信息增益最高！
    - $g(D, I) = 0.764 - 0 = \mathbf{0.764}$
  - 然而此时的决策树显然不具备任何泛化能力！
  - 信息增益准则倾向于选择具备更多取值可能的属性
  - 如何解决?

序号 I	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

划分准则2：增益率

- 解决方法：采用增益率作为划分准则。
- 属性A的增益率定义为：

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

  - 其中，$H_A(D)=-\sum_{i\in[m]}\frac{|D^{A=a_i}|}{|D|}\log\frac{|D^{A=a_i}|}{|D|}$ 描述**属性A本身的熵**（不确定程度）
  - $H_A(D)$ 不同于 $H(D^{A=a_i})$！前者以A作为"标签"衡量A的纯度，后者还是在衡量标签y本身的纯度（在$D^{A=a_i}$ 上）
  - 当属性A具备较多的取值时，其本身在D上的不确定性较高，$H_A(D)$较大，使得对应增益率较低；一定程度上缓解了属性本身带来的高信息增益
- 在候选属性集合F中选取使得划分后增益率最高的属性，即:

$$A_*=argmax_{A\in F}\,g_R(D,A)$$

序号	认真工作A	视野B	喜欢香蕉C	好科学家y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

划分准则2：增益率

在上述例子中，

$$g(D,I) = 0.764 > g(D,A) = 0.320$$

然而

$$g_R(D,I) = \frac{0.764}{(-\frac{1}{9}\log\frac{1}{9}) * 9} = 0.241$$

$$g_R(D,A) = \frac{0.320}{(-\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9})} = 0.323$$

$$g_R(D,I) = 0.241 < g_R(D,A) = 0.323$$

序号 I	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

- 根据增益率划分，仍旧应当选择A作为最优划分属性
- 增益率准则对可取值数目较少的属性有所偏好

划分准则3：基尼指数

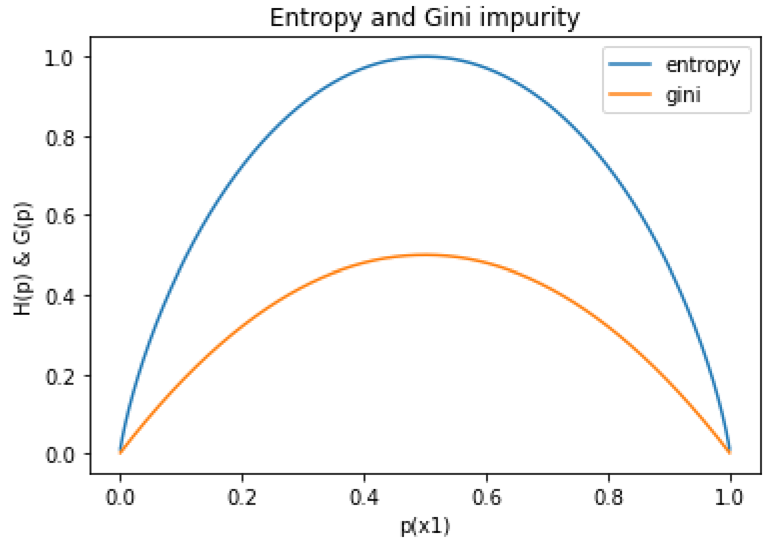
- 基尼指数 (Gini index)：另一种度量样本集合D纯度的方法

$$\mathrm{ini}\ \ (D) = \sum_{k\in[K]} \frac{|D_k|}{|D|}\left(1 - \frac{|D_k|}{|D|}\right) = 1 - \sum_{k\in[K]} \left(\frac{|D_k|}{|D|}\right)^2$$

  - $|\cdot|$表示集合元素数目，D_k表示集合D中标签为k的子集，$[K]$为标签集（K分类）

$$\mathrm{ini}\ \ (D) \approx \sum_{k\in[K]} P(y=k)(1-(P(y=k))$$

  - 反映随机抽取2个样本，其类别不一样的概率
  - 类似于信息熵，当只有2种可能取值时，基尼指数和信息熵随取值概率的变化变化曲线如右图



划分准则3：基尼指数

- 集合D的基尼指数：

$$\text{ini}\ \ (D) = \sum_{k\in[K]} \frac{|D_k|}{|D|}\left(1-\frac{|D_k|}{|D|}\right) = 1 - \sum_{k\in[K]}\left(\frac{|D_k|}{|D|}\right)^2$$

- 属性A对集合D的基尼指数：

$$\text{ini}\ \ (D, A) = \sum_{i\in[m]} \frac{|D^{A=a_i}|}{|D|} Gini(D^{A=a_i})$$

  - 其中，属性A的可能取值为$\{a_i \mid i \in [m]\}$，$D^{A=a_i}$表示集合D中样本属性A取a_i的子集

- 在候选属性集合F中选取使得划分后基尼指数最小的属性(平均纯度最大)，即：

$$A_* = argmin_{A\in F}\ \text{ini}\ \ (D, A)$$

划分准则3：基尼指数

- 对于该例，在根节点时选择属性时：

$$D = \{\mathrm{data}_1 \sim \mathrm{data}_9\}$$

$$\mathrm{ini}\ (D) = 1 - \left(\frac{2}{9}\right)^2 - \left(\frac{7}{9}\right)^2 = 0.346$$

$$\mathrm{ini}\ (D, A) = \left(\frac{4}{9}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{5}{9}\left(1 - \left(\frac{5}{5}\right)^2\right)\right)$$

$$= 0.222 + 0$$

$$= \mathbf{0.222}$$

$$\mathrm{ini}\ (D, B) = 0.267$$

$$\mathrm{ini}\ (D, C) = 0.344$$

序号	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

- 因此选择属性A作为划分属性

属性的离散到连续

北京大学 PEKING UNIVERSITY

- 数据的属性可以连续吗？
  - 是否认真工作 -> 每天工作时长

序号	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

序号	平均每天工作时长 A_c	视野 B	喜欢香蕉 C	好科学家 y
1	10	√	√	√
2	8	√	×	√
3	0	×	√	×
4	0	×	×	×
5	4	√	×	×
6	2	×	√	×
7	6	√	×	×
8	0	×	√	×
9	4	√	×	×

属性的离散到连续

- 使用离散化技术：二分法
  - 假设连续属性A在训练样本集D上有m个取值，升序排序为$\{a^1, a^2, \ldots, a^m\}$
  - 可建立$m-1$个候选划分点，合集为
    $T_A = \{\frac{a^i + a^{i+1}}{2} | i = 1, \ldots, m-1\}$
  - 选取一个划分点，将属性A二值化
  - 对于每个候选属性，选择最大化信息增益的候选划分点作为实际划分点，得到该候选属性的信息增益；再选取信息增益最大的候选属性

$$A_*, t_* = \operatorname*{argmax}_{A \in F, t \in T_A} g(D, A, t)$$

$$g(D, A, t) = H(D) - \left(\frac{|D^{A,t,+}|}{|D|} H(D^{A,t,+}) + \frac{|D^{A,t,-}|}{|D|} H(D^{A,t,-}) \right)$$

  - 其中，$D^{A,t,+}$表示集合D根据属性 A 按 t 二分后取正的子集

序号	平均每天工作时长A_c	视野 B	喜欢香蕉 C	好科学家 y
1	10	√	√	√
2	8	√	×	√
3	0	×	√	×
4	0	×	×	×
5	4	√	×	×
6	2	×	√	×
7	6	√	×	×
8	0	×	√	×
9	4	√	×	×

属性的离散到连续

- 以要求分为2类为例：只选取一个候选点，连续属性转为 $\{\pm 1\}$ 的离散属性。
- 如右图，A_c有6个取值$\{0,2,4,6,8,10\}$，5个候选划分点$\{1,3,5,7,9\}$
- 当取划分点=3时，二分为5正4负
- 当取划分点=7时，取得最大信息增益

$$D=\{\text{data}_1\sim\text{data}_9\},$$

$$H(D)=(-\frac{2}{9}\log\frac{2}{9})+(-\frac{7}{9}\log\frac{7}{9})=0.764$$

$$\begin{aligned}g(D,A,t=3)&\\&=0.764-\left(\frac{5}{9}\left(-\frac{2}{5}\log\frac{2}{5}-\frac{3}{5}\log\frac{3}{5}\right)+\frac{4}{9}\left(-\frac{4}{4}\log\frac{4}{4}-0\right)\right)\\&=0.225\end{aligned}$$

$$\begin{aligned}g(D,A,t=7)&\\&=0.764-\left(\frac{2}{9}\left(-\frac{2}{2}\log\frac{2}{2}-0\right)+\frac{7}{9}\left(-\frac{7}{7}\log\frac{7}{7}-0\right)\right)=\mathbf{0.764}\end{aligned}$$

序号	平均每天工作时长A_c	认真工作 $A_{t=3}$(取划分点3)	认真工作 $A_{t=7}$(取划分点7)	好科学家 y
1	10	√	√	√
2	8	√	√	√
3	0	×	×	×
4	0	×	×	×
5	4	√	×	×
6	2	×	×	×
7	6	√	×	×
8	0	×	×	×
9	4	√	×	×

标签的离散到连续

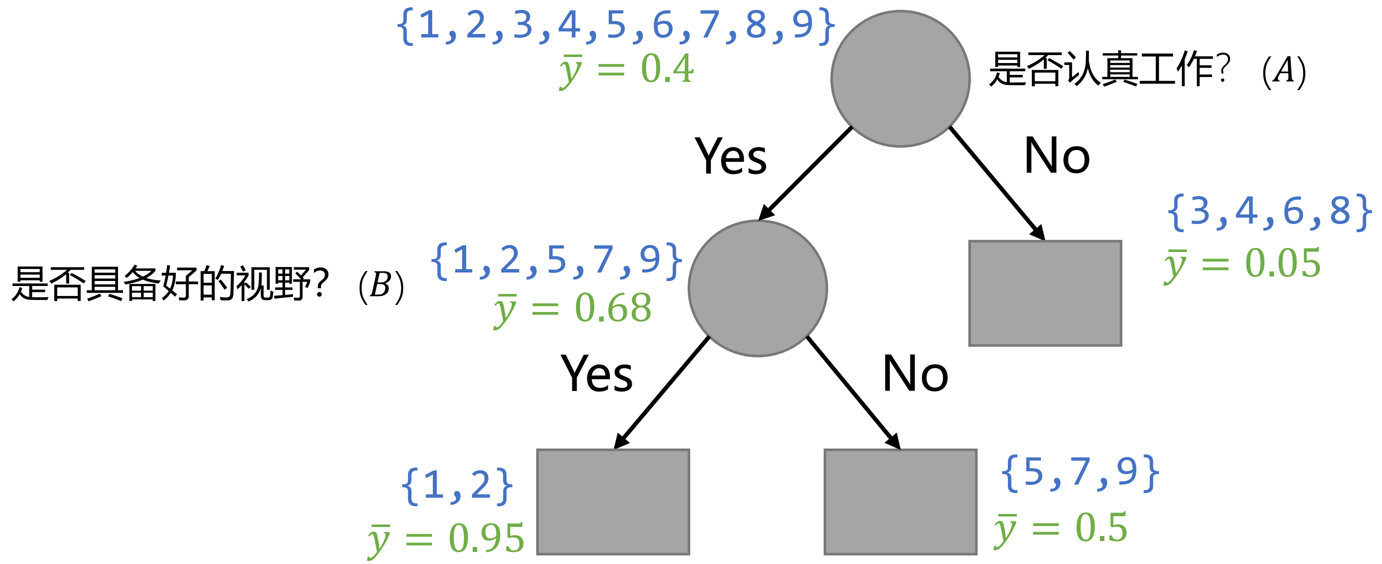
- 数据的标签可以连续吗？
  - 是否是一个好的科学家 -> 一个研究水平多高的科学家

序号	认真工作 A	视野 B	喜欢香蕉 C	好科学家 y
1	√	√	√	√
2	√	√	×	√
3	√	×	√	×
4	×	×	×	×
5	×	√	×	×
6	×	×	√	×
7	×	√	×	×
8	√	×	√	×
9	×	√	×	×

序号	认真工作 A	视野 B	喜欢香蕉 C	科学家水平打分 y
1	√	√	√	1
2	√	√	×	0.9
3	×	√	√	0.1
4	×	×	×	0
5	√	×	×	0.5
6	×	×	√	0
7	√	×	×	0.5
8	×	√	√	0.1
9	√	×	×	0.5

标签的离散到连续

- 数据的标签可以连续吗？
  - 是否是一个好的科学家 -> 一个研究水平多高的科学家
- 分类问题 -> 回归问题
- 决策树 -> 回归树



序号	认真工作 A	视野 B	喜欢香蕉 C	科学家水平打分 y
1	√	√	√	1
2	√	√	×	0.9
3	×	√	√	0.1
4	×	×	×	0
5	√	×	×	0.5
6	×	×	√	0
7	√	×	×	0.5
8	×	√	√	0.1
9	√	×	×	0.5

注： {1,2...}表示data point序号 $\bar{y}$表示该节点预测的水平打分均值

标签的离散到连续

- 数据的标签可以连续吗?
  - 是否是一个好的科学家 -> 一个研究水平多高的科学家
- 分类问题 -> 回归问题
- 决策树 -> 回归树

	离散标签（分类问题）	连续标签（回归问题）
离散属性	决策树	回归树
连续属性	决策树+离散化技术	回归树+离散化技术

回归树

- 保留决策树的结构
- 通过 L2 Loss 重新定义集合D的纯度
  - 离散：信息熵、基尼指数
  - 连续：

样本均值 $$\bar{y}_D = \frac{1}{|D|}\sum_{j\in D} y_j$$

样本方差（未除以$|D|$） $$L(D) = \sum_{j\in D}(y_j - \bar{y}_D)^2 \qquad L(D,A) = \sum_{i\in[m]} L(D^{A=a_i})$$

$$A_* = \underset{A\in F}{\operatorname{argmin}}\, L(D,A)$$

  - 其中，属性A的可能取值为$\{a_i \mid i \in [m]\}$，$D^{A=a_i}$表示集合D中样本属性A取a_i的子集。$\bar{y}_D$作为某节点标签的平均值，也表示该节点的预测标签。

回归树

- 对于该例，在根节点选择属性时：$D = \{\mathrm{data}_1 \sim \mathrm{data}_9\}$

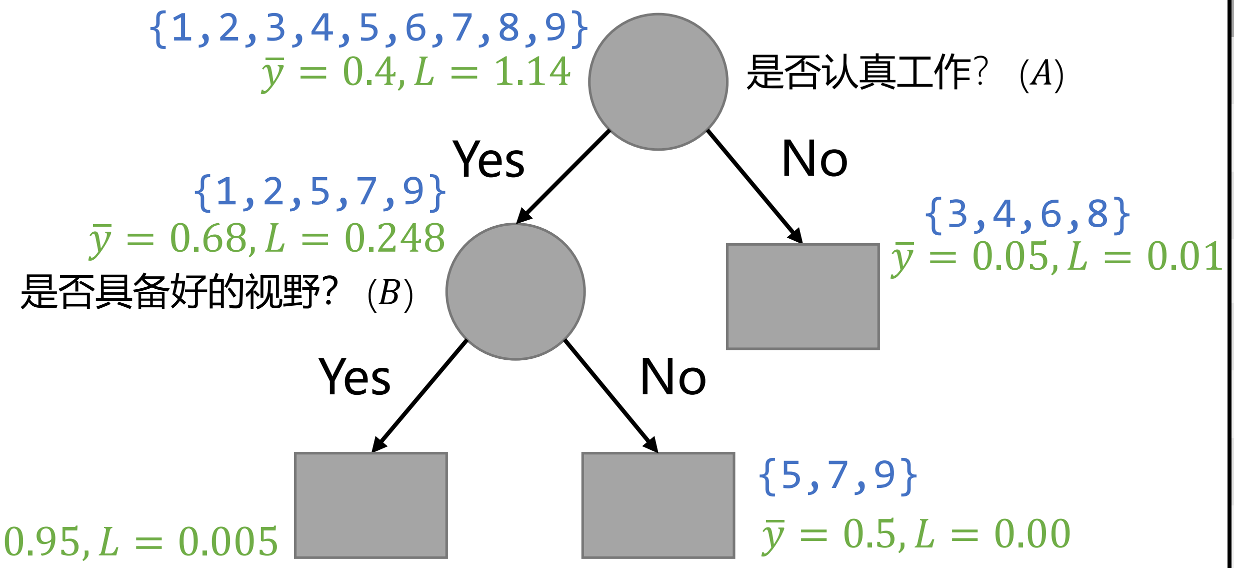
$$L(D) = (1-0.4)^2 + (0.9-0.4)^2 + 3\times(0.5-0.4)^2 + 2\times(0.1-0.4)^2 + 2\times(0-0.4)^2 = 1.14$$

$$L(D,A) = (1-0.68)^2 + (0.9-0.68)^2 + 3\times(0.5-0.68)^2 + 2\times(0.1-0.05)^2 + 2\times(0-0.05)^2 = \mathbf{0.258}$$

$$L(D,B) = 1.0275$$

$$L(D,C) = 1.068$$

- 因此选择属性A作为划分属性。



序号	认真工作 A	视野 B	喜欢香蕉 C	科学家水平打分 y
1	√	√	√	1
2	√	√	×	0.9
3	×	√	√	0.1
4	×	×	×	0
5	√	×	×	0.5
6	×	×	√	0
7	√	×	×	0.5
8	×	√	√	0.1
9	√	×	×	0.5

注：

{1,2...}表示data point序号

$\bar{y}$表示该节点预测的水平打分， L表示Loss值

树模型训练伪代码

Algorithm 1 How to build a decison tree

Input: Dataset D, Feature set $\mathcal{F}$

Output: Root Node T = Build_Decision_Tree($D, \mathcal{F}$)

```
function Build_Decision_Tree(D, F)
    if F = ∅ Or |D| = 1 then
        S ← Purity_Score(D)
        return Leaf_Node N = Make_Node(D, S, NULL)
    end if
    S_best ← NULL, A_best ← NULL
    for A ∈ F do
        S = Purity_Score(D, A)
        (S_best, A_best) ← Better_Score(S_best, S, A_best, A)
    end for
    Array_D ← Partition_Dataset(D, A_best)
    Node N = Make_Node(D, S_best, A_best)
    for i = 0 → |Array_D| do
        T ← Build_Decision_Tree(Array_D[i], F − A_best)
        Node N = Add_Child(N, T)
    end for
    return N
end function

function Purity_Score(D, A)
    return Corresponding Purity_Score formula with D and A(omited if NULL)
end function
```

终止条件：属性集F为空，或只剩一个样本

也可以用一些提前终止条件，如深度达到某一上限、或纯度分数好于某一阈值

遍历所有当前属性集中的属性，计算纯度分数，选择分数最好的属性A

按选择的属性划分样本集D, 在每个子集$D^{A=a_i}$上递归使用$F-A$构建决策树

树模型训练伪代码

```
function BETTER_SCORE(S_best, S, A_best, A)
    if S_best = NULL then
        return (S, A)
    end if
    if S is better than S_best according to Purity_Score formula then
        return (S, A)
    else
        return (S_best, A_best)
    end if
end function

function PARTITION_DATASET(D, A)
    return Array of D partitioned by A
end function

function MAKE_NODE(D, S, A)
    return Node N with current dataset D, Purity_Score S, partition feature A(omited if NULL)
end function

function ADD_CHILD(N, T)
    return Node N with child T
end function
```

集成学习(ensemble learning)

- 集成学习 = 多个个体学习器 + 结合策略
- 个体学习器
  - 例子：线性模型，决策树
  - 要求： “好而不同”
- 例子：随机森林

随机森林(Random forest)

- 个体学习器：决策树
- 结合策略：样本扰动+属性扰动
- 步骤：
  - 对于每一个决策树，对训练集进行随机采样得到一个独立的训练集（样本扰动）
  - 对于每一个决策树，只选择数据集中的一部分样本特征进行子树划分训练（属性扰动）
  - 最后，用所有训练好的决策树的平均预测（如多数类）作为对测试样本的输出

总结

- **决策树**
  - 比线性模型决策逻辑更复杂：非线性分类边界
  - 划分准则：信息增益、增益率、基尼系数
  - 连续属性（离散化，二分法）
  - 连续标签（回归树，L2 Loss衡量纯度）
  - 随机森林（集成多棵决策树）
- **下节课：神经网络**

谢谢

北京大學
PEKING UNIVERSITY