1 説明

決定木とは，節点に属性名，枝に属性値，葉にクラスが与えられた木である．与えられたデータを分類する際は，木の根から順に節点の属性に対応するデータの属性値を調べ，対応する属性値の枝を辿る．葉まで達したとき，そのデータを葉のクラスに分類する．

ID3 は，情報量に基づき分類力の高い属性を選択することにより，簡潔な決定木を生成することができる．

- エントロピー (平均情報量) の定義
  エントロピーは以下のように定義できる．

  $\mathrm{Entropy}(X) = -\sum_{x \in X} p(x) \log p(x)$

  ただし，X は事象の集合，x は事象，$p(x)$ は事象 x の生起確率である．

  以下では簡単のために，事象の代わりに，生起確率を直接記述した式も用いる．

  $e(p) = -p \log_2 p - (1-p) \log_2 (1-p)$
  e(0)=e(1)=0, e(1/4)=e(3/4)=0.8113, e(1/3)=e(2/3)=0.9183, e(1/2)=1

- Information Gain (情報利得，情報獲得 (量))
  事例集合 S をある属性 A で分類したときに獲得される情報量を information gain といい，$G(S, A)$ と表現する．

$$
\begin{aligned}
G(S, A) &= \mathrm{Entropy}(S) - \text{集合 } S \text{ を属性 } A \text{ で分類後の情報量} && (1)\\
&= \mathrm{Entropy}(S) - \sum_{v \in A} \frac{|S_{A=v}|}{|S|} \mathrm{Entropy}(S_{A=v}) && (2)\\
S_{A=v} &= \{s | s \in S \text{ and } s \text{ の属性 } A \text{ の値が } v\} && (3)
\end{aligned}
$$

  ただし，$|S|$,$|S_{A=v}|$ は集合 S,$S_{A=v}$ の要素数である．

- 分類属性の選択
  $G(S, A)$ が最大となる属性 A を S の分類属性として選択する．なお，$G(S, A)/|A|$ を基準とする場合もある．ここでは $G(S, A)$ を基準とし，$G(S, A)$ が同じ場合，属性値の数が少ない属性を用いる．

2 課題

以下の事例集合 D に対して，ID3 アルゴリズムにより，決定木を学習せよ．

- 訓練事例集合 D

No	Examples	Yes/No
1	< 晴, 暖, 並, 強, 温, 同 >	Yes
2	< 晴, 暖, 高, 強, 温, 同 >	Yes
3	< 雨, 寒, 高, 強, 温, 変 >	No
4	< 晴, 暖, 高, 強, 冷, 変 >	Yes
5	< 晴, 寒, 高, 弱, 冷, 変 >	No
6	< 雨, 暖, 並, 弱, 冷, 同 >	No

3 解答

- D における分類前の情報量
  分類前の情報量はその集合のエントロピーと等しい．Entropy(D)=e(3/6)=1

- D における情報利得と分類属性の選択

分類属性 A	分類後の情報量	情報利得 $G(D,A)$
天気	4/6 Entropy(天気=晴 in D) + 2/6 Entropy(天気=雨 in D) = 4/6 e(3/4) + 2/6 e(0/2) = 4/6 ∗ 0.8113 = 0.5409	0.4591
気温	4/6 Entropy(気温=暖 in D) + 2/6 Entropy(気温=寒 in D) = 4/6 e(3/4) + 2/6 e(0/2) = 4/6 ∗ 0.8113 = 0.5409	0.4591
湿度	2/6 Entropy(湿度=並 in D) + 4/6 Entropy(湿度=高 in D) = 2/6 e(1/2) + 4/6 e(2/4) = 2/6 + 4/6 = 1	0.0
風	4/6 Entropy(風=強 in D) + 2/6 Entropy(風=弱 in D) = 4/6 e(3/4) + 2/6 e(0/2) = 4/6 ∗ 0.8113 = 0.5409	0.4591
水温	3/6 Entropy(水温=温 in D) + 3/6 Entropy(水温=冷 in D) 3/6 e(2/3) + 3/6 e(1/3) = 0.9183	0.0817
予報	3/6 Entropy(予報=同 in D) + 3/6 Entropy(予報=変 in D) 3/6 e(2/3) + 3/6 e(1/3) = 0.9183	0.0817

  したがって，分類のための属性値としては，天気, 気温, 風 のいずれかを選べばよい．ここでは，属性値の数が少ない “気温” を選択したことにする．

- 気温 により D を分類後の集合

$$S_1 = D_{\text{気温=暖}} = \{D_1(Yes), D_2(Yes), D_4(Yes), D_6(No)\} \quad (4)$$

$$S_2 = D_{\text{気温=寒}} = \{D_3(No), D_5(No)\} \quad (5)$$

- S_1 における分類前の情報量
  Entropy(S_1)=e(3/4)=0.8113

- S_1 における情報利得と分類属性の選択

分類属性 A	分類後の情報量	情報利得 $G(S_1,A)$
天気	3/4 Entropy(天気=晴 in S_1) + 1/4 Entropy(天気=雨 in S_1) = 3/4 e(3/3) + 1/4 e(0/1) = 0.0	0.8113
湿度	2/4 Entropy(湿度=並 in S_1) + 2/4 Entropy(湿度=高 in S_1) = 2/6 e(1/2) + 4/6 e(2/2) = 2/6 = 0.3333	0.4780
風	3/4 Entropy(風=強 in S_1) + 1/4 Entropy(風=弱 in S_1) = 3/4 e(3/3) + 1/4 e(0/1) = 0.0	0.8113
水温	2/4 Entropy(水温=温 in S_1) + 2/4 Entropy(水温=冷 in S_1) 2/4 e(2/2) + 2/4 e(1/2) = 0.5	0.3113
予報	3/4 Entropy(予報=同 in S_1) + 1/4 Entropy(予報=変 in S_1) 3/4 e(2/3) + 1/4 e(1/1) = 3/4 ∗ 0.9183 = 0.6887	0.1226

  したがって，分類のための属性値としては，天気, 風 のいずれかを選べばよい．ここでは，属性値の数が少ない “風” を選択したことにする．

- 風 により S_1 を分類後の集合

$$S_{11} = S_{1\text{風=強}} = \{D_1(Yes), D_2(Yes), D_4(Yes)\} \quad (6)$$

$$S_{12} = S_{1\text{風=弱}} = \{D_6(No)\} \quad (7)$$

  S_{11}, S_{12} はいずれも正事例あるいは負事例のみなので，Yes あるいは No のラベルを付けて終了．

- S_2 における分類前の情報量
  Entropy(S_1)=e(0/2)=0
  S_2 いずれも負事例なので，No のラベルを付けて終了.
- 決定木