ID3 AIの例子

例子：根据天气决定是否打网球

假设我们有以下数据集，记录了不同天气状况下，人们是否去打网球的历史记录：

Day	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

目标：构建一个决策树，根据 `Outlook`，`Temperature`，`Humidity`，`Wind` 这四个特征来预测 `Play Tennis?`。

ID3 算法核心：信息增益

ID3 算法的核心是**选择信息增益最大的特征**作为当前节点的划分标准。信息增益基于**熵**的概念。

1. 计算公式

- **熵 (Entropy)**：度量样本集合纯度的指标。
  - 对于一个有 p 个正样本和 n 个负样本的集合 S，其熵定义为：
$$I(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$
- **信息增益 (Information Gain)**：使用某个特征 A 对集合 S 进行划分后，熵的减少量。增益越大，意味着用该特征划分后纯度提升越高。
  - $Gain(S, A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$
  - 其中 $Values(A)$ 是特征 A 的所有取值集合，S_v 是 S 中特征 A 的值为 v 的样本子集。

构建决策树的过程

我们的根节点包含所有 14 个样本。我们先计算根节点的熵。

Step 1: 计算整个数据集的熵

- 总样本数：14
- 正样本 (`Yes`)：9
- 负样本 (`No`)：5
- 熵 $I(S)$：$I(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \approx 0.940$

Step 2: 计算每个特征的信息增益

我们需要分别计算 `Outlook`，`Temperature`，`Humidity`，`Wind` 的信息增益。

a) 计算特征 `Outlook` 的信息增益

`Outlook` 有三个取值： `Sunny`，`Overcast`，`Rain`。

- `Sunny`：样本数 = 5。其中 `Yes` =2, `No` =3。
  - 熵 $I(S_{Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.971$
- `Overcast`：样本数 = 4。其中 `Yes` =4, `No` =0。
  - 熵 $I(S_{Overcast}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$ (规定 $0\log_2 0 = 0$)
- `Rain`：样本数 = 5。其中 `Yes` =3, `No` =2。
  - 熵 $I(S_{Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$

信息增益：

$$Gain(S, Outlook) = I(S) - \left[\frac{5}{14} \times I(S_{Sunny}) + \frac{4}{14} \times I(S_{Overcast}) + \frac{5}{14} \times I(S_{Rain})\right]$$

$$Gain(S, Outlook) = 0.940 - \left[\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971\right]$$

$$Gain(S, Outlook) = 0.940 - [0.347 + 0 + 0.347] = 0.940 - 0.694 = 0.246$$

b) 计算特征 `Humidity` 的信息增益

`Humidity` 有两个取值： `High` , `Normal` 。

- `High` : 样本数 = 7。其中 `Yes` =3, `No` =4。
  - 熵 $I(S_{High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} \approx 0.985$
- `Normal` : 样本数 = 7。其中 `Yes` =6, `No` =1。
  - 熵 $I(S_{Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} \approx 0.592$

信息增益：

$$Gain(S, Humidity) = I(S) - \left[\frac{7}{14} \times I(S_{High}) + \frac{7}{14} \times I(S_{Normal})\right]$$

$$Gain(S, Humidity) = 0.940 - \left[\frac{1}{2} \times 0.985 + \frac{1}{2} \times 0.592\right]$$

$$Gain(S, Humidity) = 0.940 - [0.4925 + 0.296] = 0.940 - 0.7885 = 0.1515$$

c) 计算特征 `Wind` 的信息增益

`Wind` 有两个取值： `Weak` , `Strong` 。

- `Weak` : 样本数 = 8。其中 `Yes` =6, `No` =2。
  - 熵 $I(S_{Weak}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \approx 0.811$
- `Strong` : 样本数 = 6。其中 `Yes` =3, `No` =3。
  - 熵 $I(S_{Strong}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.0$

信息增益：

$$Gain(S, Wind) = I(S) - \left[\frac{8}{14} \times I(S_{Weak}) + \frac{6}{14} \times I(S_{Strong})\right]$$

$$Gain(S, Wind) = 0.940 - \left[\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1.0\right]$$

$$Gain(S, Wind) = 0.940 - [0.463 + 0.429] = 0.940 - 0.892 = 0.048$$

d) 计算特征 `Temperature` 的信息增益

`Temperature` 有三个取值： `Hot` , `Mild` , `Cool` 。（计算过程类似，此处简化）

- `Hot` : 样本数=4, `Yes` =2, `No` =2。熵=1.0
- `Mild` : 样本数=6, `Yes` =4, `No` =2。熵≈0.918

- `Cool`：样本数=4，`Yes`=3，`No`=1。熵≈0.811

信息增益：

$$Gain(S, Temperature) = 0.940 - \left[\frac{4}{14} \times 1.0 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811\right] \approx 0.029$$

Step 3: 选择根节点

比较四个特征的信息增益：

- `Gain(Outlook)` = **0.246** (最大)
- `Gain(Humidity)` = 0.1515
- `Gain(Wind)` = 0.048
- `Gain(Temperature)` ≈ 0.029

因此，我们选择 `Outlook` 作为根节点进行划分。 此时的树结构如下：

```
代码块
1            Outlook
2          /    |    \
3     Sunny   Overcast    Rain
4     [2,3]    [4,0]     [3,2]
5    (No/Yes)  (Pure Yes) (No/Yes)
```

- `Overcast` 分支的样本已经是纯的（全部为 `Yes`），所以它直接成为一个**叶子节点**，决策为 `Yes`。
- `Sunny` 和 `Rain` 分支的样本不纯，需要进一步划分。

Step 4: 划分 `Sunny` 分支

现在我们在 `Outlook=Sunny` 的这个子集（5个样本）上重复上面的过程，但不再考虑已经用过的 `Outlook` 特征。

计算 `Sunny` 子集的熵 $I(S_{Sunny}) \approx 0.971$ (之前算过)。

分别计算 `Temperature`，`Humidity`，`Wind` 在这个子集上的信息增益。

（计算过程省略，与之前方法完全相同）

- `Gain(S_Sunny, Humidity)` 的信息增益会最大。
  - `Humidity=High` (3个样本，全是 `No`，熵=0)
  - `Humidity=Normal` (2个样本，全是 `Yes`，熵=0)
  - $Gain = 0.971 - [\frac{3}{5} \times 0 + \frac{2}{5} \times 0] = 0.971$

所以，我们选择 `Humidity` 来划分 `Sunny` 分支。划分后产生两个纯的叶子节点。

Step 5: 划分 `Rain` 分支

在 `Outlook=Rain` 的子集（5个样本，3Yes/2No，熵≈0.971）上计算。

（计算过程省略）

- `Gain(S_Rain, Wind)` 的信息增益会最大。
  - `Wind=Weak` (3个样本，全是 `Yes`，熵=0)
  - `Wind=Strong` (2个样本，全是 `No`，熵=0)
  - $Gain = 0.971 - [\frac{3}{5} \times 0 + \frac{2}{5} \times 0] = 0.971$

所以，我们选择 `Wind` 来划分 `Rain` 分支。划分后产生两个纯的叶子节点。

最终决策树

通过以上过程，我们得到了最终的决策树：

代码块

```
                Outlook
              /     |    \
         Sunny /   Overcast \   Rain
             /       |       \
       Humidity              Wind
        /   \                /  \
   High /     \ Normal    Weak/   \Strong
      /        \           /       \
    No         Yes       Yes        No
```

决策规则举例：

- 如果 `Outlook=Overcast` -> **Yes**
- 如果 `Outlook=Sunny` 且 `Humidity=High` -> **No**
- 如果 `Outlook=Sunny` 且 `Humidity=Normal` -> **Yes**
- 如果 `Outlook=Rain` 且 `Wind=Weak` -> **Yes**
- 如果 `Outlook=Rain` 且 `Wind=Strong` -> **No**