

機械学習 決定木

管理工学科
篠沢 佳久

資料の内容

- 決定木
 - 分類のための指標(特徴選択)
 - エントロピー, ゲイン, ジニ係数
- 決定木作成アルゴリズム
 - CART
- 回帰木
- 実習
 - Irisデータセット(クラス分類)
 - Bostonデータセット(回帰木)

決定木

決定木 (decision Tree)

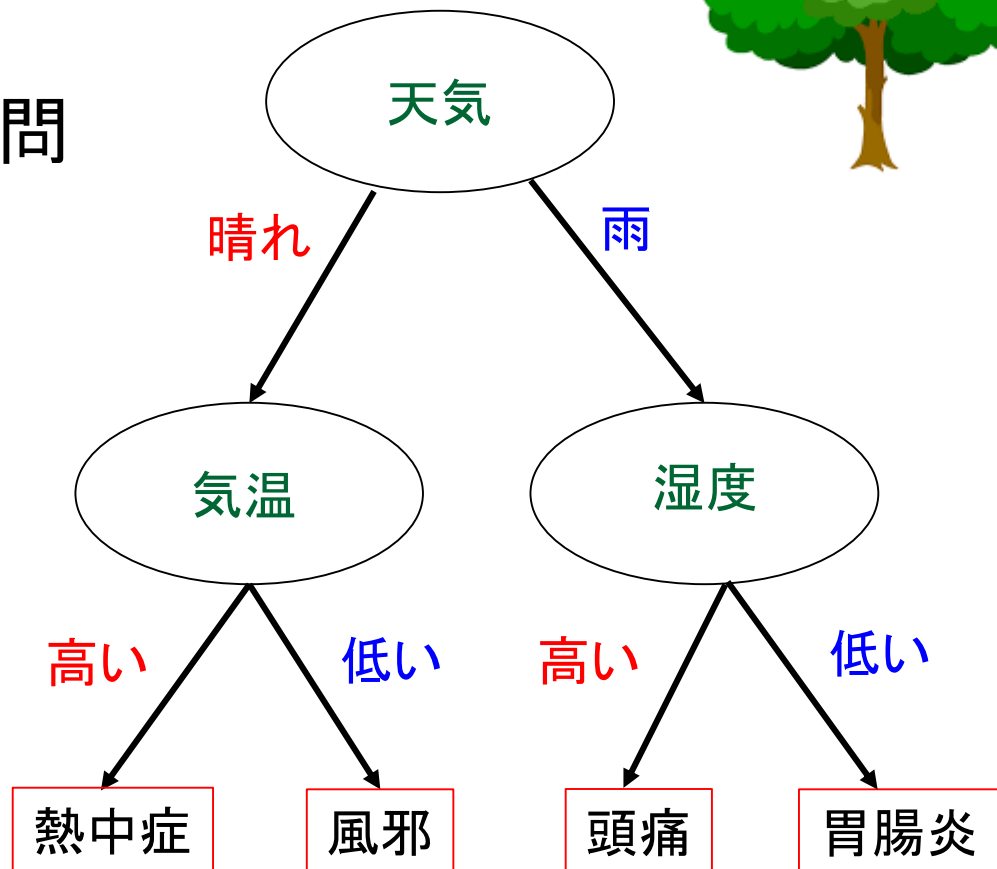


■ ノード

- 分類のための質問

■ 葉

- 分類結果



決定木の目的

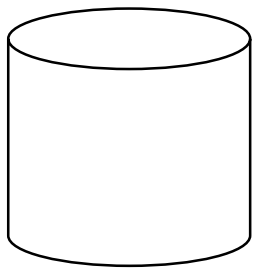
決定木

① 分類精度の向上

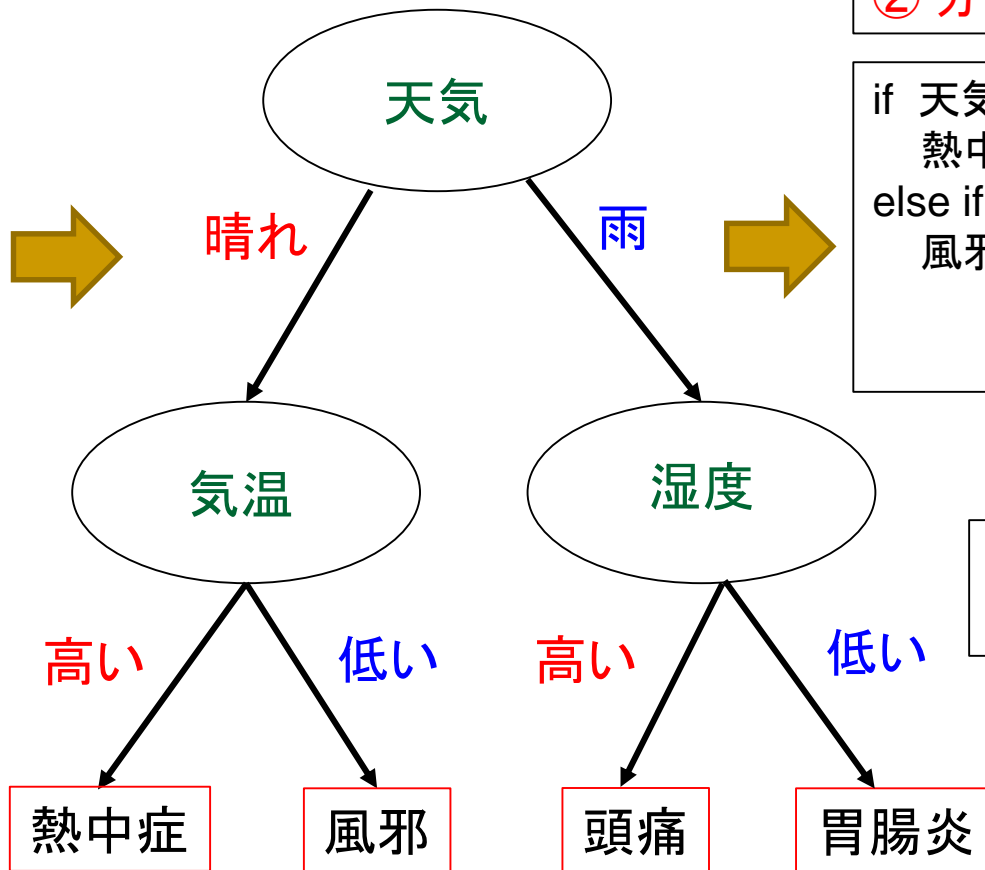
② 分類のためのルールの抽出

```
if 天気 == 晴れ and 気温 == 高い
    熱中症
else if 天気 == 晴れ and 気温 == 低い
    風邪
.
.
```

なぜ、その結果になるのか
説明が可能



データ



決定木の例

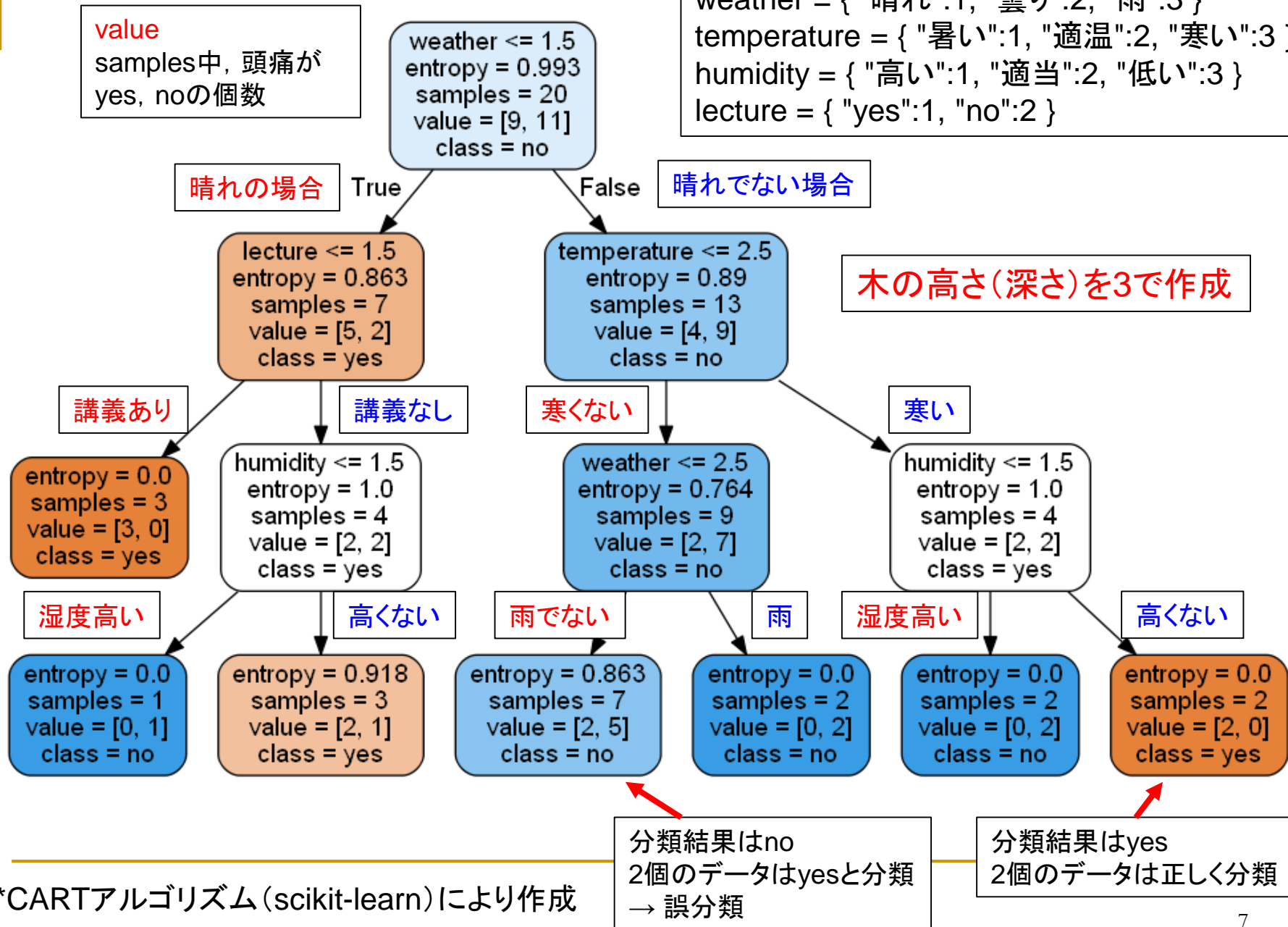
- 「頭痛」の日の判定
- 目的変数
 - 頭痛 … yes, no
- 説明変数(特徴量)
 - 天気 … 晴れ, 曇り, 雨
 - 気温 … 暑い, 適温, 寒い
 - 湿度 … 高い, 適当, 低い
 - 講義 … yes, no

学習データ

	天気	気温	湿度	講義	頭痛
1	晴れ	寒い	低い	yes	yes
2	曇り	暑い	高い	no	yes
3	晴れ	暑い	低い	no	no
4	雨	適温	高い	yes	no
5	雨	寒い	高い	no	no
6	晴れ	適温	適当	no	yes
7	曇り	暑い	低い	yes	no
8	雨	寒い	高い	no	no
9	曇り	適温	低い	yes	yes
10	曇り	適温	適当	no	no
11	晴れ	暑い	適当	yes	yes
12	晴れ	適温	高い	no	no
13	雨	暑い	低い	no	no
14	雨	寒い	適当	yes	yes
15	曇り	適温	低い	yes	no
16	晴れ	暑い	適当	no	yes
17	晴れ	寒い	高い	yes	yes
18	曇り	適温	高い	yes	no
19	曇り	適温	適当	no	no
20	雨	寒い	適当	no	yes

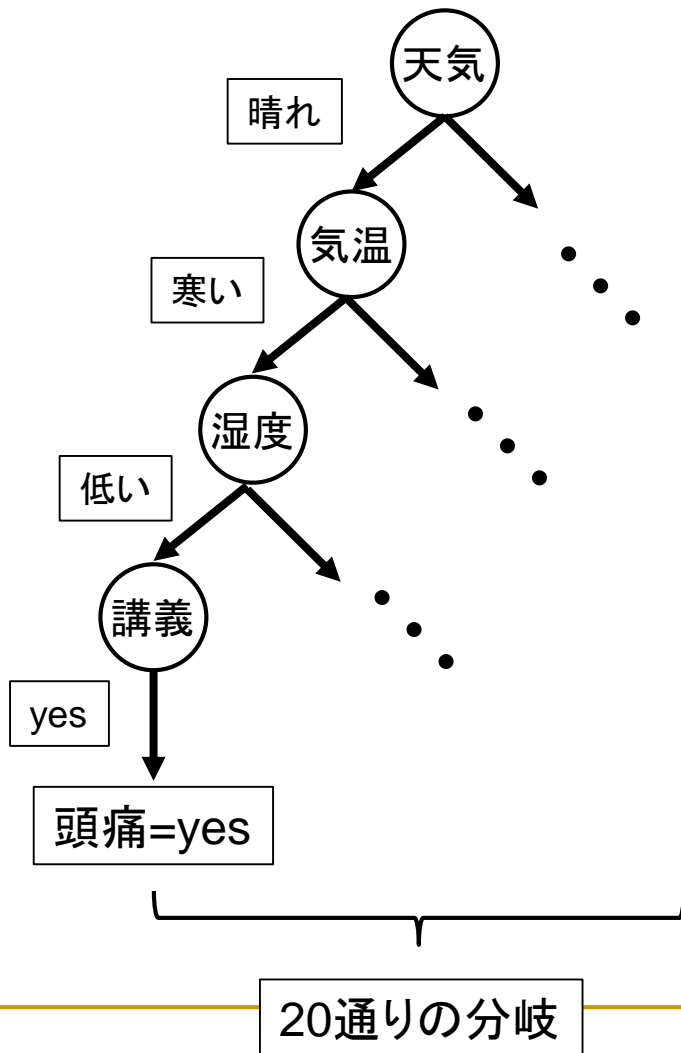
value
samples中, 頭痛が
yes, noの個数

weather = { "晴れ":1, "曇り":2, "雨":3 }
temperature = { "暑い":1, "適温":2, "寒い":3 }
humidity = { "高い":1, "適当":2, "低い":3 }
lecture = { "yes":1, "no":2 }



*CARTアルゴリズム (scikit-learn) により作成

「悪い(?)」決定木



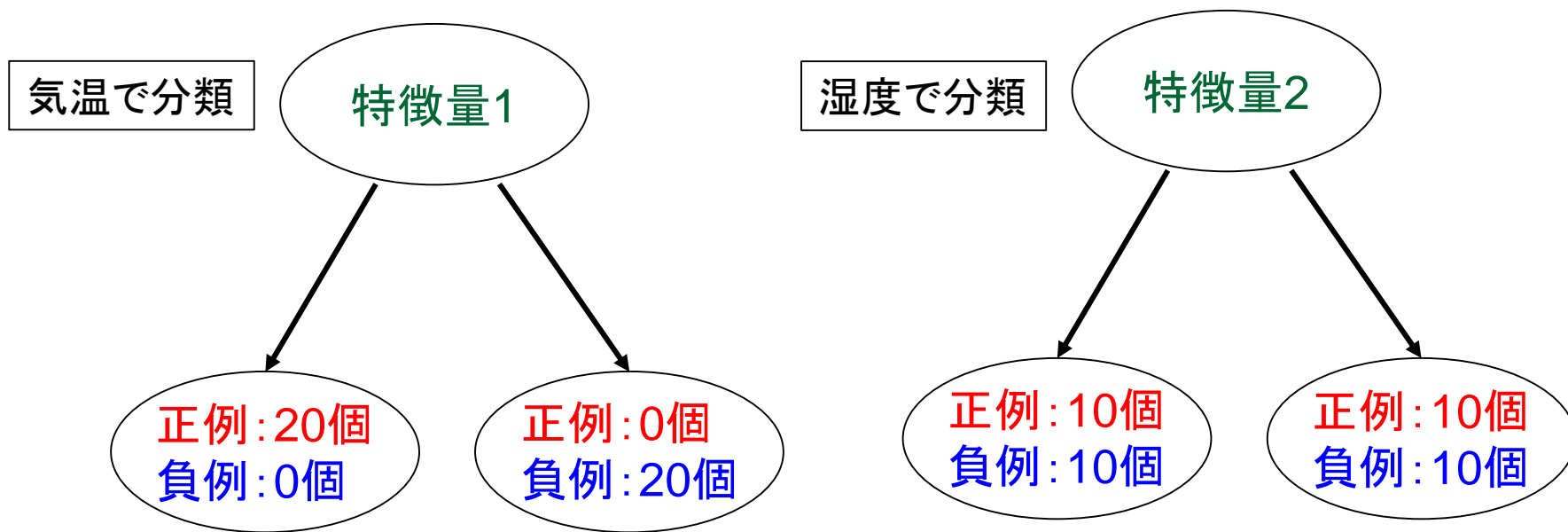
- 可能な分岐を全て考慮した場合
→ 学習データの精度は100%
→ 一般的なルールを抽出したことにはならない



学習データを「きれいに」分類し、
一般的なルールを抽出

「きれいに」分類するとは？

- 二値分類
- データ
 - 正例 20個, 負例 20個
 - 特徴量1, 特徴量2で分類した結果

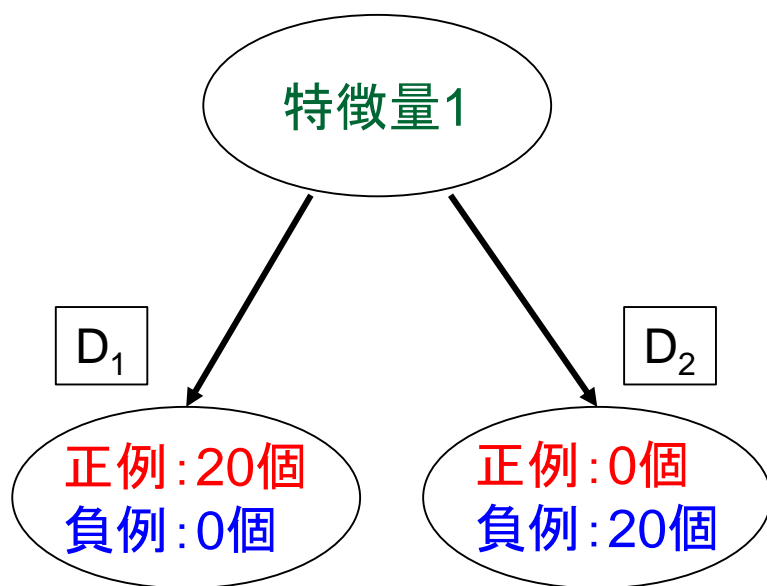


分類のための指標①

■ エントロピー(平均情報量)

$$E(D) = -P_+ \log P_+ - P_- \log P_-$$

P_+ : 分類後の正例の割合
 P_- : 分類後の負例の割合



D_1

$$E(D_1) = -\frac{20}{20} \log \frac{20}{20} - \frac{0}{20} \log \frac{0}{20} = 0$$

D_2

$$E(D_2) = -\frac{0}{20} \log \frac{0}{20} - \frac{20}{20} \log \frac{20}{20} = 0$$

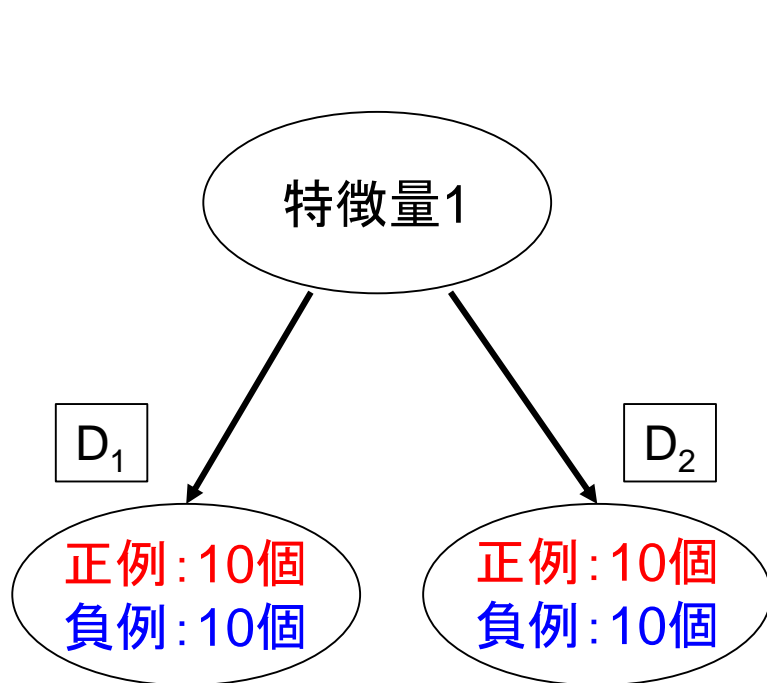
エントロピーが0に近いほど、「きれいに」
分類される

分類のための指標②

■ エントロピー(平均情報量)

$$E(D) = -P_+ \log P_+ - P_- \log P_-$$

P_+ : 分類後の正例の割合
 P_- : 分類後の負例の割合



D_1

$$E(D_1) = -\frac{10}{20} \log \frac{10}{20} - \frac{10}{20} \log \frac{10}{20} = 1$$

D_2

$$E(D_2) = -\frac{10}{20} \log \frac{10}{20} - \frac{10}{20} \log \frac{10}{20} = 1$$

エントロピーが1に近いほど、「きれいに」
分類されない

頭痛の場合

- データ数(20個)

- yes:9個, no:11個

- 開始時のエントロピー

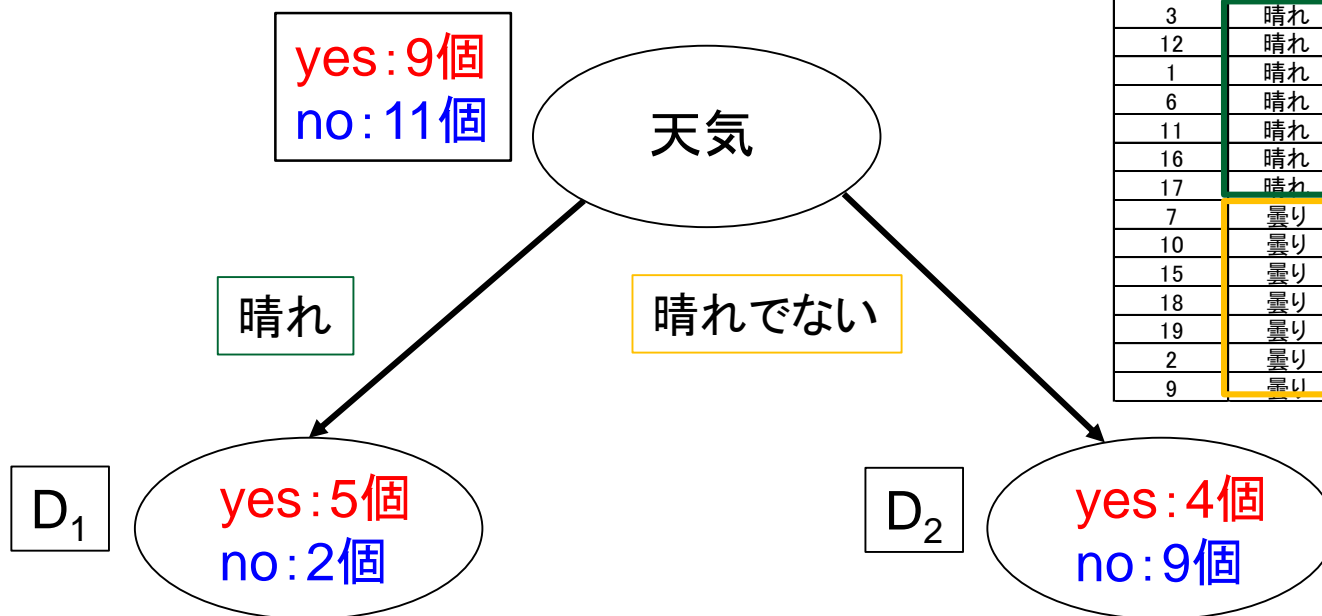
$$E(D) = -\frac{9}{20} \log \frac{9}{20} - \frac{11}{20} \log \frac{11}{20} = 0.992$$

- 分岐数は2*

	天気	気温	湿度	講義	頭痛
1	晴れ	寒い	低い	yes	yes
2	曇り	暑い	高い	no	yes
6	晴れ	適温	適当	no	yes
9	曇り	適温	低い	yes	yes
11	晴れ	暑い	適当	yes	yes
14	雨	寒い	適当	yes	yes
16	晴れ	暑い	適当	no	yes
17	晴れ	寒い	高い	yes	yes
20	雨	寒い	適当	no	yes
3	晴れ	暑い	低い	no	no
4	雨	適温	高い	yes	no
5	雨	寒い	高い	no	no
7	曇り	暑い	低い	yes	no
8	雨	寒い	高い	no	no
10	曇り	適温	適当	no	no
12	晴れ	適温	高い	no	no
13	雨	暑い	低い	no	no
15	曇り	適温	低い	yes	no
18	曇り	適温	高い	yes	no
19	曇り	適温	適当	no	no

*分岐数を2以上にする場合, 特徴量が連続値の場合, 後ほど説明します

天気で分類した場合



	天気	気温	湿度	講義	頭痛
4	雨	適温	高い	yes	no
5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
13	雨	暑い	低い	no	no
14	雨	寒い	適当	yes	yes
20	雨	寒い	適当	no	yes
3	晴れ	暑い	低い	no	no
12	晴れ	適温	高い	no	no
1	晴れ	寒い	低い	yes	yes
6	晴れ	適温	適当	no	yes
11	晴れ	暑い	適当	yes	yes
16	晴れ	暑い	適当	no	yes
17	晴れ	寒い	高い	yes	yes
7	曇り	暑い	低い	yes	no
10	曇り	適温	適当	no	no
15	曇り	適温	低い	yes	no
18	曇り	適温	高い	yes	no
19	曇り	適温	適当	no	no
2	曇り	暑い	高い	no	yes
9	曇り	適温	低い	yes	yes

晴れ

$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

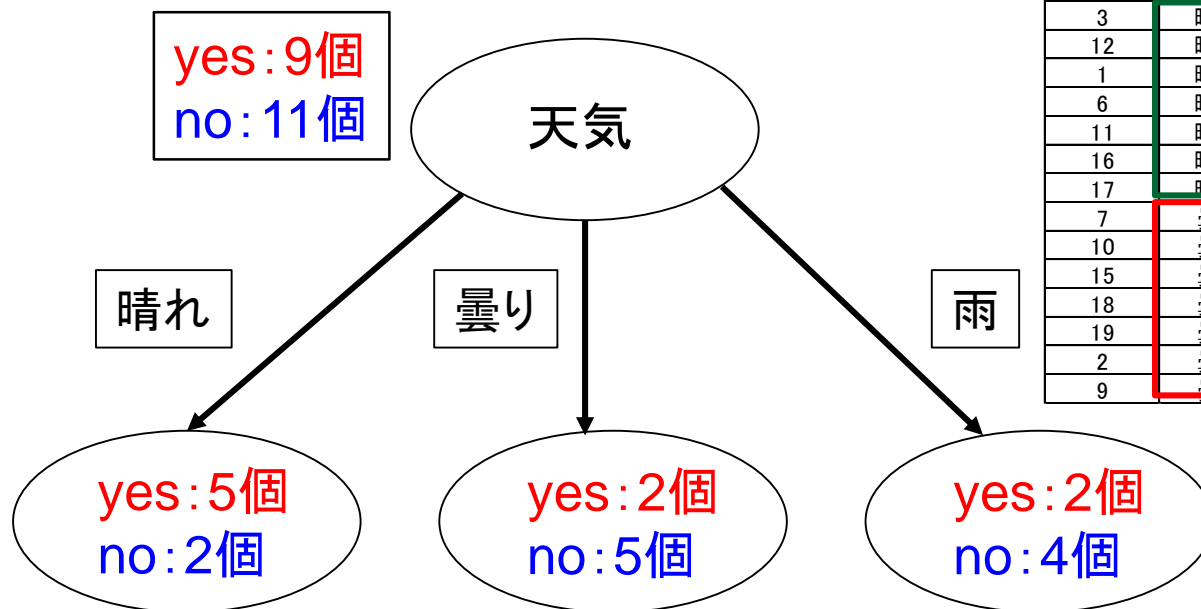
晴れでない

$$E(D_2) = -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} = 0.890$$

*「晴れと晴れでない」以外に、「雨と雨でない」、「曇りと曇りでない」とも分岐できます
分岐については後ほど、CARTアルゴリズムで説明します

分岐数は3も可能

	天気	気温	湿度	講義	頭痛
4	雨	適温	高い	yes	no
5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
13	雨	暑い	低い	no	no
14	雨	寒い	適当	yes	yes
20	雨	寒い	適当	no	yes
3	晴れ	暑い	低い	no	no
12	晴れ	適温	高い	no	no
1	晴れ	寒い	低い	yes	yes
6	晴れ	適温	適当	no	yes
11	晴れ	暑い	適当	yes	yes
16	晴れ	暑い	適当	no	yes
17	晴れ	寒い	高い	yes	yes
7	曇り	暑い	低い	yes	no
10	曇り	適温	適当	no	no
15	曇り	適温	低い	yes	no
18	曇り	適温	高い	yes	no
19	曇り	適温	適当	no	no
2	曇り	暑い	高い	no	yes
9	曇り	適温	低い	yes	yes



晴れ

$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

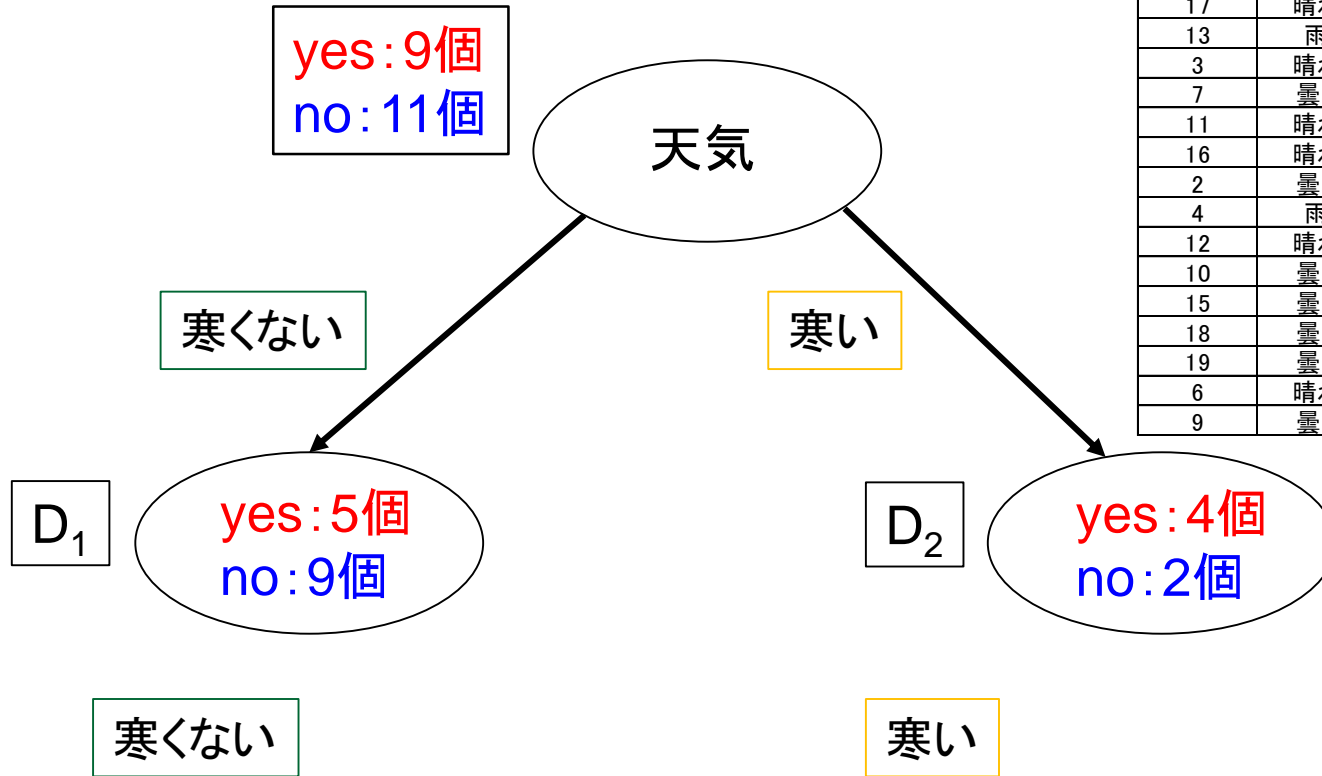
雨

$$E(D_3) = -\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6} = 0.918$$

曇り

$$E(D_2) = -\frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7} = 0.863$$

気温で分類した場合



	天気	気温	湿度	講義	頭痛
5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
14	雨	寒い	適当	yes	yes
20	雨	寒い	適当	no	yes
1	晴れ	寒い	低い	yes	yes
17	晴れ	寒い	高い	yes	yes
13	雨	暑い	低い	no	no
3	晴れ	暑い	低い	no	no
7	曇り	暑い	低い	yes	no
11	晴れ	暑い	適当	yes	yes
16	晴れ	暑い	適当	no	yes
2	曇り	暑い	高い	no	yes
4	雨	適温	高い	yes	no
12	晴れ	適温	高い	no	no
10	曇り	適温	適当	no	no
15	曇り	適温	低い	yes	no
18	曇り	適温	高い	yes	no
19	曇り	適温	適当	no	no
6	晴れ	適温	適当	no	yes
9	曇り	適温	低い	yes	yes

$$E(D_1) = -\frac{5}{14} \log \frac{5}{14} - \frac{9}{14} \log \frac{9}{14} = 0.940$$

$$E(D_2) = -\frac{4}{6} \log \frac{4}{6} - \frac{2}{6} \log \frac{2}{6} = 0.918$$

湿度で分類した場合

	天気	気温	湿度	講義	頭痛
5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
4	雨	適温	高い	yes	no
12	晴れ	適温	高い	no	no
18	曇り	適温	高い	yes	no
17	晴れ	寒い	高い	yes	yes
2	曇り	暑い	高い	no	yes
13	雨	暑い	低い	no	no
3	晴れ	暑い	低い	no	no
7	曇り	暑い	低い	yes	no
15	曇り	適温	低い	yes	no
1	晴れ	寒い	低い	yes	yes
9	曇り	適温	低い	yes	yes
10	曇り	適温	適当	no	no
19	曇り	適温	適当	no	no
14	雨	寒い	適当	yes	yes
20	雨	寒い	適当	no	yes
11	晴れ	暑い	適当	yes	yes
16	晴れ	暑い	適当	no	yes
6	晴れ	適温	適当	no	yes

yes: 9個
no: 11個

湿度

高い

高くない

D₁

yes: 2個
no: 5個

D₂

yes: 7個
no: 6個

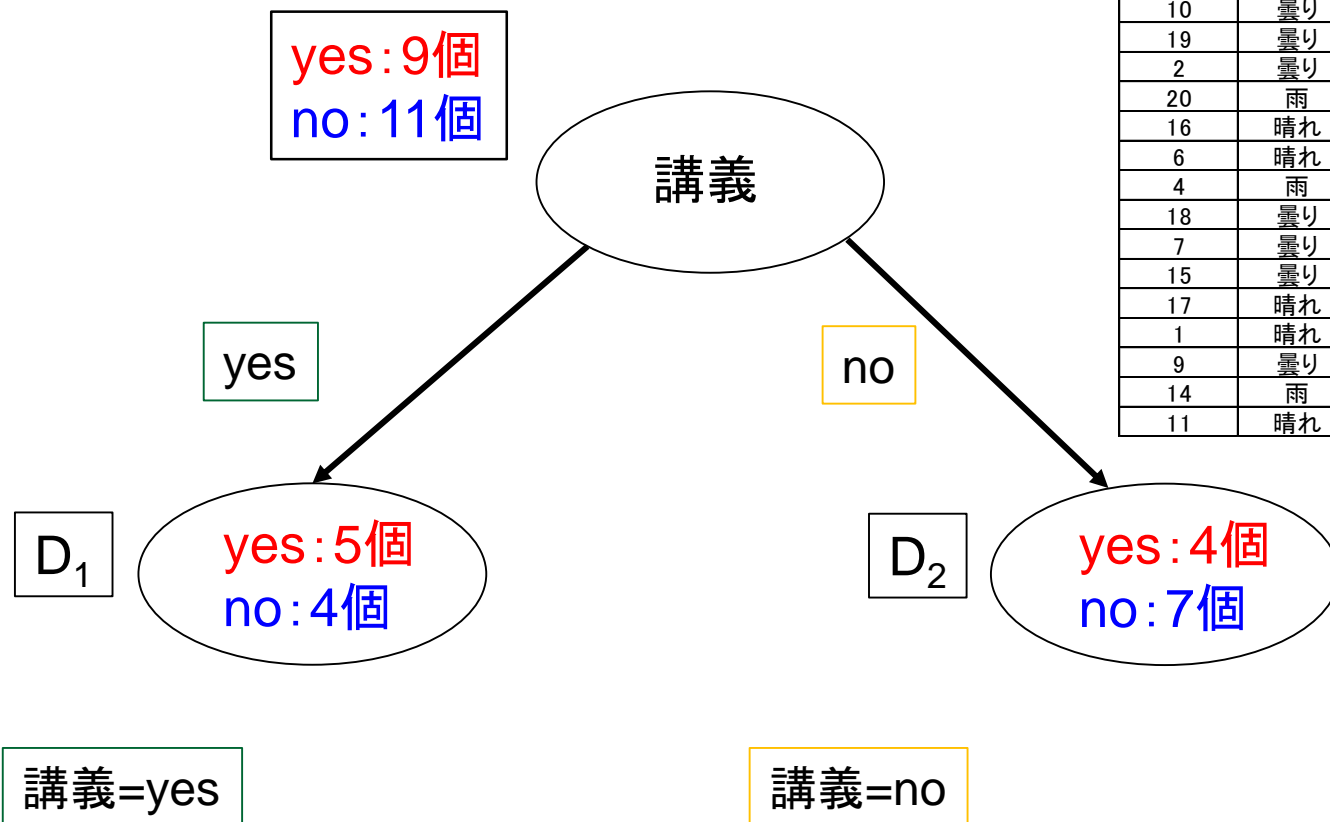
高い

高くない

$$E(D_1) = -\frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7} = 0.863$$

$$E(D_2) = -\frac{7}{13} \log \frac{7}{13} - \frac{6}{13} \log \frac{6}{13} = 0.995$$

講義で分類した場合



	天気	気温	湿度	講義	頭痛
5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
12	晴れ	適温	高い	no	no
13	雨	暑い	低い	no	no
3	晴れ	暑い	低い	no	no
10	曇り	適温	適当	no	no
19	曇り	適温	適当	no	no
2	曇り	暑い	高い	no	yes
20	雨	寒い	適当	no	yes
16	晴れ	暑い	適当	no	yes
6	晴れ	適温	適当	no	yes
4	雨	適温	高い	yes	no
18	曇り	適温	高い	yes	no
7	曇り	暑い	低い	yes	no
15	曇り	適温	低い	yes	no
17	晴れ	寒い	高い	yes	yes
1	晴れ	寒い	低い	yes	yes
9	曇り	適温	低い	yes	yes
14	雨	寒い	適当	yes	yes
11	晴れ	暑い	適当	yes	yes

$$E(D_1) = -\frac{5}{9} \log \frac{5}{9} - \frac{4}{9} \log \frac{4}{9} = 0.991$$

$$E(D_2) = -\frac{4}{11} \log \frac{4}{11} - \frac{7}{11} \log \frac{7}{11} = 0.945$$

特徴の選択方法

- エントロピーが最も小さくなる特徴を選択
- ゲイン*

$$Gain(D, a) = E(D) - \sum_{c \in Value(a)} \frac{|D_c|}{|D|} E(D_c)$$

現在のエントロピー

分類後のエントロピー

a: 特徴量

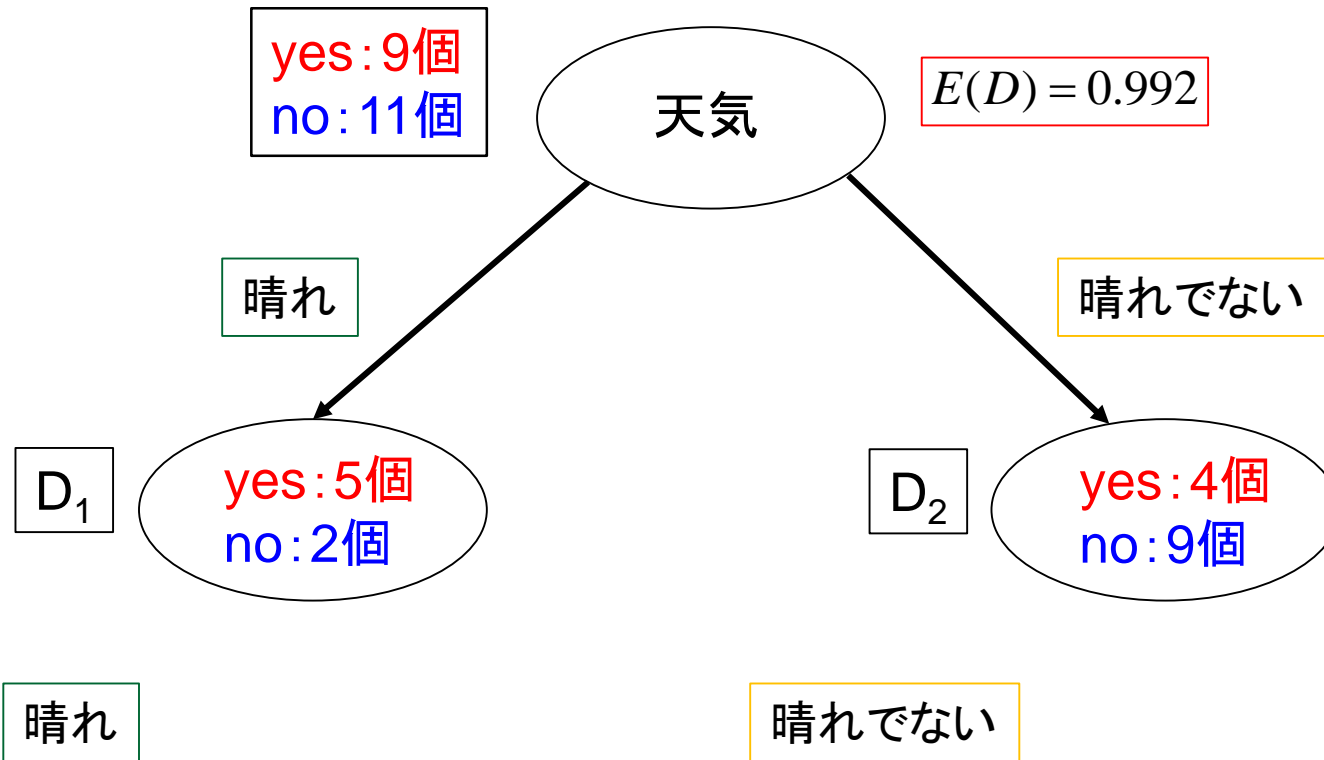
Value(a): 特徴量aの値

D_c : 特徴量aの値を持つデータの集合

- ゲインが最大となる特徴を選択

*情報利得とも呼ばれます

天気で分類した場合

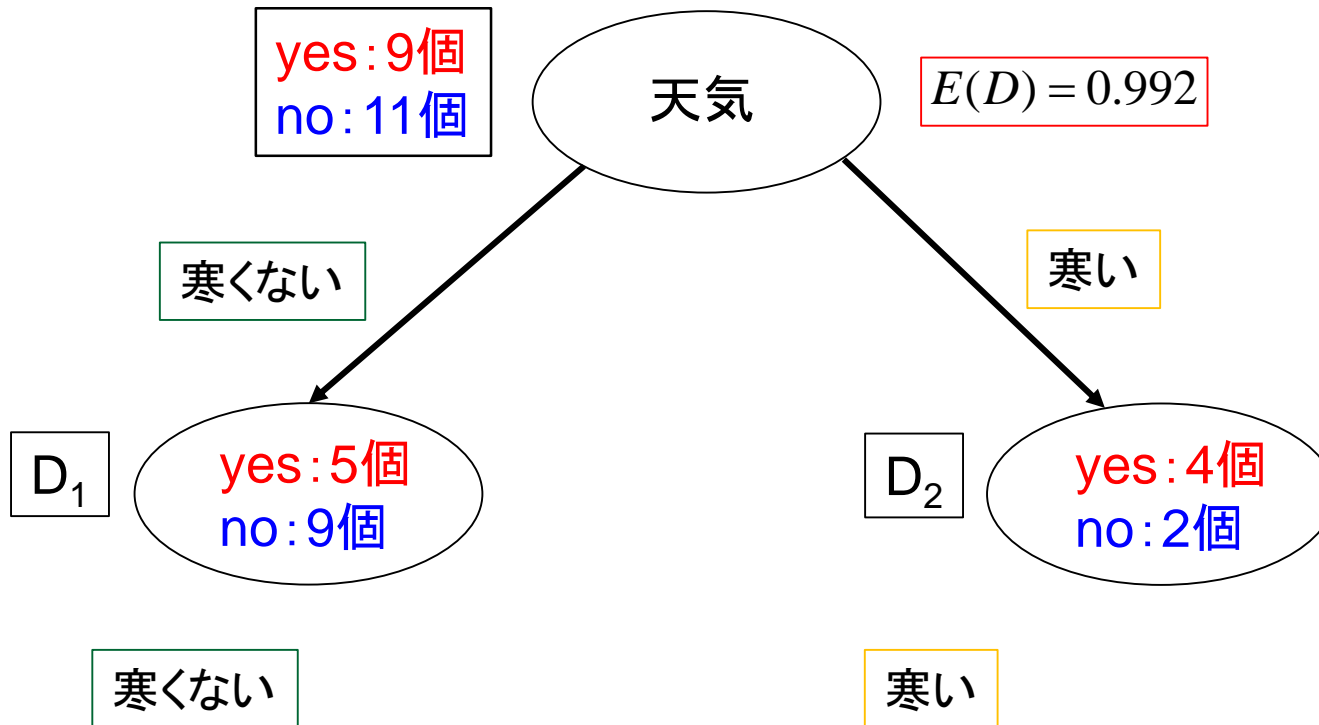


$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

$$E(D_2) = -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} = 0.890$$

$$\begin{aligned} \text{Gain}(D) &= E(D) - \frac{7}{20} E(D_1) - \frac{13}{20} E(D_2) \\ &= 0.992 - \frac{7}{20} \times 0.863 - \frac{13}{20} \times 0.890 = 0.111 \end{aligned}$$

気温で分類した場合

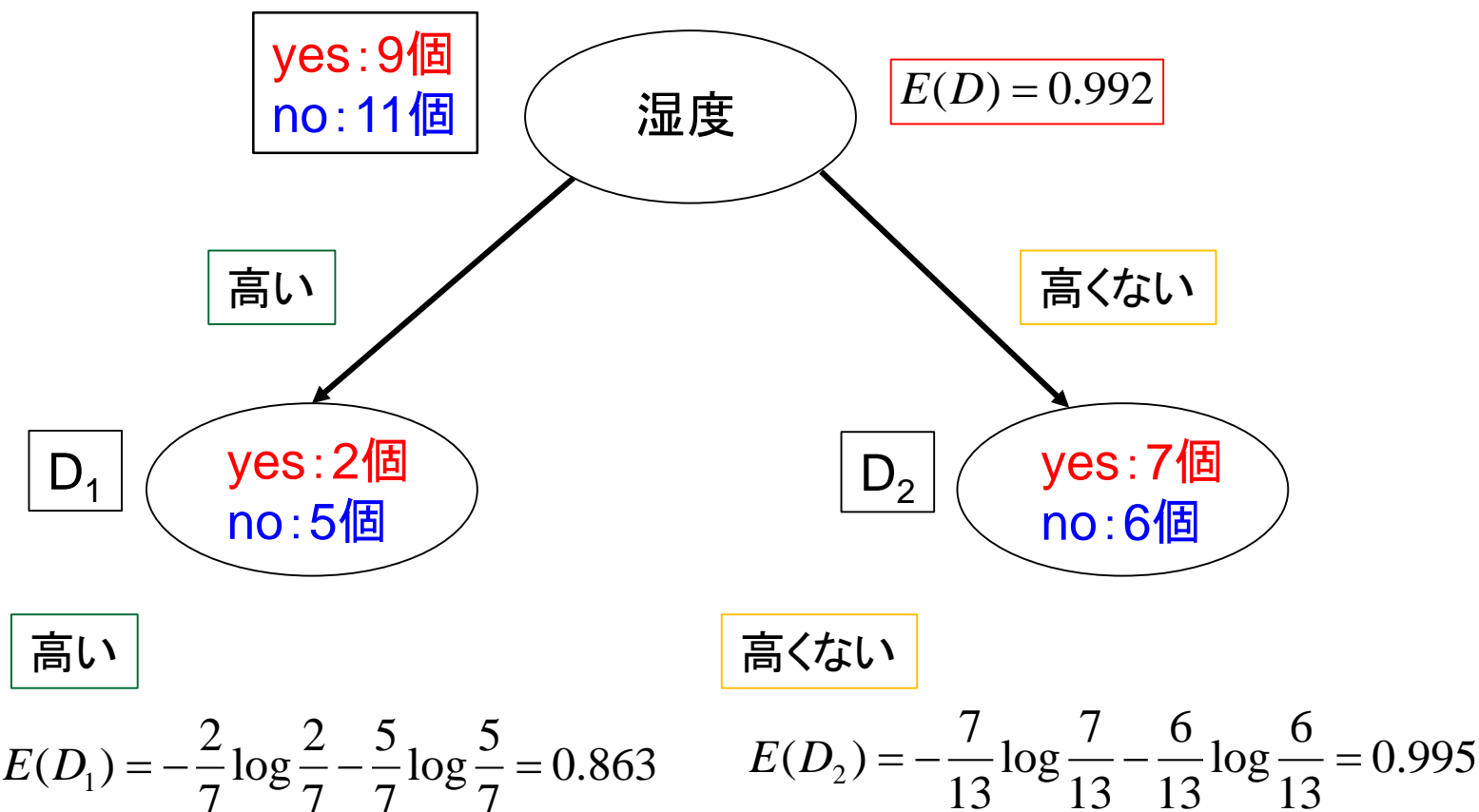


$$E(D_1) = -\frac{5}{14} \log \frac{5}{14} - \frac{9}{14} \log \frac{9}{14} = 0.940$$

$$E(D_2) = -\frac{4}{6} \log \frac{4}{6} - \frac{2}{6} \log \frac{2}{6} = 0.918$$

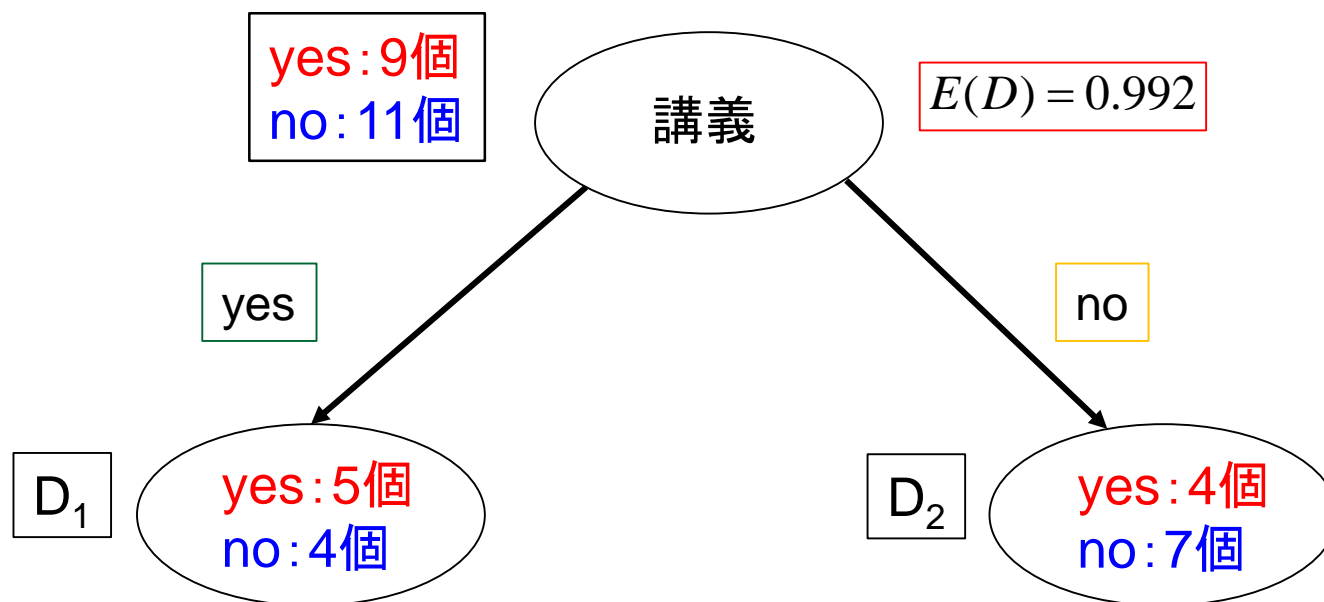
$$\begin{aligned} \text{Gain}(D) &= E(D) - \frac{14}{20} E(D_1) - \frac{6}{20} E(D_2) \\ &= 0.992 - \frac{14}{20} \times 0.979 - \frac{6}{20} \times 0.811 = 0.059 \end{aligned}$$

湿度で分類した場合



$$\begin{aligned} \text{Gain}(D) &= E(D) - \frac{7}{20} E(D_1) - \frac{13}{20} E(D_2) \\ &= 0.992 - \frac{7}{20} \times 0.863 - \frac{13}{20} \times 0.995 = 0.043 \end{aligned}$$

講義で分類した場合



講義=yes

$$E(D_1) = -\frac{5}{9} \log \frac{5}{9} - \frac{4}{9} \log \frac{4}{9} = 0.991$$

講義=no

$$E(D_2) = -\frac{4}{11} \log \frac{4}{11} - \frac{7}{11} \log \frac{7}{11} = 0.945$$

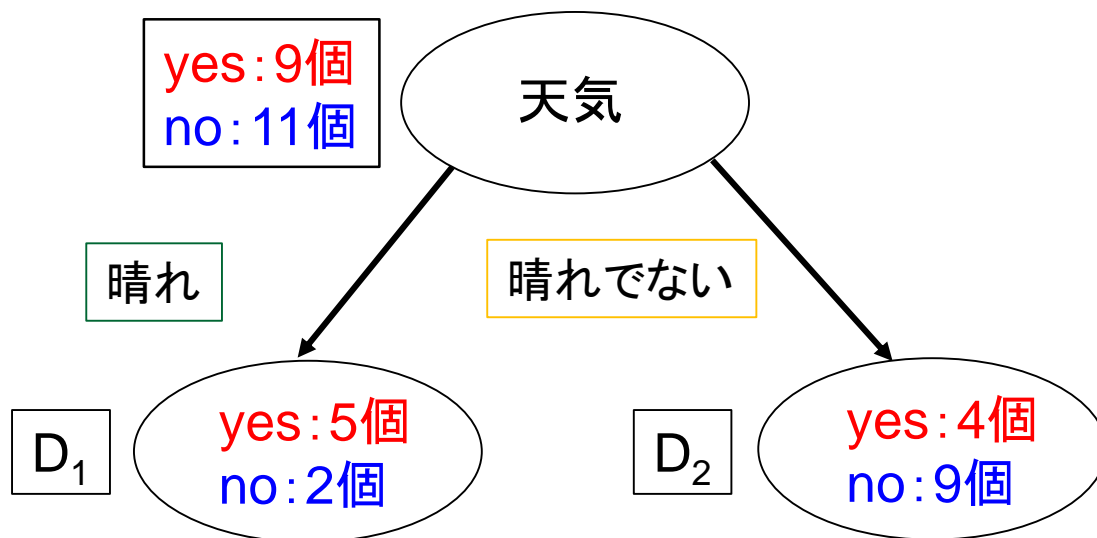
$$\begin{aligned} \text{Gain}(D) &= E(D) - \frac{9}{20} E(D_1) - \frac{11}{20} E(D_2) \\ &= 0.992 - \frac{9}{20} \times 0.991 - \frac{11}{20} \times 0.945 = 0.026 \end{aligned}$$

分類するための特徴選択

- 天気で分類した場合 → ゲイン = 0.111 最大
- 気温で分類した場合 → ゲイン = 0.059
- 湿度で分類した場合 → ゲイン = 0.043
- 講義で分類した場合 → ゲイン = 0.026



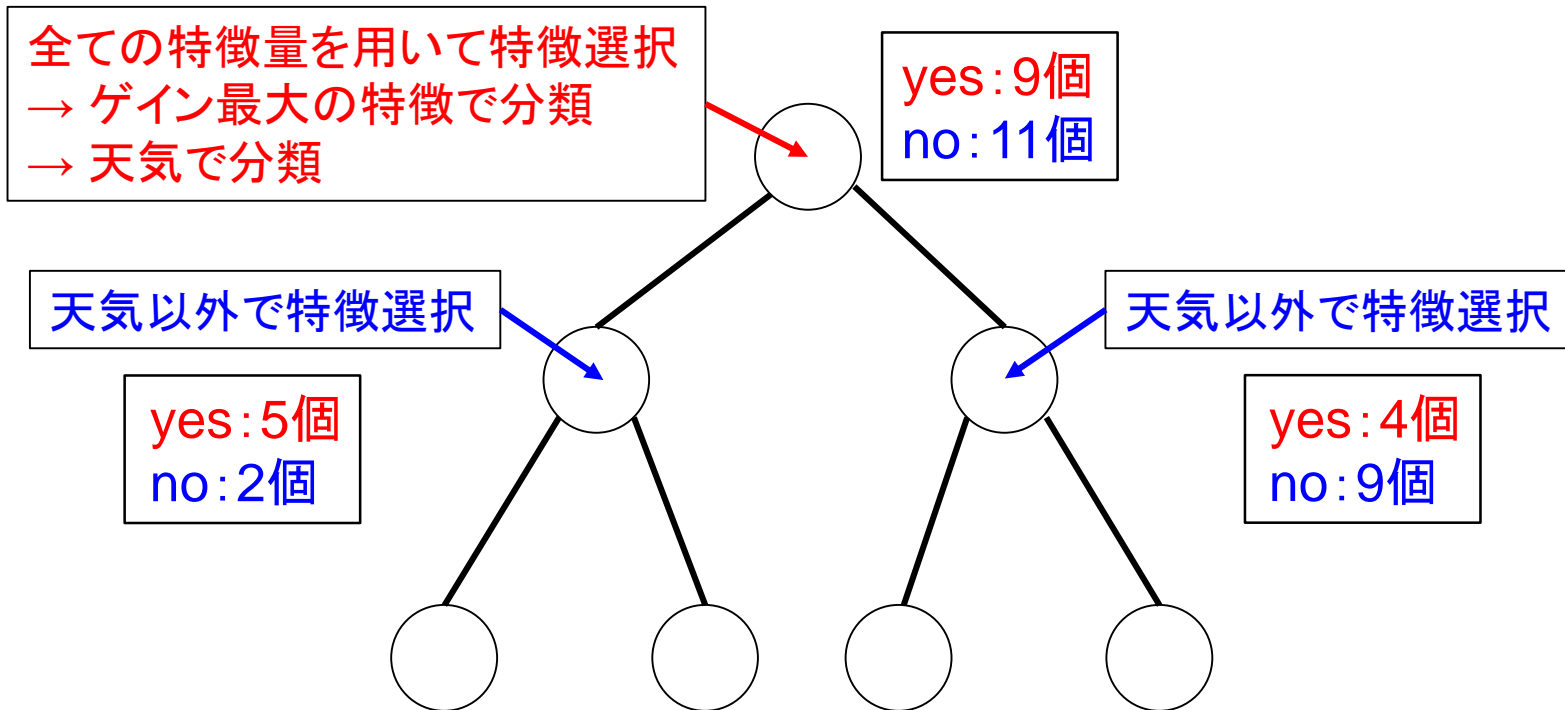
- 天気で分類



次の特徴選択は？

■ Greedyアルゴリズム

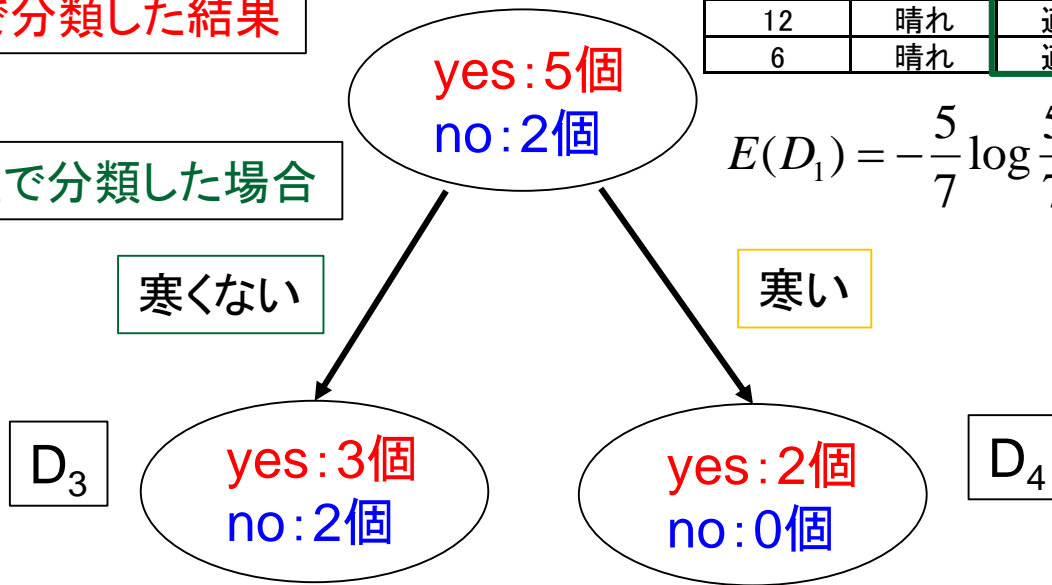
- 現時点でゲインが最大となる特徴を選択



さらに特徴選択①

晴れで分類した結果

気温で分類した場合



17	晴れ	寒い	高い	yes	yes
1	晴れ	寒い	低い	yes	yes
3	晴れ	暑い	低い	no	no
16	晴れ	暑い	適当	no	yes
11	晴れ	暑い	適当	yes	yes
12	晴れ	適温	高い	no	no
6	晴れ	適温	適当	no	yes

$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

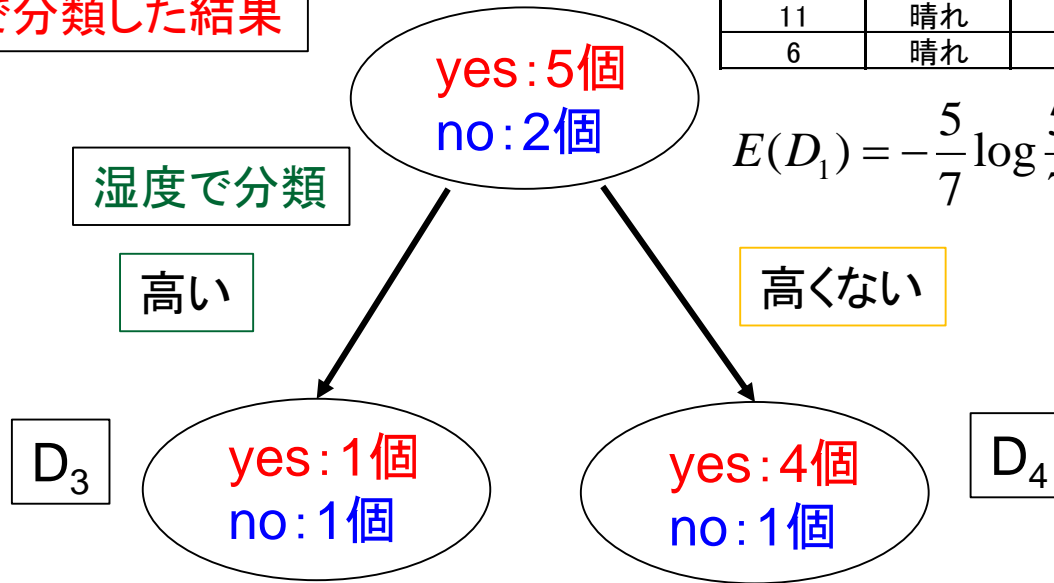
$$E(D_3) = -\frac{3}{5} \log \frac{3}{5} - \frac{2}{5} \log \frac{2}{5} = 0.971$$

$$E(D_4) = -\frac{2}{2} \log \frac{2}{2} - \frac{0}{2} \log \frac{0}{2} = 0$$

$$\begin{aligned}
 \text{Gain}(D_1) &= E(D_1) - \frac{5}{7} E(D_3) - \frac{2}{7} E(D_4) \\
 &= 0.863 - \frac{5}{7} \times 0.971 - \frac{2}{7} \times 0.0 = 0.169
 \end{aligned}$$

さらに特徴選択②

晴れで分類した結果



12	晴れ	適温	高い	no	no
17	晴れ	寒い	高い	yes	yes
3	晴れ	暑い	低い	no	no
1	晴れ	寒い	低い	yes	yes
16	晴れ	暑い	適当	no	yes
11	晴れ	暑い	適当	yes	yes
6	晴れ	適温	適当	no	yes

$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

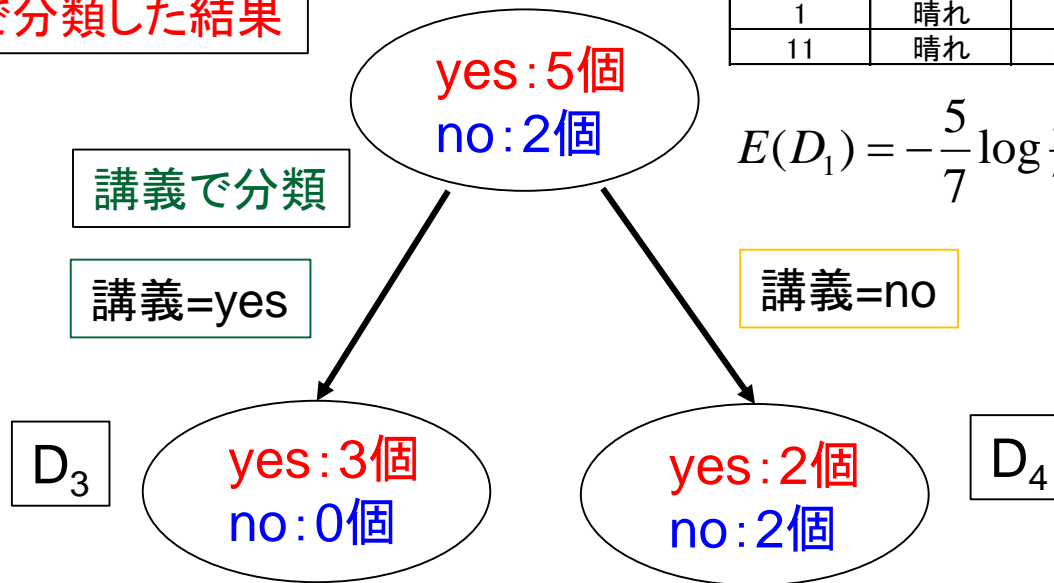
$$E(D_3) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = 1$$

$$E(D_4) = -\frac{4}{5} \log \frac{4}{5} - \frac{1}{5} \log \frac{1}{5} = 0.721$$

$$\begin{aligned}
 \text{Gain}(D_1) &= E(D_1) - \frac{2}{7} E(D_3) - \frac{5}{7} E(D_4) \\
 &= 0.863 - \frac{2}{7} \times 1 - \frac{5}{7} \times 0.721 = 0.061
 \end{aligned}$$

さらに特徴選択③

晴れで分類した結果



$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

$$E(D_3) = -\frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 0$$

$$E(D_4) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} = 1$$

$$\begin{aligned} \text{Gain}(D_1) &= E(D_1) - \frac{3}{7} E(D_3) - \frac{4}{7} E(D_4) \\ &= 0.863 - \frac{3}{7} \times 0 - \frac{4}{7} \times 1 = 0.291 \end{aligned}$$

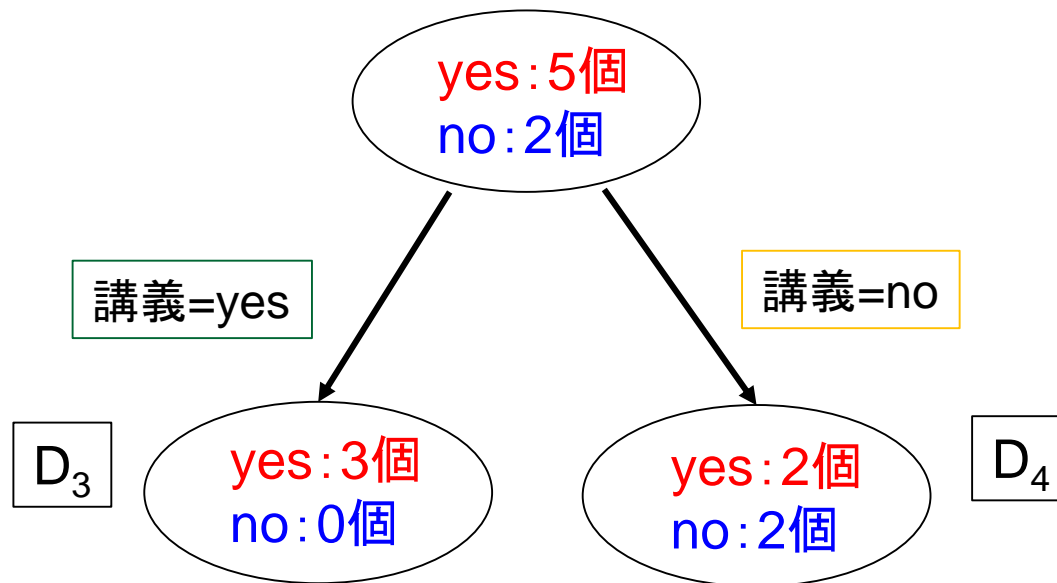
12	晴れ	適温	高い	no	no
3	晴れ	暑い	低い	no	no
16	晴れ	暑い	適当	no	yes
6	晴れ	適温	適当	no	yes
17	晴れ	寒い	高い	yes	yes
1	晴れ	寒い	低い	yes	yes
11	晴れ	暑い	適当	yes	yes

特徴選択の結果

- 気温で分類した場合 → ゲイン = 0.169
- 湿度で分類した場合 → ゲイン = 0.061
- 講義で分類した場合 → **ゲイン = 0.291** 最大



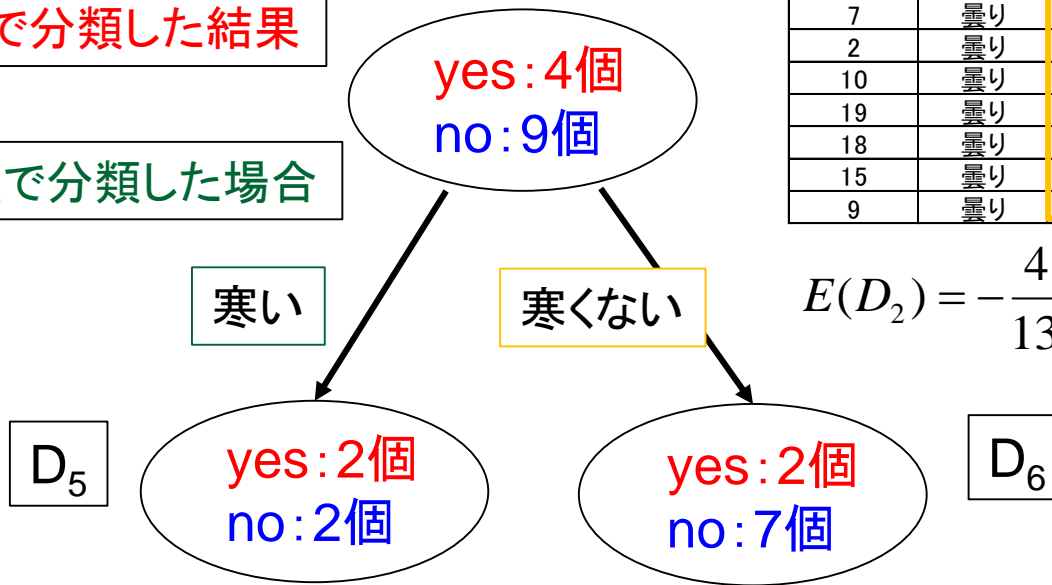
- **講義で分類**



さらに特徴選択④

晴れでないで分類した結果

気温で分類した場合



5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
20	雨	寒い	適当	no	yes
14	雨	寒い	適当	yes	yes
13	雨	暑い	低い	no	no
4	雨	適温	高い	yes	no
7	曇り	暑い	低い	yes	no
2	曇り	暑い	高い	no	yes
10	曇り	適温	適当	no	no
19	曇り	適温	適当	no	no
18	曇り	適温	高い	yes	no
15	曇り	適温	低い	yes	no
9	曇り	適温	低い	yes	yes

$$E(D_2) = -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} = 0.890$$

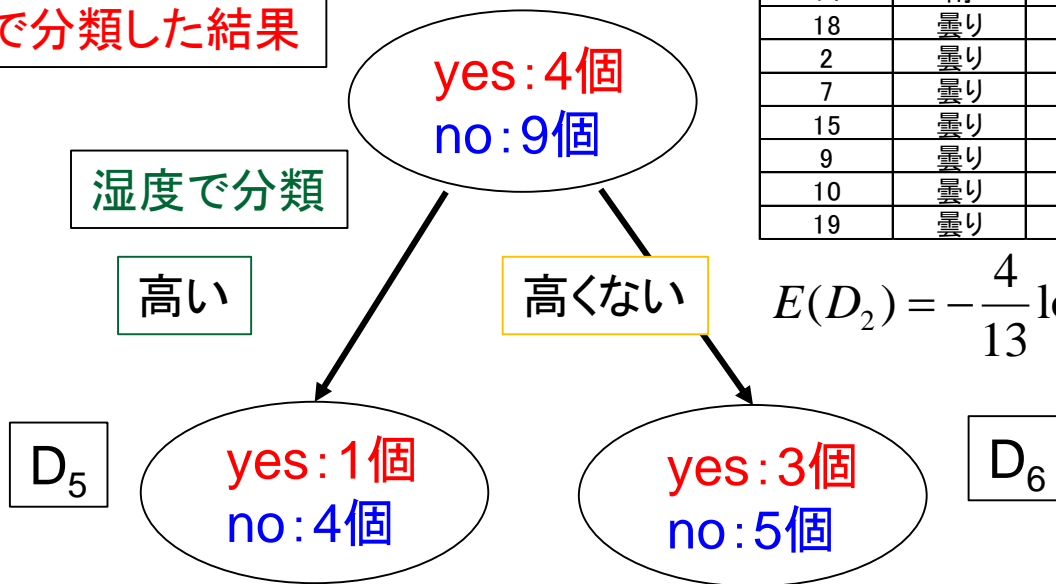
$$E(D_5) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} = 1$$

$$E(D_6) = -\frac{2}{9} \log \frac{2}{9} - \frac{7}{9} \log \frac{7}{9} = 0.764$$

$$\begin{aligned}
 \text{Gain}(D_2) &= E(D_2) - \frac{4}{13} E(D_5) - \frac{9}{13} E(D_6) \\
 &= 0.890 - \frac{4}{13} \times 1.0 - \frac{9}{13} \times 0.764 = 0.053
 \end{aligned}$$

さらに特徴選択⑤

晴れでないで分類した結果



5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
4	雨	適温	高い	yes	no
13	雨	暑い	低い	no	no
20	雨	寒い	適当	no	yes
14	雨	寒い	適当	yes	yes
18	曇り	適温	高い	yes	no
2	曇り	暑い	高い	no	yes
7	曇り	暑い	低い	yes	no
15	曇り	適温	低い	yes	no
9	曇り	適温	低い	yes	yes
10	曇り	適温	適当	no	no
19	曇り	適温	適当	no	no

$$E(D_2) = -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} = 0.890$$

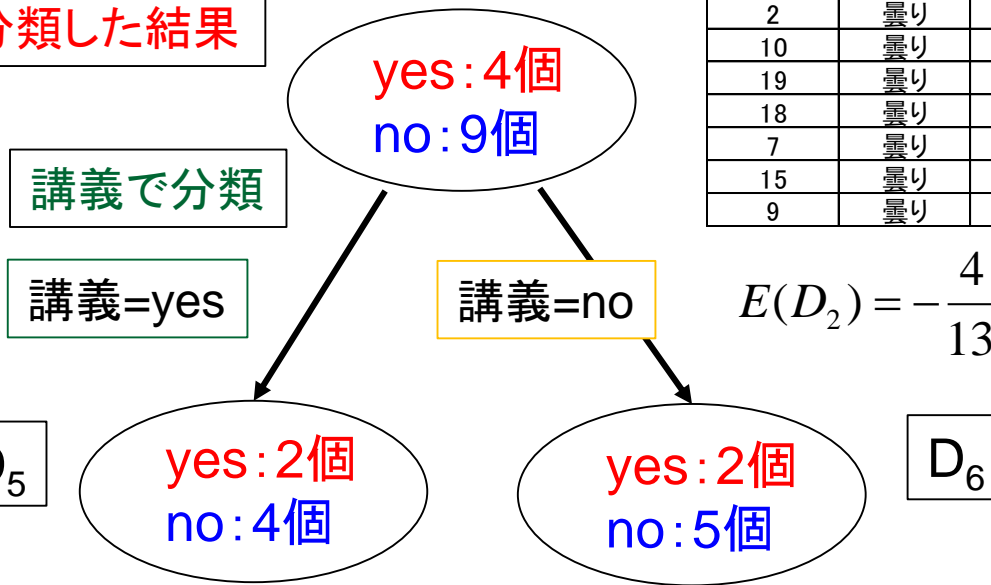
$$E(D_5) = -\frac{1}{5} \log \frac{1}{5} - \frac{4}{5} \log \frac{4}{5} = 0.721$$

$$E(D_6) = -\frac{3}{8} \log \frac{3}{8} - \frac{5}{8} \log \frac{5}{8} = 0.954$$

$$\begin{aligned} \text{Gain}(D_2) &= E(D_2) - \frac{5}{13} E(D_5) - \frac{8}{13} E(D_6) \\ &= 0.890 - \frac{5}{13} \times 0.721 - \frac{8}{13} \times 0.954 = 0.025 \end{aligned}$$

さらに特徴選択⑥

晴れでないで分類した結果



5	雨	寒い	高い	no	no
8	雨	寒い	高い	no	no
13	雨	暑い	低い	no	no
20	雨	寒い	適当	no	yes
4	雨	適温	高い	yes	no
14	雨	寒い	適当	yes	yes
2	曇り	暑い	高い	no	yes
10	曇り	適温	適当	no	no
19	曇り	適温	適当	no	no
18	曇り	適温	高い	yes	no
7	曇り	暑い	低い	yes	no
15	曇り	適温	低い	yes	no
9	曇り	適温	低い	yes	yes

$$E(D_2) = -\frac{4}{13} \log \frac{4}{13} - \frac{9}{13} \log \frac{9}{13} = 0.890$$

$$E(D_5) = -\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6} = 0.918$$

$$E(D_6) = -\frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7} = 0.863$$

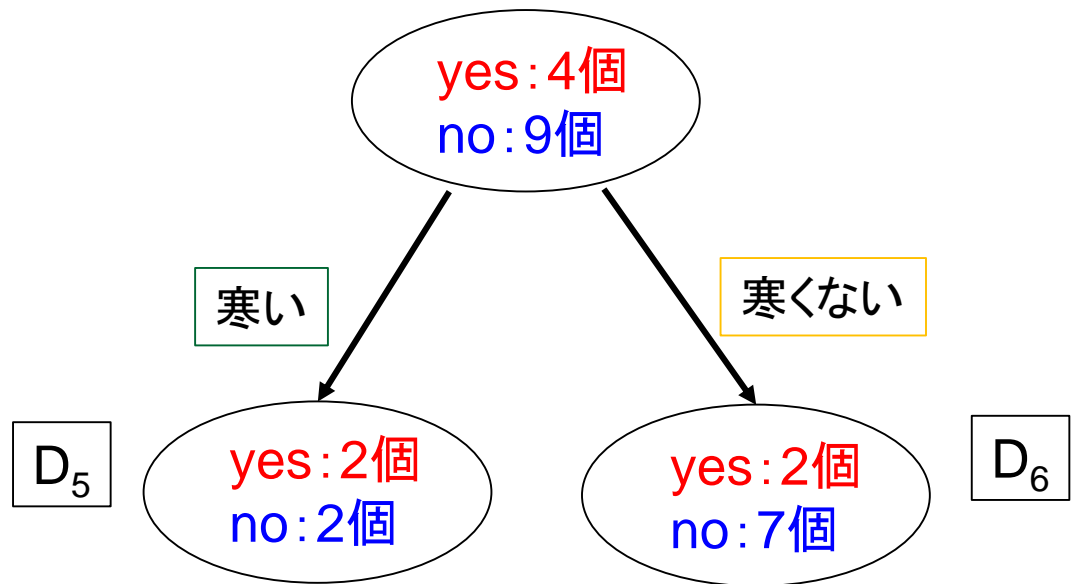
$$\begin{aligned} \text{Gain}(D_2) &= E(D_2) - \frac{5}{13} E(D_5) - \frac{8}{13} E(D_6) \\ &= 0.890 - \frac{6}{13} \times 0.918 - \frac{7}{13} \times 0.863 = 0.002 \end{aligned}$$

特徴選択の結果

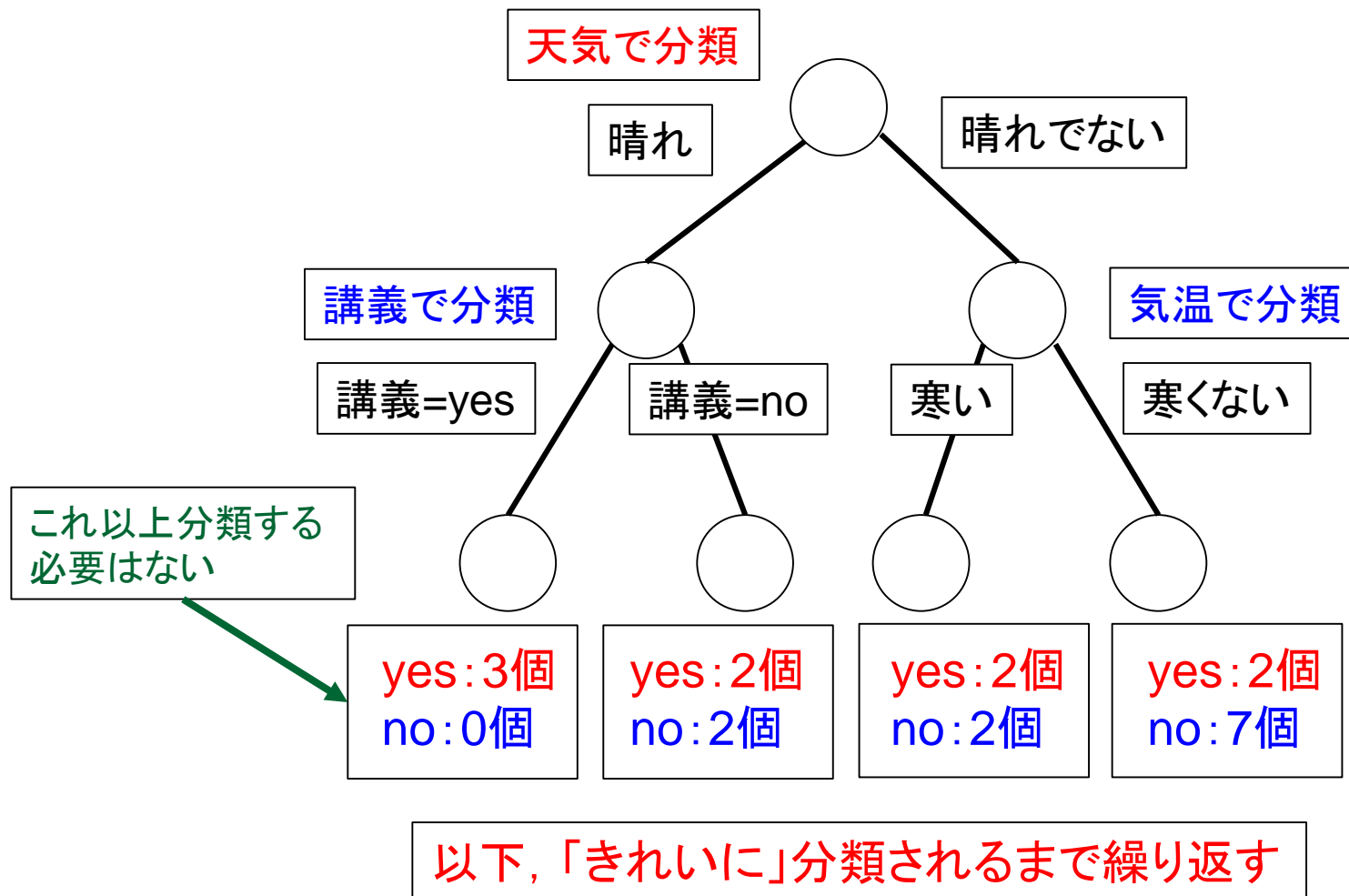
- 気温で分類した場合 → ゲイン = 0.053 最大
- 湿度で分類した場合 → ゲイン = 0.025
- 講義で分類した場合 → ゲイン = 0.002



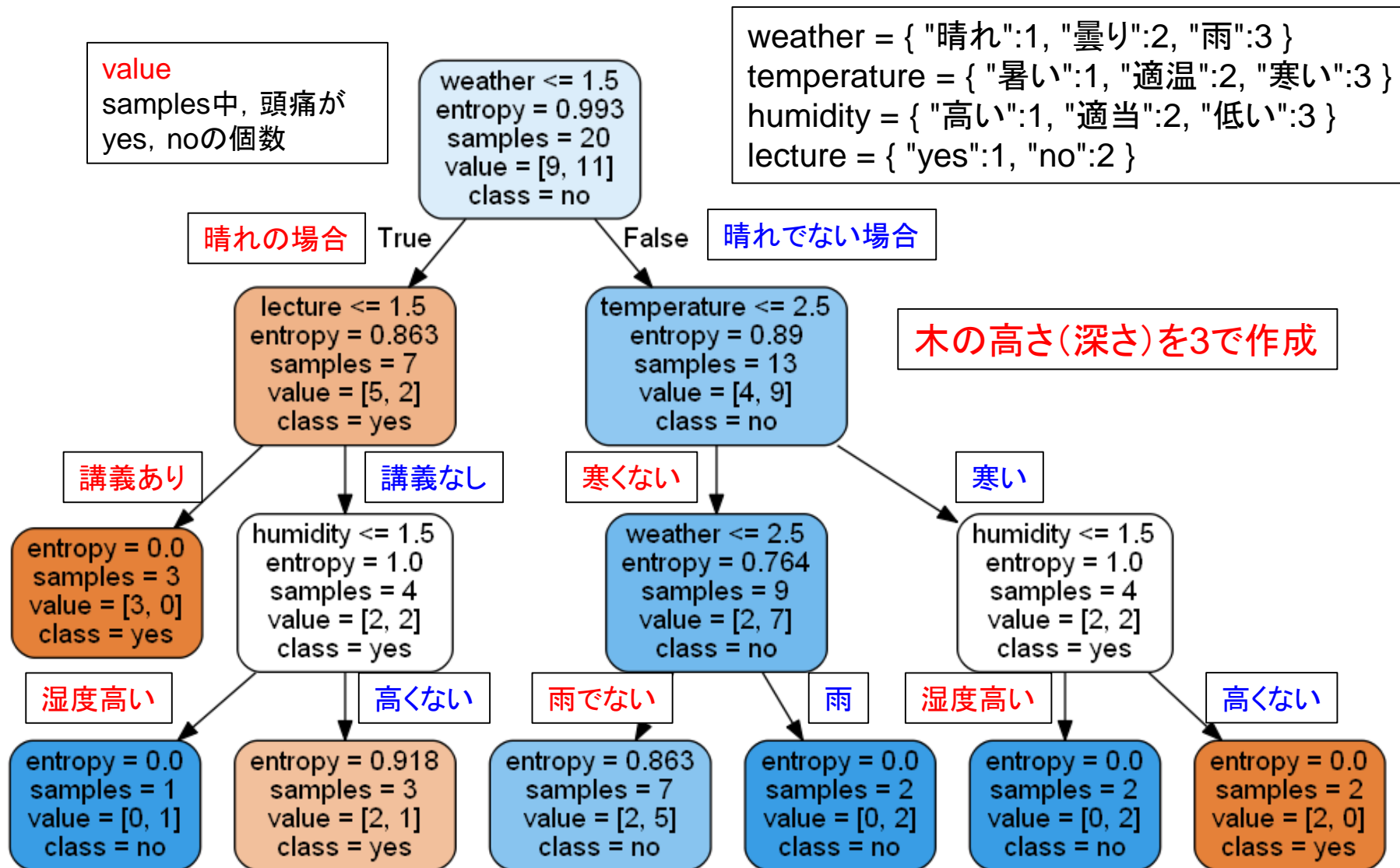
- 気温で分類



生成された決定木



最終的に生成された決定木



決定木の学習アルゴリズム

D: 学習データ

決定木の学習(D):

if 停止条件を満たしている:
停止

理想はエントロピーが0

Dを分類するための特徴選択

ゲイン最大となる特徴

$D_L, D_R \leftarrow$ 分類結果

決定木の学習(D_L)

再帰的に学習

決定木の学習(D_R)

D_L : 左ノードで対象となるデータ

D_R : 右ノードで対象となるデータ

分類のための指標

- エントロピー(平均情報量)

- クラス $c=1,2,\dots,C$

- クラス c_i の生起確率 p_i

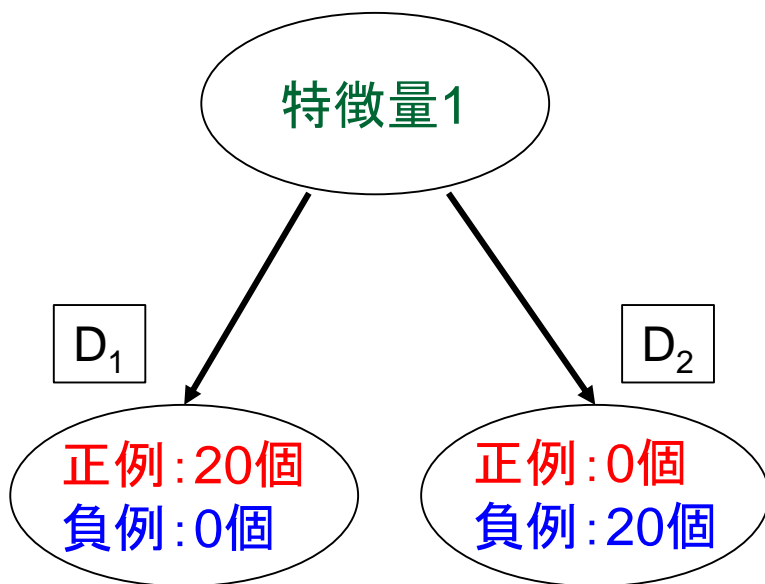
$$E = -\sum_{i=1}^C p_i \log p_i$$

- ジニ係数

$$G = 1 - \sum_{i=1}^C (p_i)^2$$

ジニ係数①

$$G = 1 - \sum_{i=1}^C (p_i)^2$$



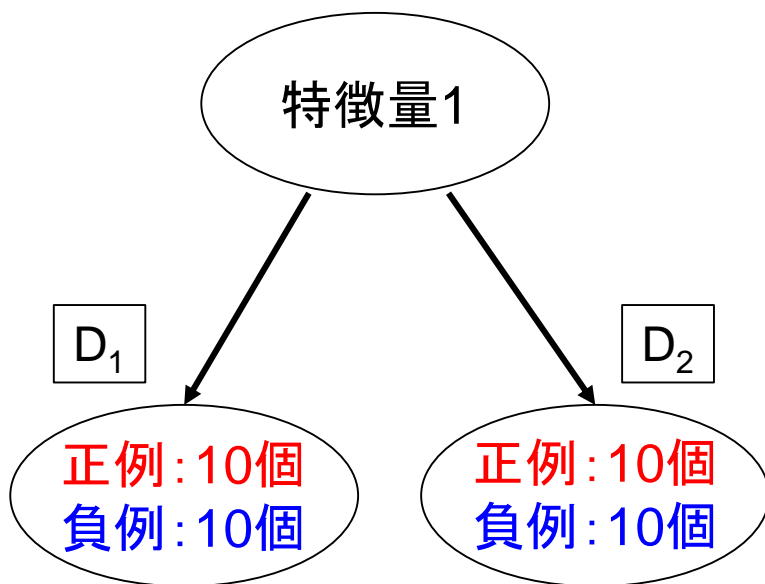
$$D_1 \quad G(D_1) = 1 - \left(\frac{20}{20}\right)^2 - \left(\frac{0}{20}\right)^2 = 0$$

$$D_2 \quad G(D_2) = 1 - \left(\frac{0}{20}\right)^2 - \left(\frac{20}{20}\right)^2 = 0$$

ジニ係数が0に近いほど、「きれいに」分類される

ジニ係数②

$$G = 1 - \sum_{i=1}^C (p_i)^2$$

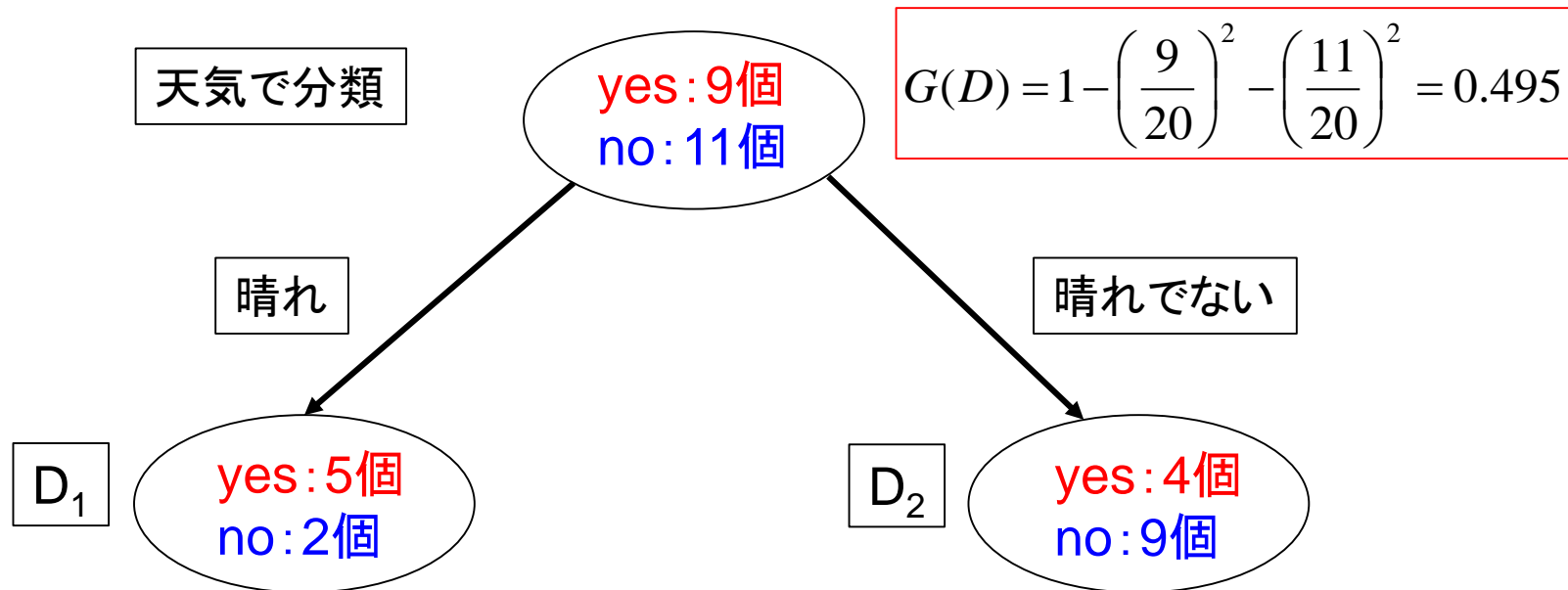


$$D_1 \quad G(D_1) = 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = 0.5$$

$$D_2 \quad G(D_2) = 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2 = 0.5$$

ジニ係数が大きいほど、「きれいに」分類されない

ジニ係数を用いる場合①



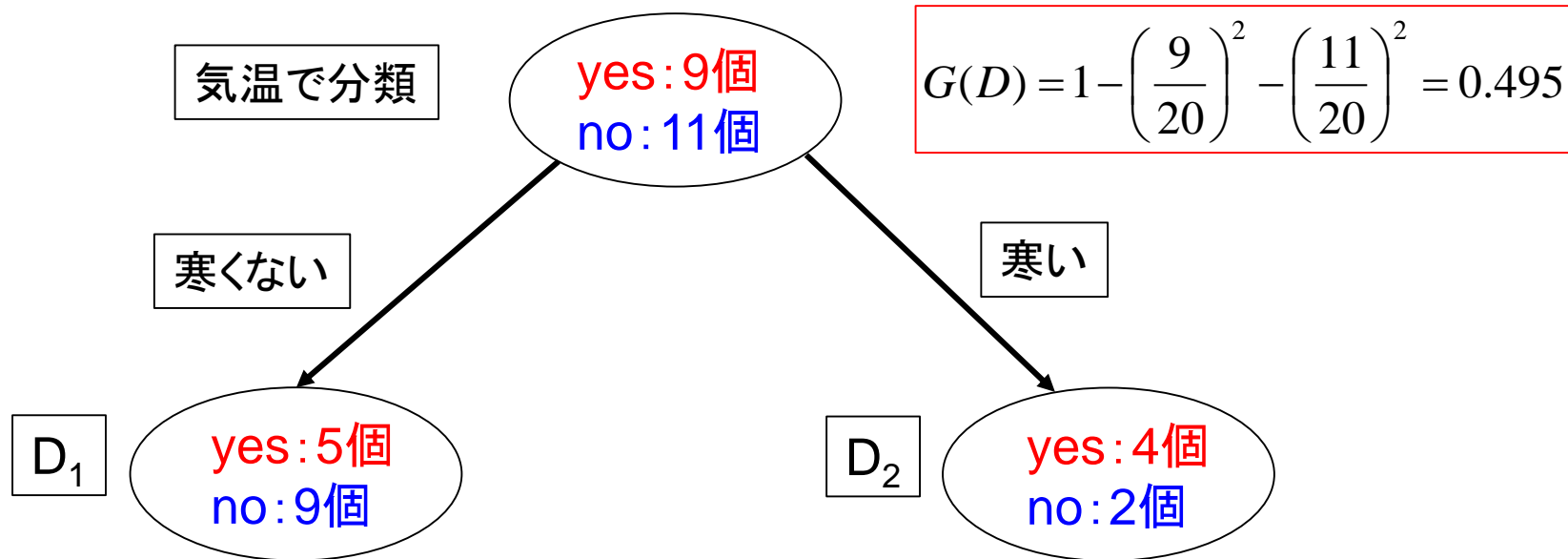
$$G(D_1) = 1 - \left(\frac{5}{7}\right)^2 - \left(\frac{2}{7}\right)^2 = 0.408$$

$$G(D_2) = 1 - \left(\frac{4}{13}\right)^2 - \left(\frac{9}{13}\right)^2 = 0.426$$

$$\begin{aligned} \text{Gain}(D) &= G(D) - \frac{7}{20} G(D_1) - \frac{13}{20} G(D_2) \\ &= 0.495 - \frac{7}{20} \times 0.408 - \frac{13}{20} \times 0.426 = 0.075 \end{aligned}$$

*ジニ係数によるゲインをジニ指標とも呼ばれますが、資料ではゲインで統一して使います

ジニ係数を用いる場合②

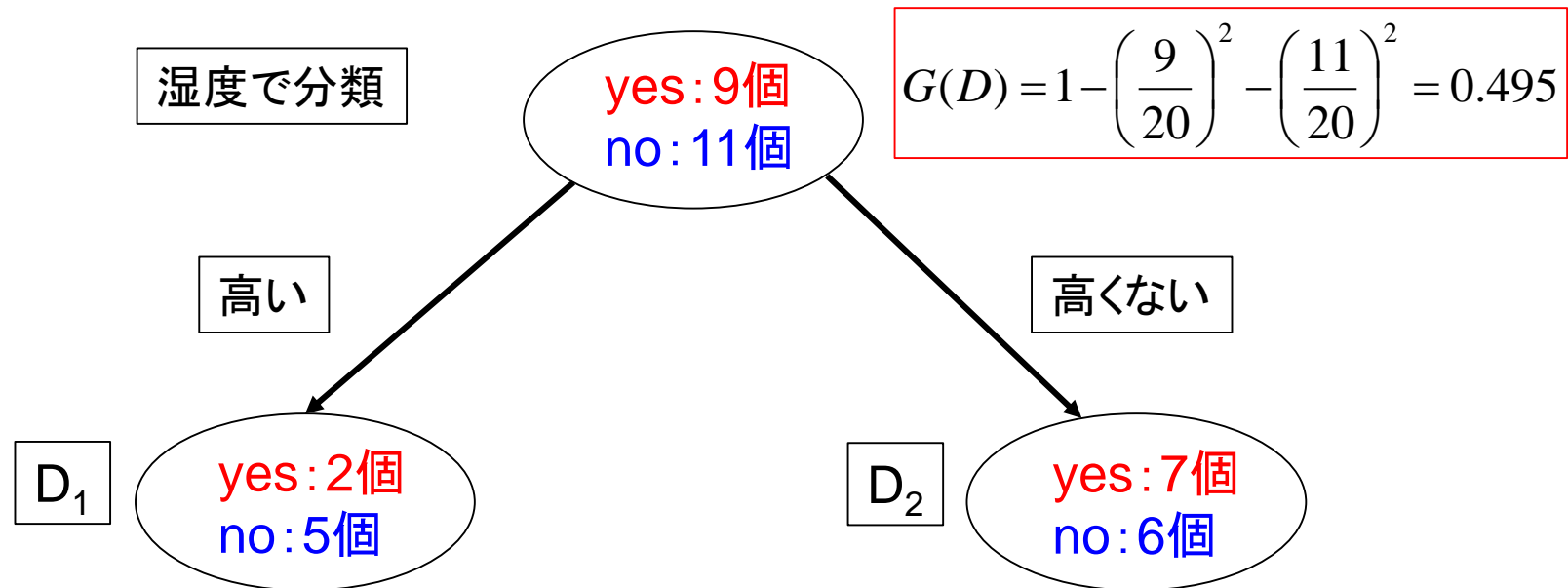


$$G(D_1) = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = 0.459$$

$$G(D_2) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.444$$

$$\begin{aligned} \text{Gain}(D) &= G(D) - \frac{14}{20} G(D_1) - \frac{6}{20} G(D_2) \\ &= 0.495 - \frac{14}{20} \times 0.459 - \frac{6}{20} \times 0.444 = 0.04 \end{aligned}$$

ジニ係数を用いる場合③



$$G(D) = 1 - \left(\frac{9}{20}\right)^2 - \left(\frac{11}{20}\right)^2 = 0.495$$

$$G(D_1) = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 = 0.408$$

$$G(D_2) = 1 - \left(\frac{7}{13}\right)^2 - \left(\frac{6}{13}\right)^2 = 0.497$$

$$\begin{aligned} \text{Gain}(D) &= G(D) - \frac{7}{20} G(D_1) - \frac{13}{20} G(D_2) \\ &= 0.495 - \frac{7}{20} \times 0.408 - \frac{13}{20} \times 0.497 = 0.029 \end{aligned}$$

ジニ係数を用いる場合④

講義で分類

yes: 9個
no: 11個

$$G(D) = 1 - \left(\frac{9}{20}\right)^2 - \left(\frac{11}{20}\right)^2 = 0.495$$

yes

no

D_1

yes: 5個
no: 4個

D_2

yes: 4個
no: 7個

$$G(D_1) = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = 0.493$$

$$G(D_2) = 1 - \left(\frac{4}{11}\right)^2 - \left(\frac{7}{11}\right)^2 = 0.462$$

$$\begin{aligned} \text{Gain}(D) &= G(D) - \frac{9}{20} G(D_1) - \frac{11}{20} G(D_2) \\ &= 0.495 - \frac{9}{20} \times 0.493 - \frac{11}{20} \times 0.462 = 0.018 \end{aligned}$$

ジニ係数を用いた特徴選択

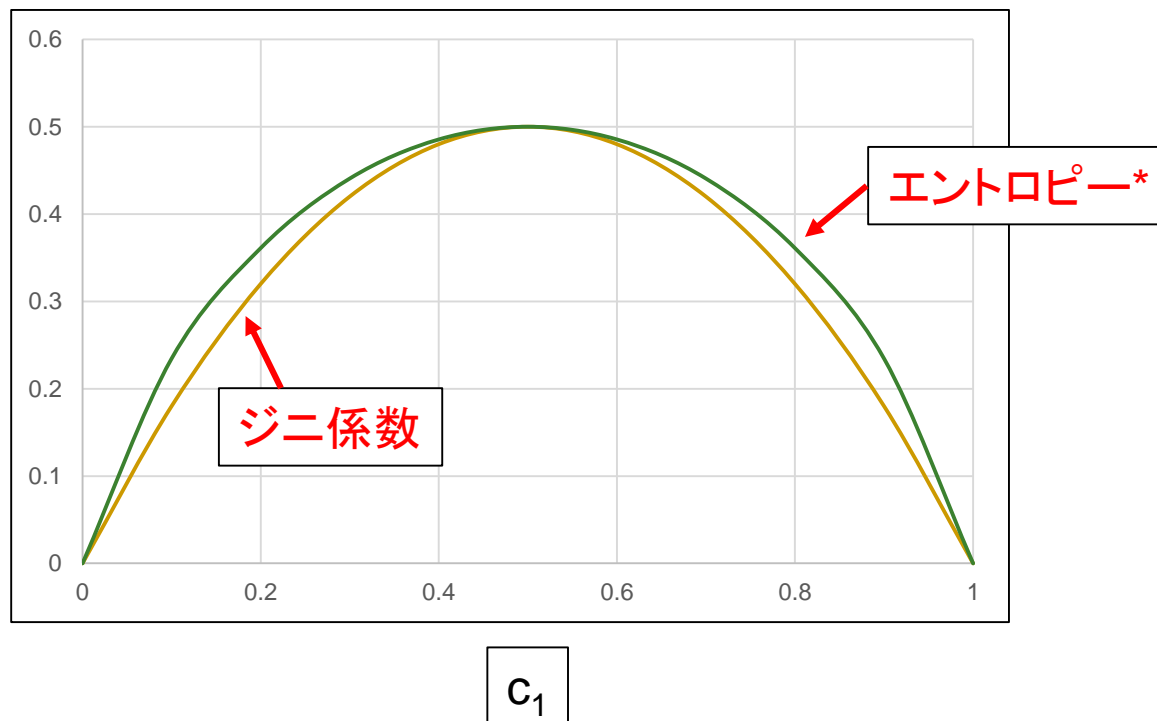
- 天気で分類した場合 → ゲイン = 0.075 最大
- 気温で分類した場合 → ゲイン = 0.04
- 湿度で分類した場合 → ゲイン = 0.029
- 講義で分類した場合 → ゲイン = 0.018



- 天気で分類
 - エントロピーによる特徴選択と同じ結果

エントロピーとジニ係数

■ 二値分類の場合

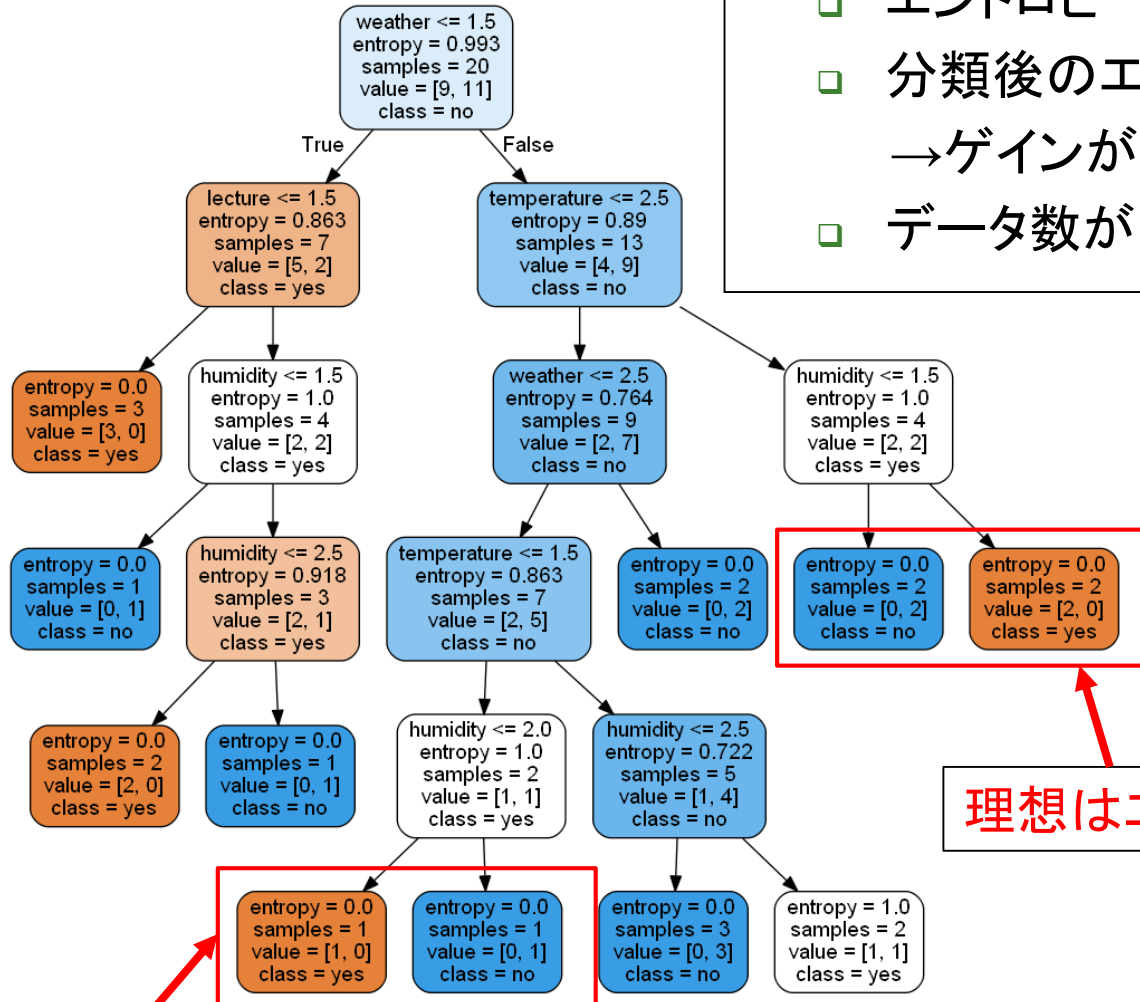


*エントロピーは0～1, ジニ係数は0～0.5なので, エントロピーの値を1/2としています

停止条件①

■ 停止条件

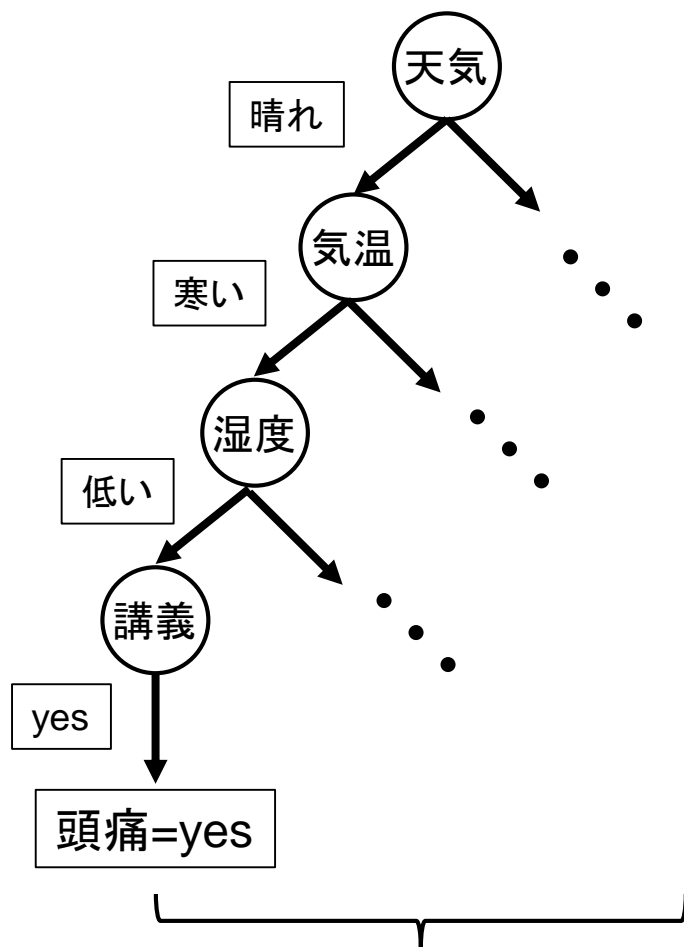
- エントロピーが0となる
- 分類後のエントロピーが下がらない
→ゲインがほぼ0
- データ数が1個になる



理想はエントロピーが0

データ数が1個になるまで繰り返すべきか

データ数が1個になるまで分類した場合

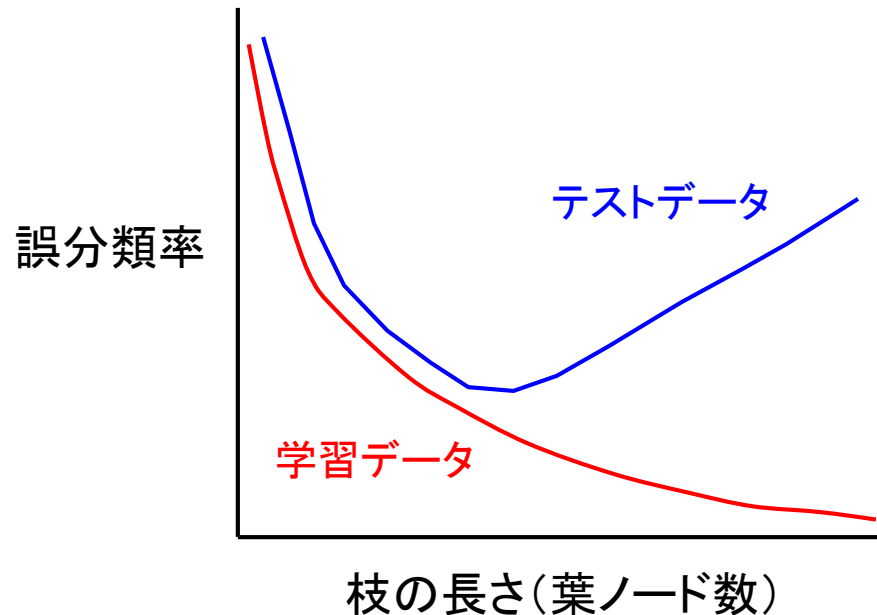


データ数が1個になるまで分類した場合
→ 学習データ特有のルールを抽出
→ 一般的なルールを抽出したことはない

20通りの分岐

停止条件②

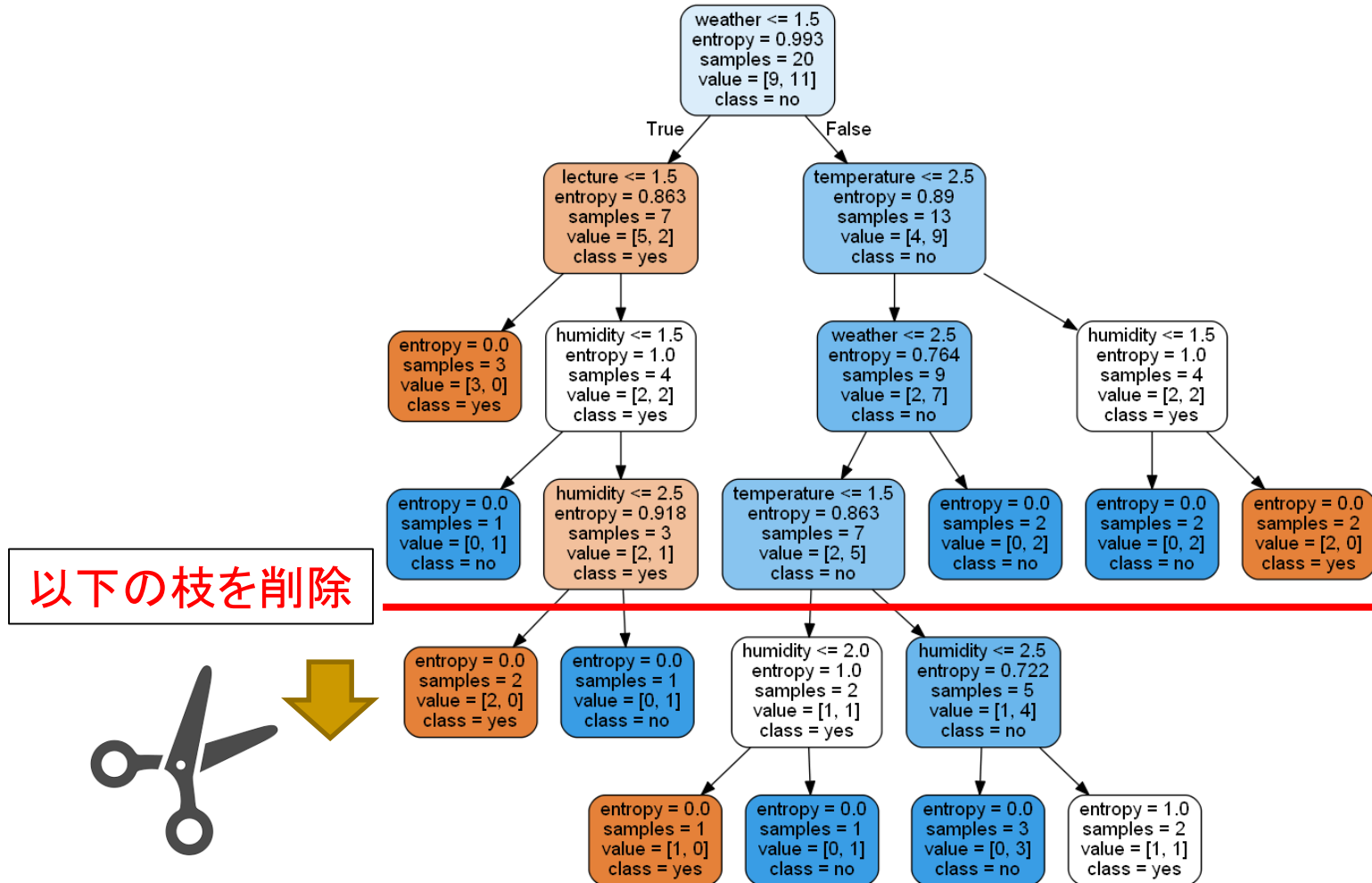
- データ数が1個になるまで繰り返した場合
 - テストデータに対する精度が低下(過学習*)



- ある程度、個数が絞られた段階で停止
 - 枝刈り

*過適合 (over fitting), 過剰適合とも呼ばれます

枝刈り (Pruning)



枝刈りの方法①

■ 評価指標による枝刈り

- 生成した決定木による誤分類率: $e(T)$
- ペナルティ: π
- 葉ノードの数: k
- データ数: N

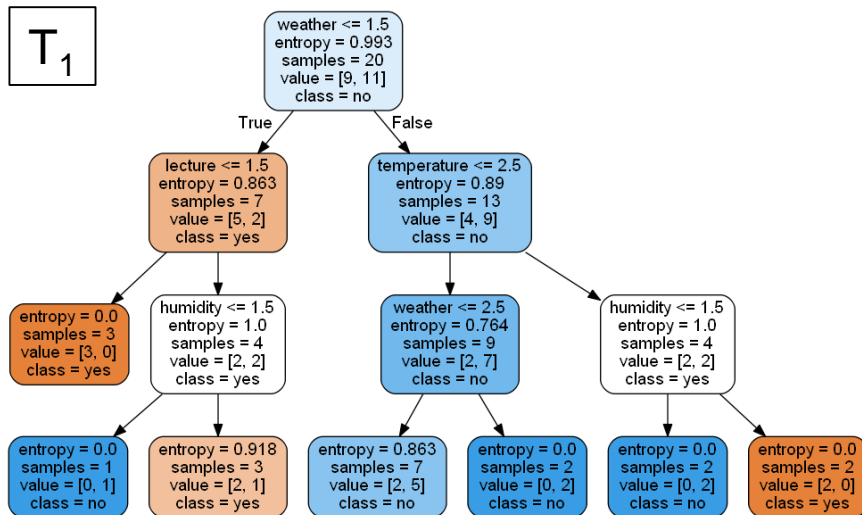
決定木の評価指標

$$\hat{e}(T) = \frac{e(T) + k\pi}{N} = \frac{e(T)}{N} + \frac{k\pi}{N}$$

決定木の複雑さ(葉ノードの個数)
をペナルティとする

評価指標による枝刈り

T_1

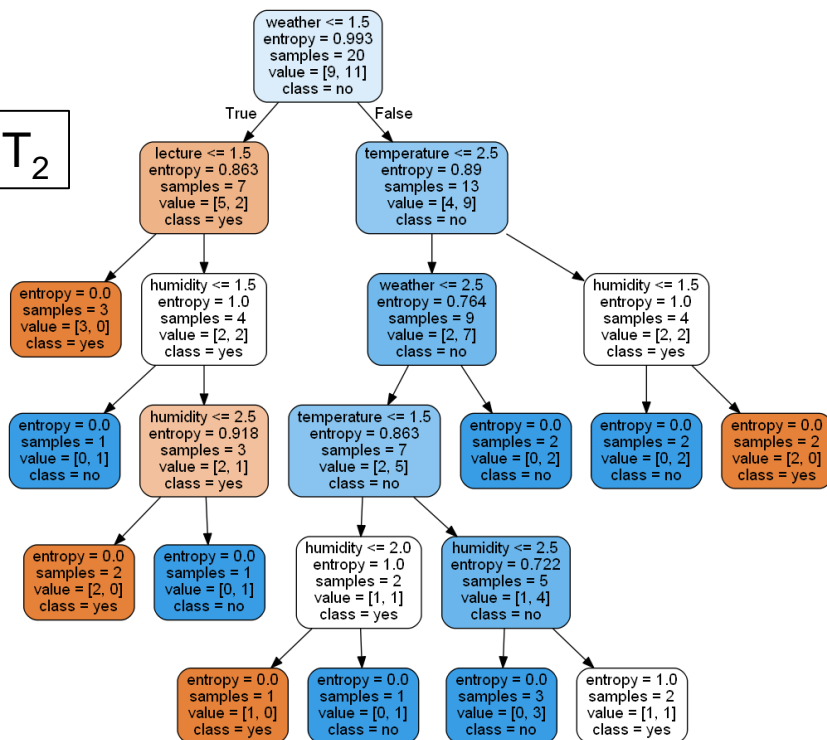


誤分類率: 3/20
ペナルティ: 0.5
葉ノードの数: 7
データ数: 20

$$\hat{e}(T_1) = \frac{3/20 + 7 \times 0.5}{20} = 0.1825$$

決定木 T_1 を選択

T_2



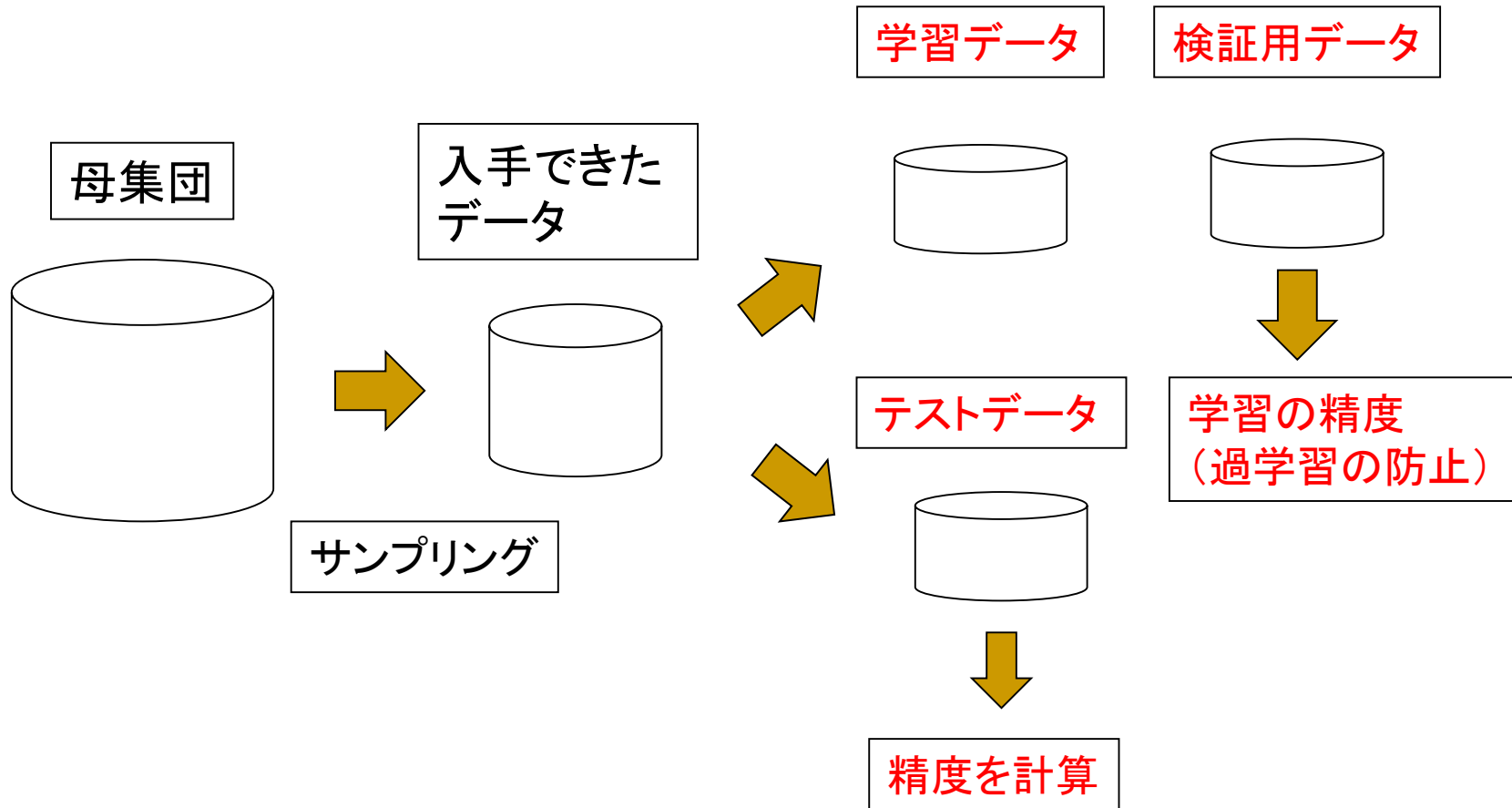
誤分類率: 1/20
ペナルティ: 0.5
葉ノードの数: 10
データ数: 20

誤分類率は T_2 の方が低い

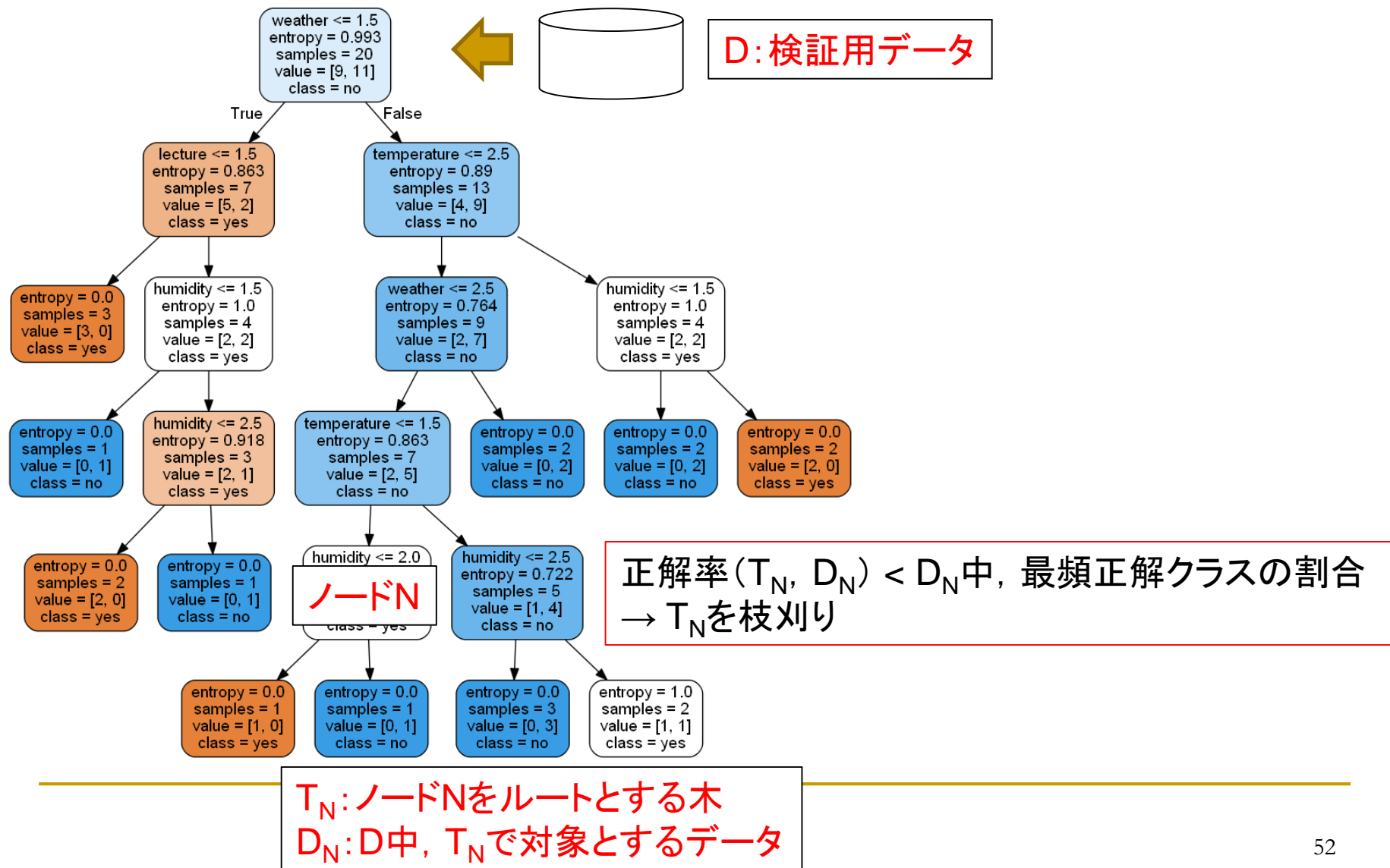
$$\hat{e}(T_2) = \frac{1/20 + 10 \times 0.5}{20} = 0.2525$$

枝刈りの方法②

■ 検証用データ (Validation) の利用



検証用データによる枝刈り



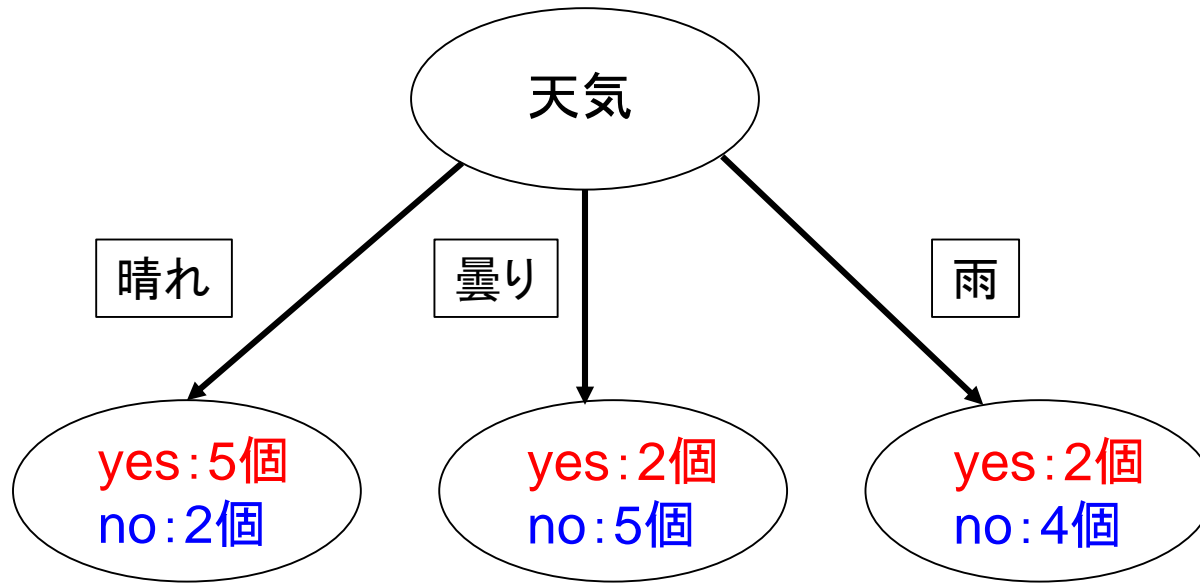
決定木作成のアルゴリズム①

- ID3 (Iterative Dichotomiser 3)
 - 説明変数はカテゴリーカル変数 (離散値)
 - 多分岐が可能
 - 分類指標はエントロピーによるゲインを利用
- C4.5 (近年はC5.0)
 - ID3の改良アルゴリズム
 - 特徴量はカテゴリーカル変数でなくてよい
 - 多分岐が可能
 - 分類前後のエントロピーの比 (ゲイン比) を利用

決定木作成のアルゴリズム②

- CART(Classification and Regression Tree)
 - 特徴量はカテゴリカル変数でなくてよい
 - 二分岐のみ
 - 分類の指標はジニ係数によるゲインを利用
 - 目的変数が数値データにも対応(回帰木)

多分岐



晴れ

$$E(D_1) = -\frac{5}{7} \log \frac{5}{7} - \frac{2}{7} \log \frac{2}{7} = 0.863$$

雨

$$E(D_3) = -\frac{2}{6} \log \frac{2}{6} - \frac{4}{6} \log \frac{4}{6} = 0.918$$

曇り

$$E(D_2) = -\frac{2}{7} \log \frac{2}{7} - \frac{5}{7} \log \frac{5}{7} = 0.863$$

CART

- 二分岐

- 分類の評価指標

- ジニ係数によるゲイン

熱中症

YES

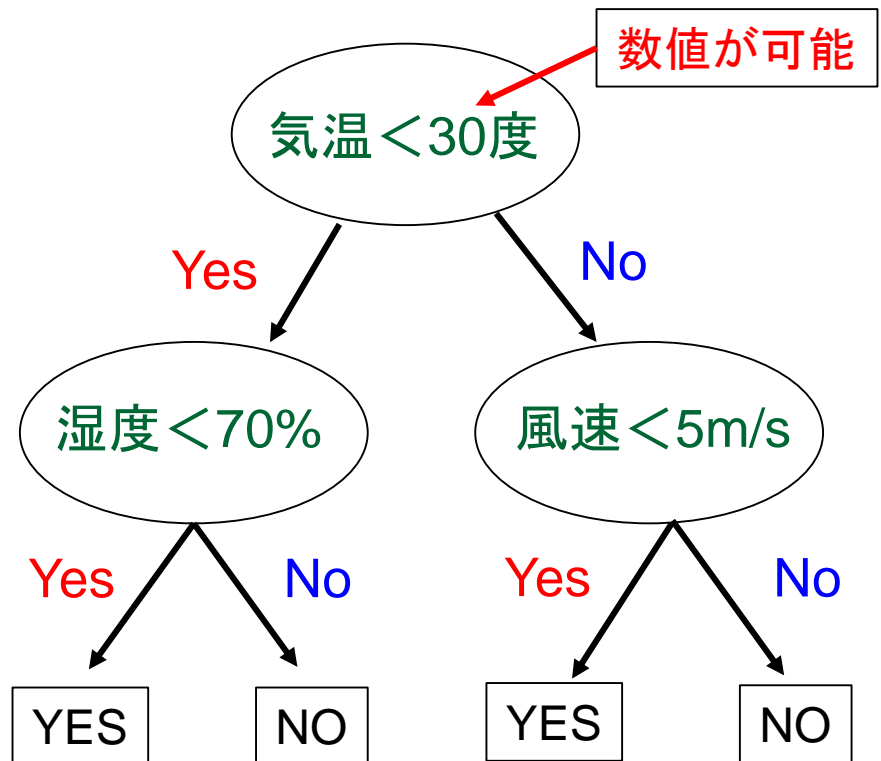
NO

YES

NO

- 特徴量はカテゴリカル変数でなくてよい(連続値が可能)

- 目的変数が数値データにも対応(回帰木)



連続値の場合の閾値の求め方①

データ	特徴量1	目的変数
1	7.2	クラス1
2	2.3	クラス2
3	4.2	クラス2
4	3.5	クラス1
5	6.1	クラス2
6	0.5	クラス1
7	2.7	クラス1
8	4.2	クラス1
9	9.1	クラス2
10	8.5	クラス2



データ	特徴量1	閾値の候補	目的変数
6	0.5		クラス1
2	2.3	1.4	クラス2
7	2.7	2.5	クラス1
4	3.5	3.1	クラス1
3	4.2	3.85	クラス2
8	4.2	4.2	クラス1
5	6.1	5.15	クラス2
1	7.2	6.65	クラス1
10	8.5	7.85	クラス2
9	9.1	8.8	クラス2

(例)
特徴量1 < 3.1

① 特徴量1でソート

② 閾値の候補を求める
上下同士の平均値(中央値)

③ 閾値の候補で特徴選択
(この場合, 9箇所)
→ ジニ係数によるゲインを求める

特徴量1 < 1.4
特徴量1 < 2.5
⋮
特徴量1 < 7.85
特徴量1 < 8.8

9箇所

連続値の場合の閾値の求め方②

データ	特徴量1	目的変数
1	7.2	クラス1
2	2.3	クラス2
3	4.2	クラス2
4	3.5	クラス1
5	6.1	クラス2
6	0.5	クラス1
7	2.7	クラス1
8	4.2	クラス1
9	9.1	クラス2
10	8.5	クラス2



データ	特徴量1	閾値の候補	目的変数
6	0.5		クラス1
2	2.3	1.4	クラス2
7	2.7	2.5	クラス1
4	3.5	3.1	クラス1
3	4.2	3.85	クラス2
8	4.2	4.2	クラス1
5	6.1	5.15	クラス2
1	7.2	6.65	クラス1
10	8.5	7.85	クラス2
9	9.1	8.8	クラス2

(例)
特徴量1 < 3.1

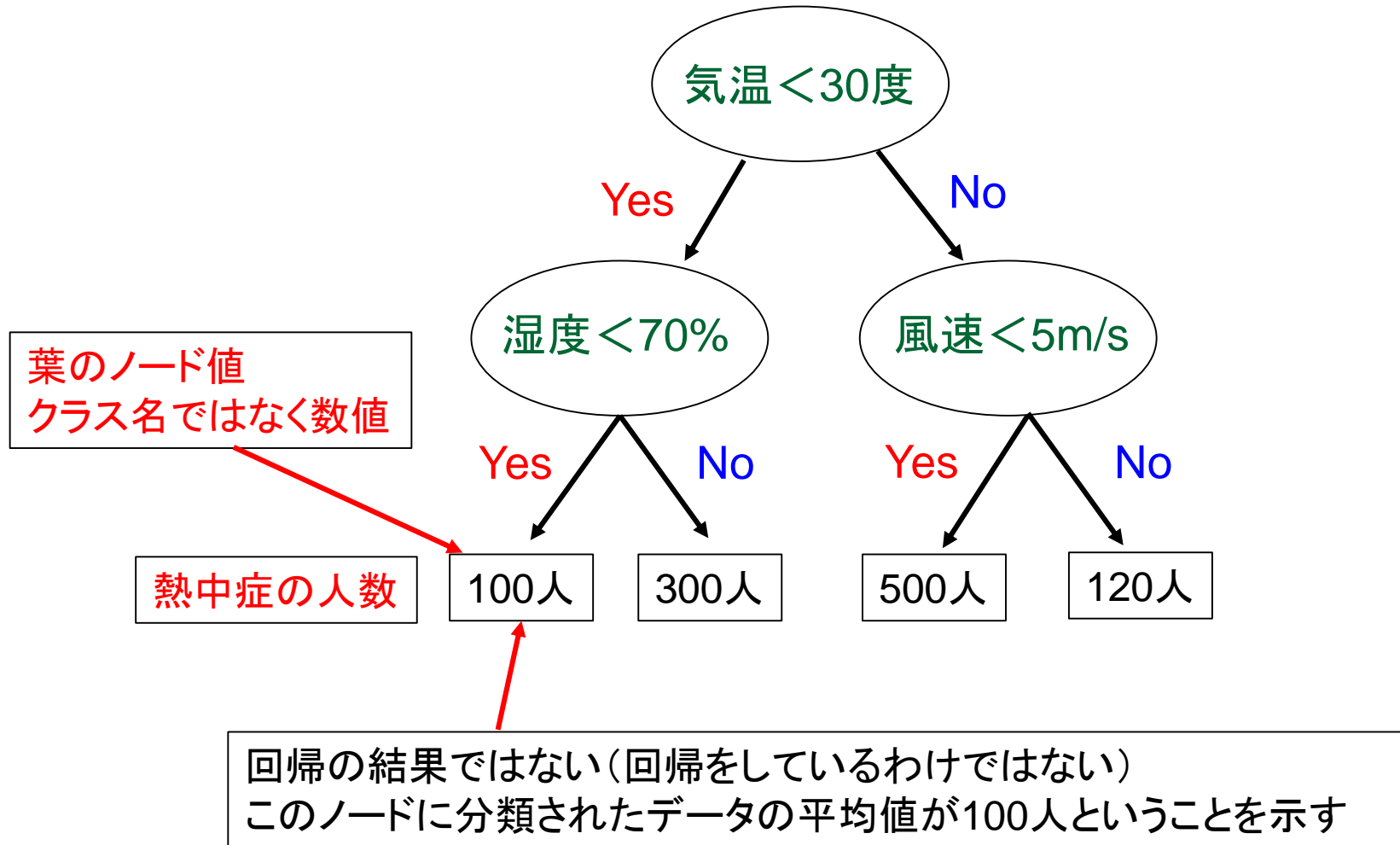
① 特徴量1でソート

② 閾値の候補を求める
上下同士の平均値(中央値)

③ 閾値の候補で特徴選択
→ ジニ係数によるゲインを求める
(この場合, 9箇所)

④ ゲインが最大となる特徴量,
閾値で分類

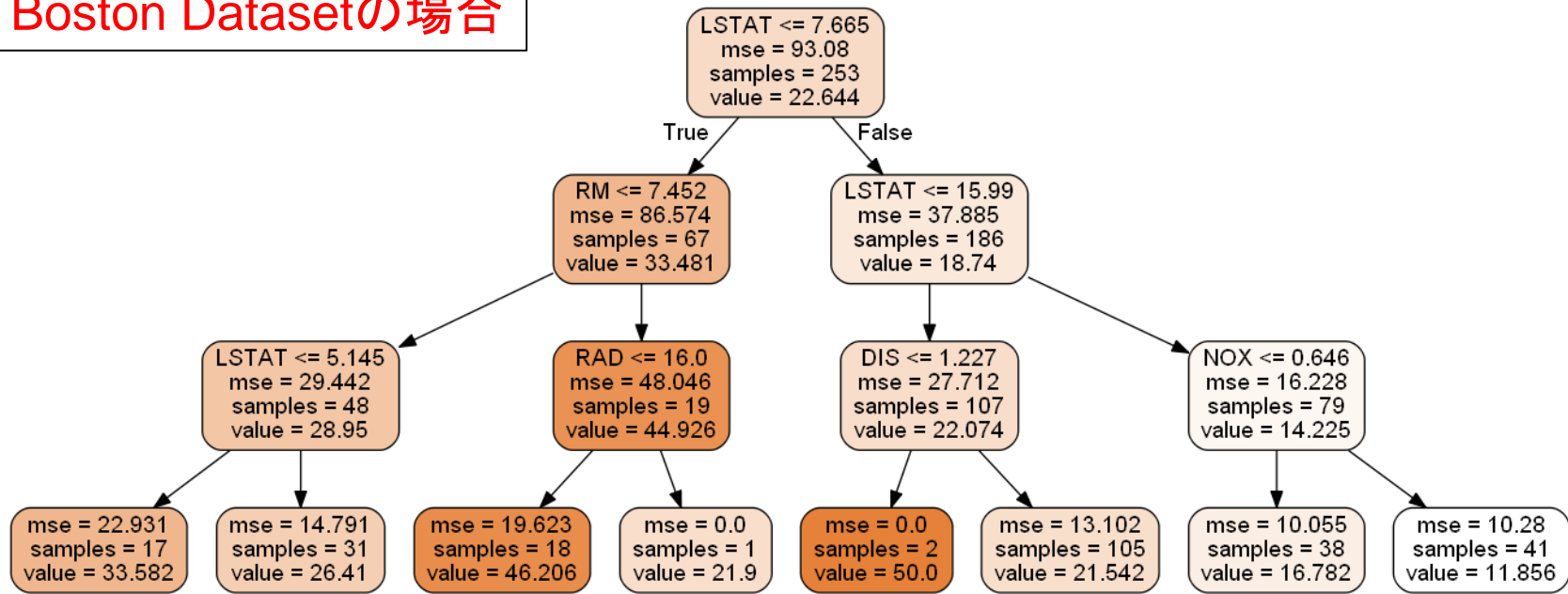
回帰木 (Regression Tree)



*クラス分類を目的とする決定木は分類木とも呼ばれます

回帰木の例

Boston Datasetの場合



この葉の場合、
値 (value) が 33.582

分類のための指標

■ 分散

$$V(D) = \frac{1}{N} \sum_{d_i \in D} (d_i - \bar{d})^2$$

D: データ
N: データ数

← 平均値

■ ゲイン

$$Gain(D, a) = V(D) - \frac{N_L}{N} V(D_L) - \frac{N_R}{N} V(D_R)$$

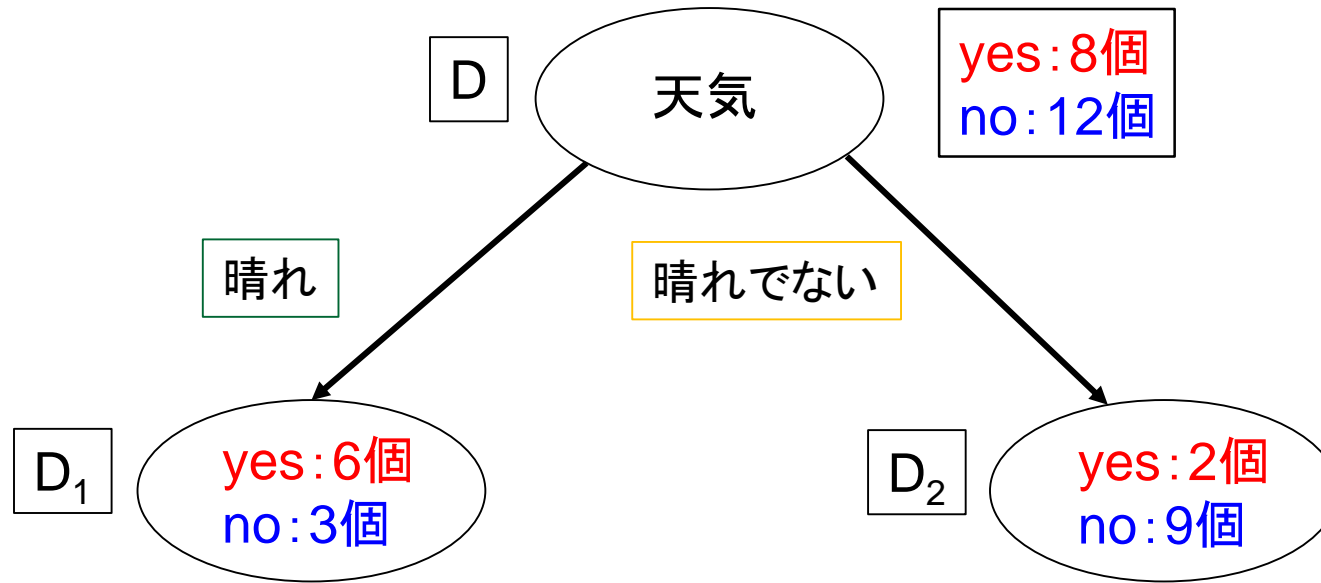
D_L : 左ノードのデータ
 N_L : 左ノードのデータ数

D_R : 右ノードのデータ
 N_R : 右ノードのデータ数

■ 分散最小

→ 平均値に近いデータが各ノードに集まる

練習問題



- ① 初期のエントロピーを求めなさい.
- ② D1のエントロピーを求めなさい.
- ③ D2のエントロピーを求めなさい.
- ④ 天気で分類した際のゲインを求めなさい.

対数の計算はExcelでして下さい*
=log(値, 2)

*特徴ごとにゲインの大小を比較するのが目的なので、対数の底は2でなくともかまいません

決定木の学習

Irisデータセット(クラス分類)

Bostonデータセット(回帰木)

Iris dataset

■ アヤメの分類問題

用途	クラス分類
データ数	150
特徴量	4
目的変数	3

クラス名	データ数
setosa	50
versicolor	50
virginica	50



Irisデータセットの分類

```
import numpy
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn import tree
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix

# データのロード
iris = datasets.load_iris()

# 種類( setosa , versicolor , virginica )
name = iris.target_names

# 特徴量
feature_name = iris.feature_names

# データ(説明変数)
data = iris.data
```

Iris_Tree.py

パッケージのimport

importが必要

```
# 目的変数 (setosa:0 , versicolor:1 , virginica:2)
```

```
label = iris.target
```

```
# 学習データ, テストデータ
```

```
train_data, test_data, train_label, test_label = train_test_split(data, label, test_size=0.5,  
random_state=None)
```

```
print( test_label )
```

クラス分類用の決定木
DecisionTreeClassifier

```
# 決定木
```

```
model = tree.DecisionTreeClassifier(criterion="gini", max_depth=3)
```

```
# 学習
```

```
model.fit(train_data, train_label)
```

指標はジニ係数
エントロピーを用いる場合
criterion="entropy"

決定木の深さ=3

```
# 予測
```

```
predict = model.predict(test_data)
```

```
print( predict )
```

テストデータの予測

```
print( " [ 予測結果 ] " )
```

```
print( classification_report(test_label, predict) )
```

```
print( "¥n [ 正解率 ]" )  
print( accuracy_score(test_label, predict) )
```

正解率の計算

```
print( "¥n [ 混同行列 ]" )  
print( confusion_matrix(test_label, predict) )
```

混同行列の表示

```
tree.export_graphviz(model, out_file="tree.dot", feature_names=feature_name,  
class_names=name, filled=True, rounded=True)
```

決定木の描画のために必要

決定木を描画(画像化)するためには...

① Graphviz (<https://www.graphviz.org/>) をインストールして下さい*

② コマンドプロンプト上で,

```
> dot -T png tree.dot -o tree.png
```

画像(png形式)で保存

上記のプログラムで指定したファイル(dot形式)

DecisionTreeClassifier

```
from sklearn import tree
```

```
tree.DecisionTreeClassifier(criterion=分類指標,  
max_depth=決定木の深さ)
```

```
model = tree.DecisionTreeClassifier(criterion="gini", max_depth=3)
```

指標はジニ係数
エントロピーを用いる場合
criterion="entropy"

決定木の深さ=3

実行結果

```
C:\Windows\system32\cmd.exe
C:\home\shino\ML-2019\Tree\program>python Iris_Tree.py
[1 2 0 0 1 2 2 1 1 1 2 0 2 0 0 0 1 0 2 1 2 1 0 0 2 0 0 1 0 2 1 0 1 2 1 0 1
 0 2 0 0 2 2 0 0 0 1 0 1 0 0 0 2 2 1 1 1 1 0 0 2 1 0 2 0 1 0 1 2 2 2 1 2 2
 1]
[1 2 0 0 1 2 2 1 1 1 2 0 1 0 0 0 1 0 1 1 2 1 0 0 2 0 0 1 0 2 1 0 1 1 1 0 1
 0 2 0 0 2 2 0 0 0 1 0 1 0 0 0 2 2 1 1 1 1 0 0 2 1 0 2 0 1 0 1 2 1 2 1 2 2
 1]
[ 予測結果 ]
      precision    recall  f1-score   support

     0         1.00        1.00        1.00         29
     1         0.86        1.00        0.92         24
     2         1.00        0.82        0.90         22

 avg / total         0.95        0.95        0.95         75

[ 正解率 ]
0.9466666666666667

[ 混同行列 ]
[[29  0  0]
 [ 0 24  0]
 [ 0  4 18]]
```

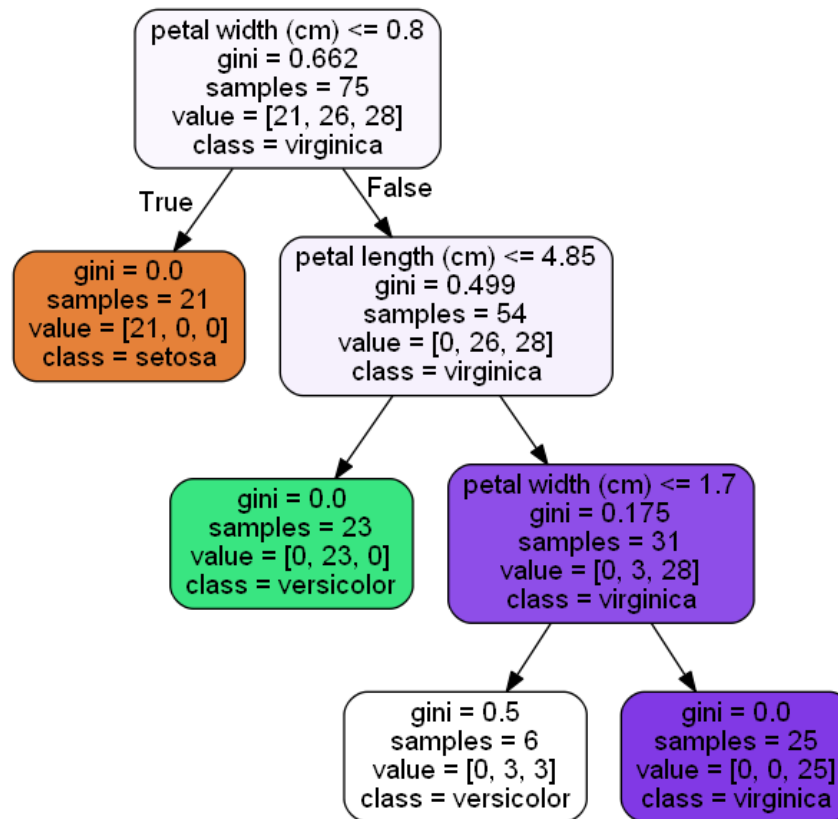
テストデータのラベル(正解値)

予測結果

生成された決定木①

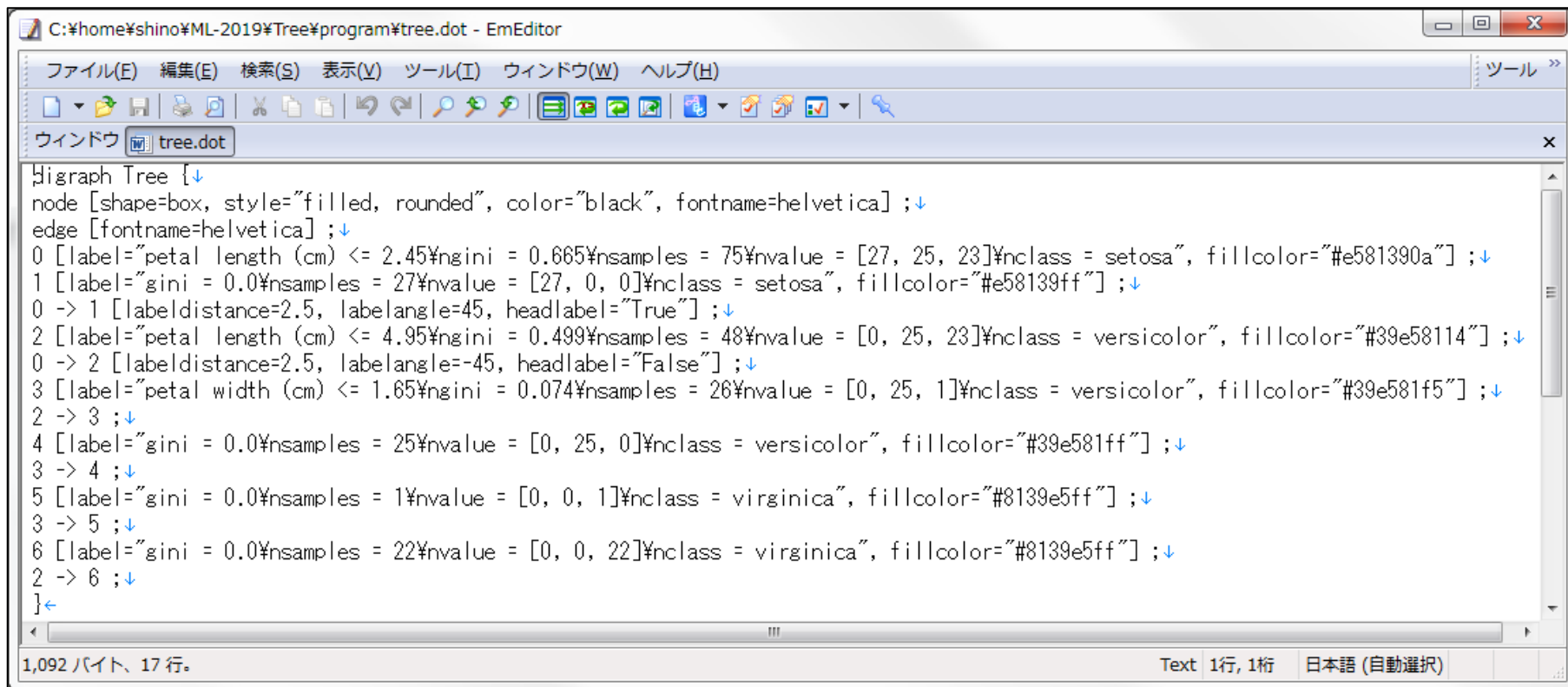
決定木の画像化

```
> dot -T png tree.dot -o tree.png
```



生成された決定木②

tree.dot



```
graph TD
    node [shape=box, style="filled, rounded", color="black", fontname=helvetica] ;
    edge [fontname=helvetica] ;
    0 [label="petal length (cm) <= 2.45%ngini = 0.665%nsamples = 75%nvalue = [27, 25, 23]%nclass = setosa", fillcolor="#e581390a"] ;
    1 [label="gini = 0.0%nsamples = 27%nvalue = [27, 0, 0]%nclass = setosa", fillcolor="#e58139ff"] ;
    0 --> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
    2 [label="petal length (cm) <= 4.95%ngini = 0.499%nsamples = 48%nvalue = [0, 25, 23]%nclass = versicolor", fillcolor="#39e58114"] ;
    0 --> 2 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
    3 [label="petal width (cm) <= 1.65%ngini = 0.074%nsamples = 26%nvalue = [0, 25, 1]%nclass = versicolor", fillcolor="#39e581f5"] ;
    2 --> 3 ;
    4 [label="gini = 0.0%nsamples = 25%nvalue = [0, 25, 0]%nclass = versicolor", fillcolor="#39e581ff"] ;
    3 --> 4 ;
    5 [label="gini = 0.0%nsamples = 1%nvalue = [0, 0, 1]%nclass = virginica", fillcolor="#8139e5ff"] ;
    4 --> 5 ;
    6 [label="gini = 0.0%nsamples = 22%nvalue = [0, 0, 22]%nclass = virginica", fillcolor="#8139e5ff"] ;
    2 --> 6 ;
    }
```

1,092 バイト、17 行。

Text 1行, 1桁 日本語 (自動選択)

Boston dataset

■ ボストンの住宅価格の回帰問題

用途	回帰
データ数	506
特徴量	13
目的変数	1



Bostonデータセットの学習(回帰木)

```
import numpy
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn import tree
from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
import matplotlib.pyplot as plt

# データのロード
boston = datasets.load_boston()

# 特徴量(13次元)
feature_name = boston.feature_names

# データ
data = boston.data

# 価格
price = boston.target
```

パッケージのimport

散布図の描画

学習データ, テストデータ

```
train_data, test_data, train_price, test_price = train_test_split(data, price, test_size=0.5, random_state=None)
```

回帰木

```
model = tree.DecisionTreeRegressor(criterion="mse", max_depth=3)
```

学習

```
model.fit(train_data, train_price)
```

指標は平均二乗誤差

決定木の深さ=3

予測(テストデータ)

```
predict = model.predict(test_data)
```

R^2 を求める

```
train_score = model.score(train_data, train_price)
```

```
test_score = model.score(test_data, test_price)
```

相関係数の二乗(R^2)の計算

```
print( " 学習データ:", train_score )
```

```
print( " テストデータ:", test_score )
```

散布図の描画(保存)

```
fig = plt.figure()  
plt.scatter( test_price , predict )  
plt.xlabel("Correct")  
plt.ylabel("Predict")  
fig.savefig("result.png")
```

x軸: 正解値
y軸: 予測値

「result.png」に保存

```
tree.export_graphviz(model, out_file="tree.dot",  
feature_names=boston.feature_name, filled=True, rounded=True)
```

回帰木の描画のために必要

DecisionTreeRegressor

```
from sklearn import tree
```

```
tree.DecisionTreeRegressor(criterion=分類指標,  
max_depth=決定木の深さ)
```

```
model = tree.DecisionTreeRegressor(criterion="mse", max_depth=3)
```

指標は平均二乗誤差

決定木の深さ=3

実行結果

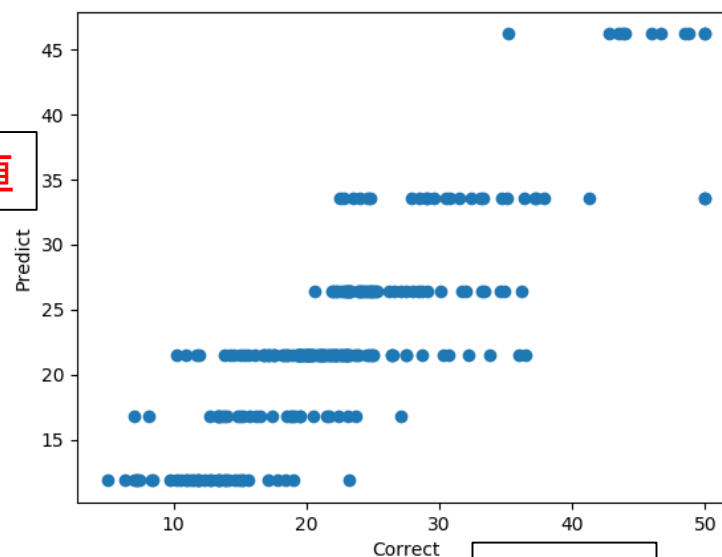
```
C:\Windows\system32\cmd.exe

C:\home\shino\ML-2019\Tree\program>python Boston_Tree.py
学習データ: 0.8564370957102287
テストデータ: 0.7075924685574708

C:\home\shino\ML-2019\Tree\program>
```

散布図

予測値

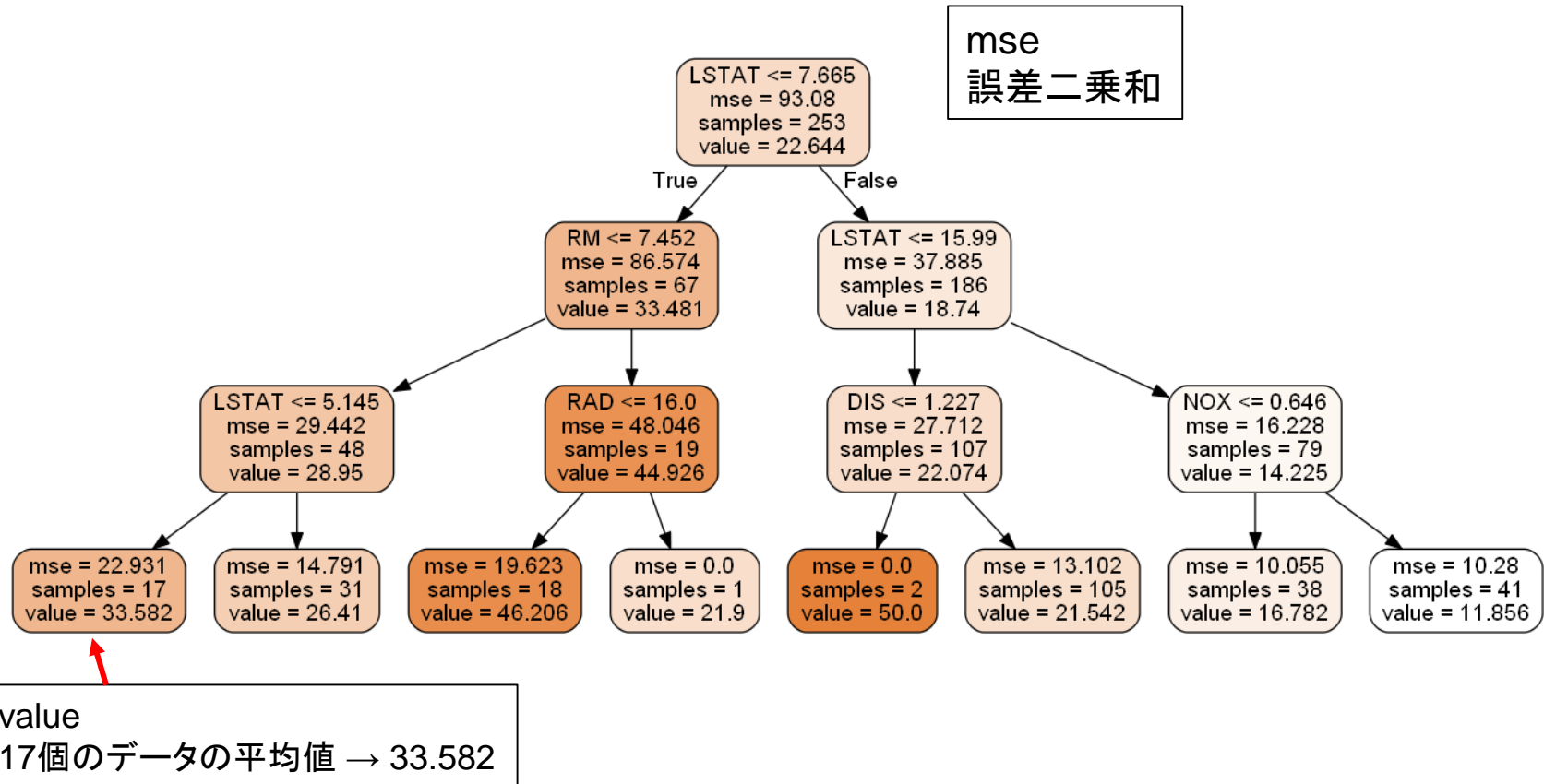


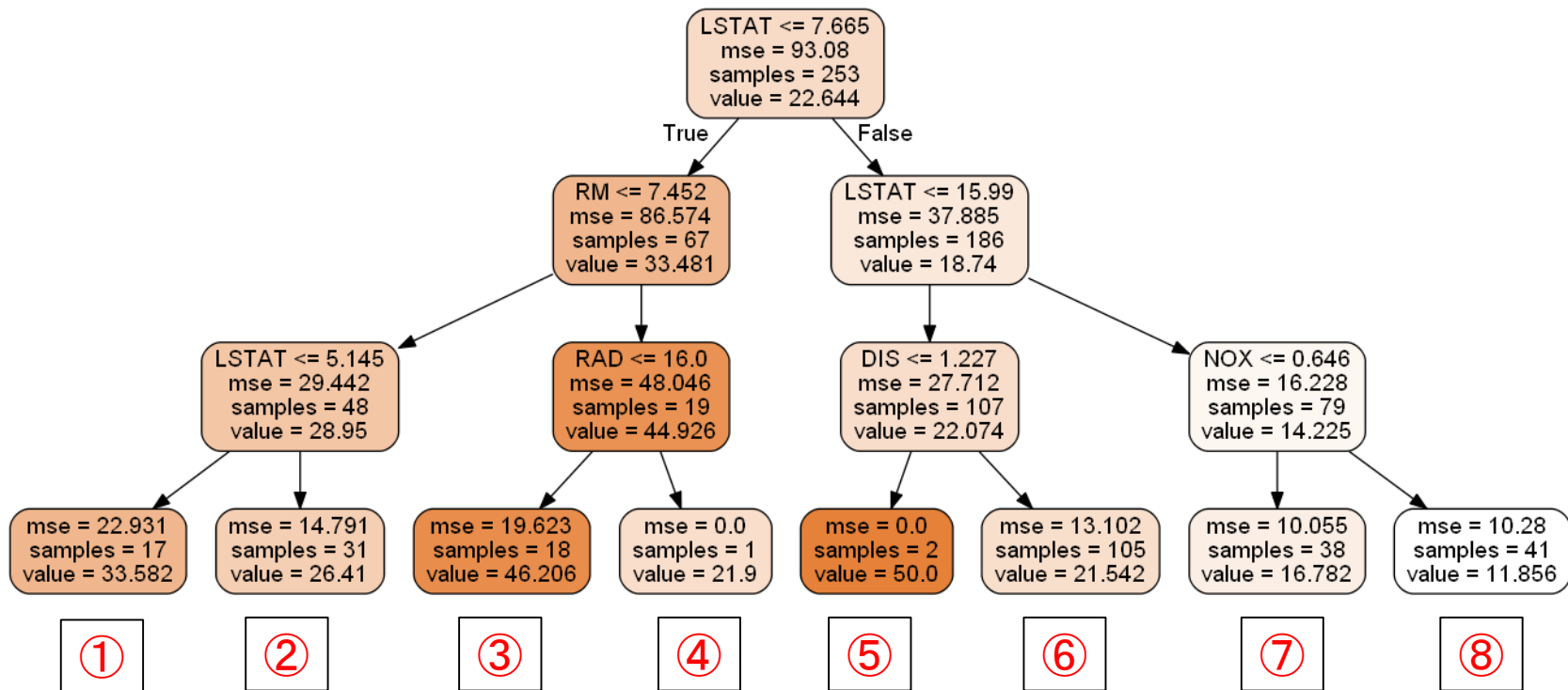
正解値

生成された回帰木

回帰木の画像化

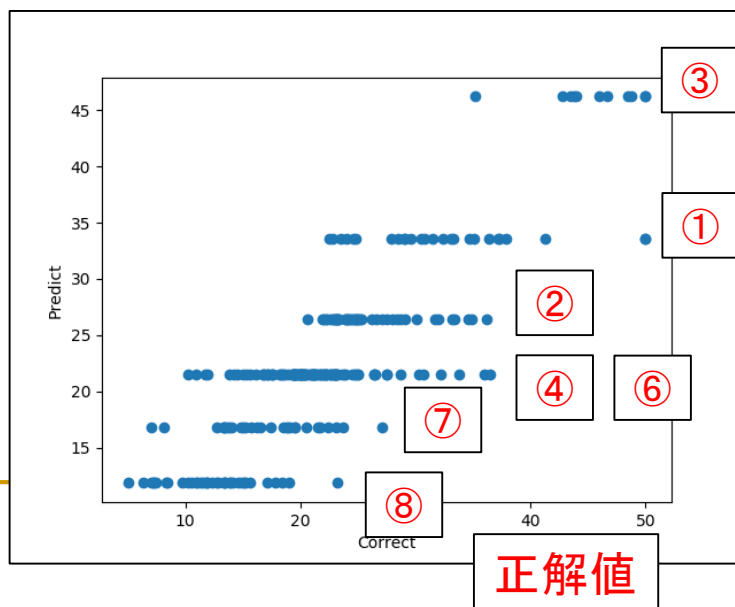
```
> dot -T png tree.dot -o tree.png
```





分類結果

予測値



回帰木での予測結果

- 予測らしい予測になっていない
- 改良方法
 - アンサンブル学習
 - ランダムフォレスト
 - 複数個の回帰木によって予測→平均値を予測値とする

参考文献

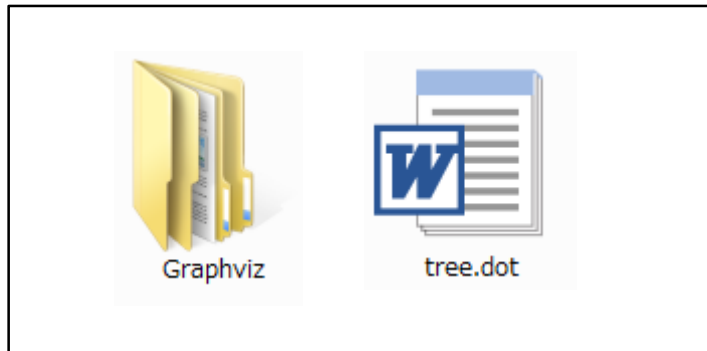
- Tom M Michell: Machine Learning, 1990
- 加藤直樹他: データマイニングとその応用, 朝倉書店, 2009
- 平井有三: はじめてのパターン認識, 森北出版株式会社, 2012
- 後藤正幸他: 入門 パターン認識と機械学習, コロナ社, 2014
- 株式会社システム計画研究所編: Pythonによる機械学習入門, オーム社, 2016
- 竹村彰通他: 機械学習, 朝倉書店, 2017
- 荒木雅弘: 機械学習入門, 森北出版株式会社, 2018

参考文献

- DecisionTreeClassifier
 - <https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
- DecisionTreeRegressor
 - <https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>

理工学ITCのPCで決定木を描画する場合

- Graphviz.zip をダウンロード
- tree.dot のあるフォルダーに展開



- コマンドプロンプト上で,
- > Graphviz¥bin¥dot.exe -T png tree.dot -o tree.png