

決定木

7月17日(日)WSL勉強会

宗政一舟

もくじ

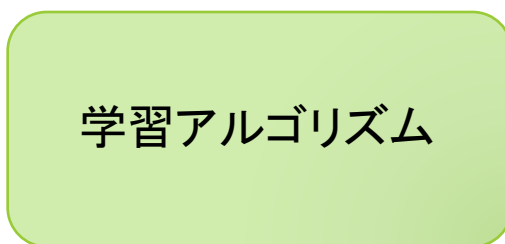
1. 決定木とは
2. 決定木学習のアルゴリズム
3. 分割の良さを表す指標
4. CARTアルゴリズム
5. ID3アルゴリズム
6. 枝刈り

1. 決定木とは

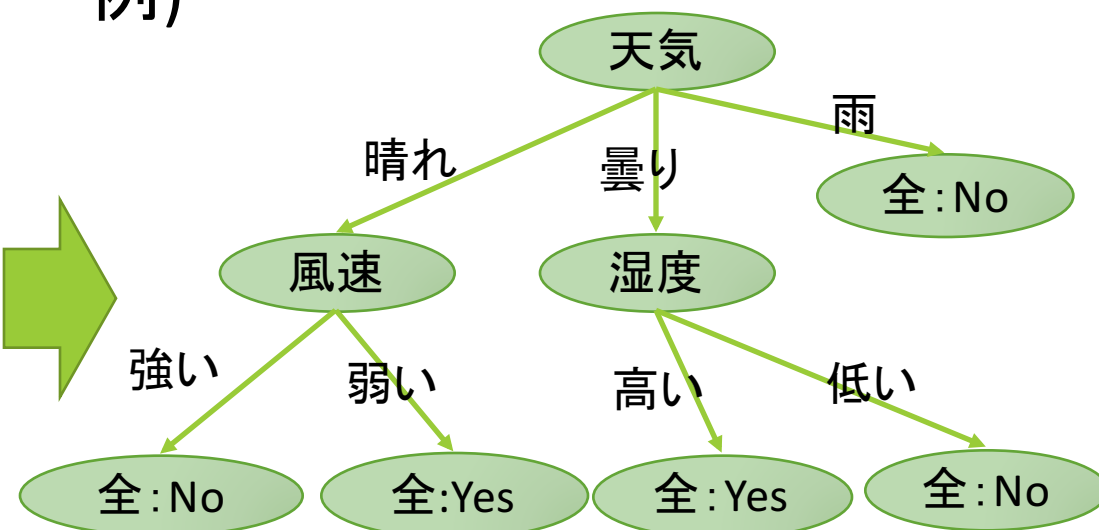
決定木とは

- 木構造をした決定を行うためのグラフ
- 与えられたデータから適切な決定木を作成することを決定木学習という

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No

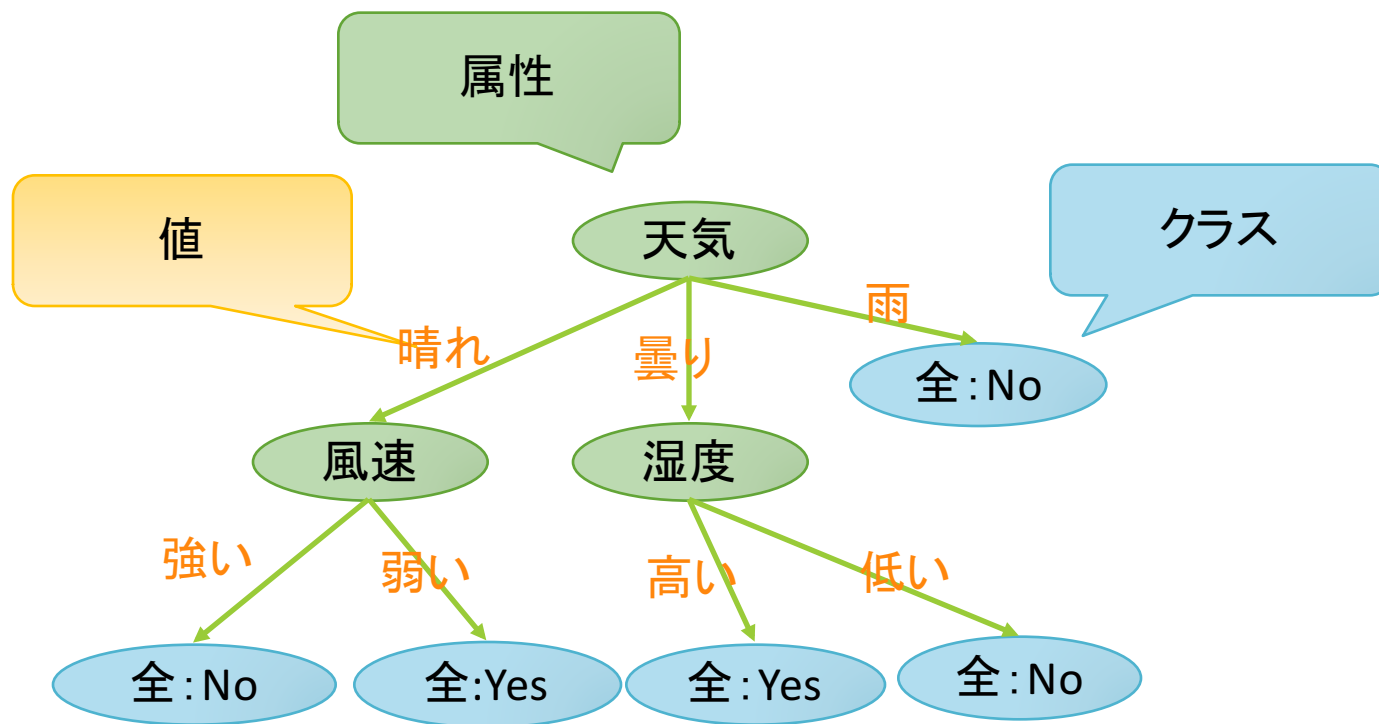


例)



決定木とは

- 属性：“根”を含む“葉”以外のノード
- 値：エッジ
- クラス：葉

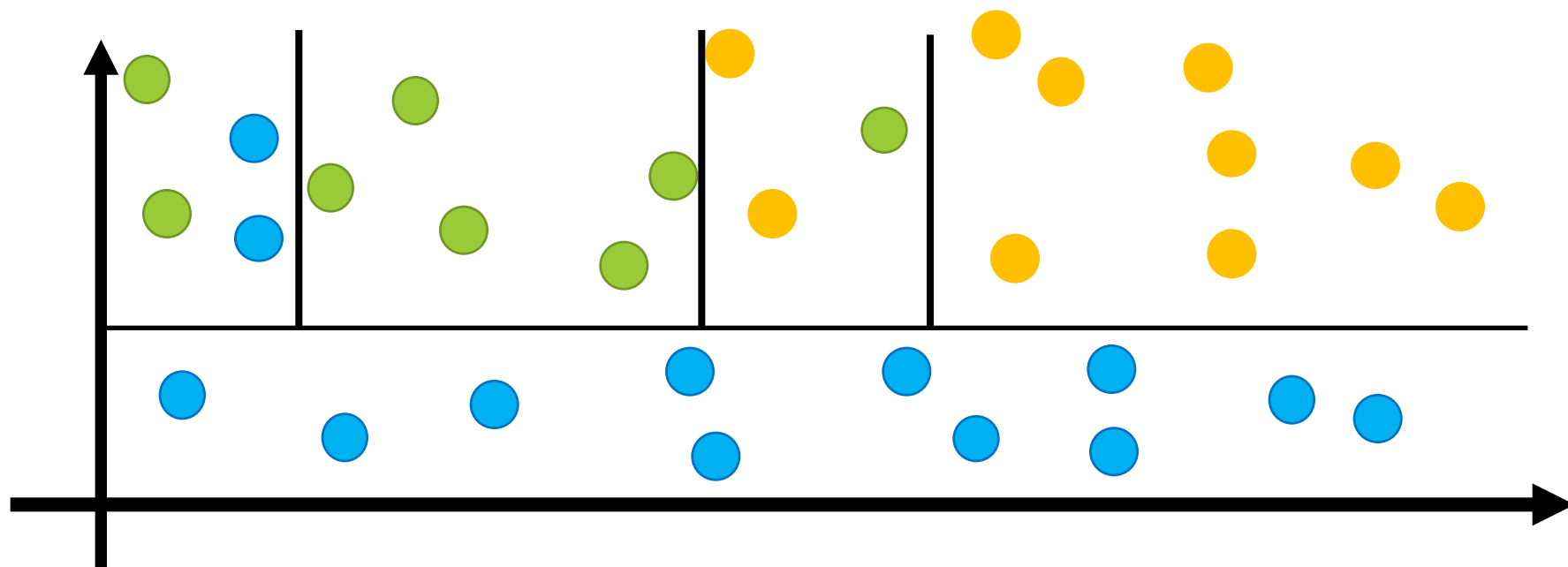


決定木の特徴

- 特定の結果をもたらす可能性の高いセグメントを見つける
 - 新しいデータに対して分類、予測を行う
- 特定の結果をもたらす要因・ルールを見つける
 - 決定木の構築

決定木の特徴

- 決定木による分類例



2.決定木学習アルゴリズム

決定木学習アルゴリズム

複数の属性とクラスを持つ事例を考え、その事例の集合を T とする。 T を学習事例と呼ぶ。

この時、事例の持つクラスは $\{C_1, C_2, \dots, C_j, \dots, C_n\}$ とする。 n はクラスの個数である。

The diagram shows a table representing a dataset T . The table has four columns: 天気 (Weather), 風速 (Wind Speed), 湿度 (Humidity), and 花見 (Hanami). The first three columns are grouped under the label '属性' (Attributes), and the last column is labeled 'クラス' (Class). The table contains five rows of data. A red bracket on the left side of the table is labeled '事例集合 T '. A red arrow points from the text ' $\{C_{Yes}, C_{No}\}$ であるから $n = 2$ となる' to the '花見' column.

属性			クラス
天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No

$\{C_{Yes}, C_{No}\}$ であるから
 $n = 2$ となる

決定木学習アルゴリズム : Step1

T がすべて同一のクラス C_j であるならば、 T に対する決定木は葉であり、そのクラスは C_j とする

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No



天気	風速	湿度	花見
曇り	弱い	高い	Yes
晴れ	弱い	高い	Yes

C_{Yes} となる



天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	強い	低い	No
雨	弱い	高い	No

C_{No} となる

決定木学習アルゴリズム : Step2

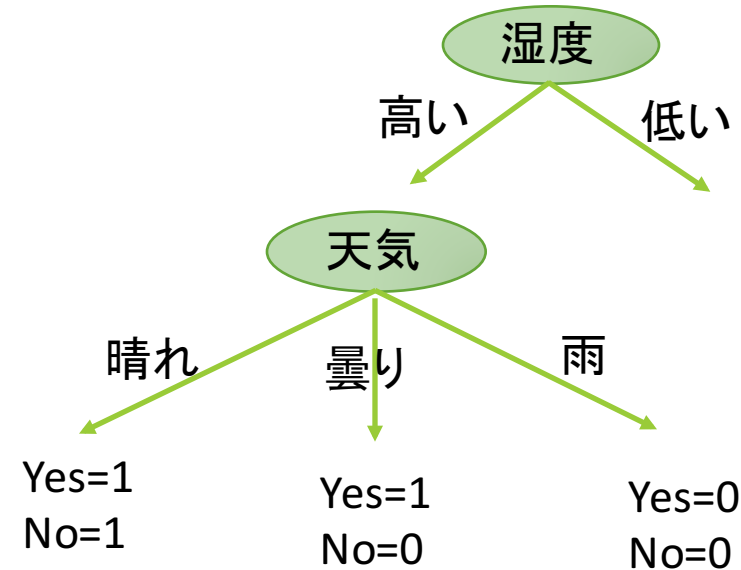
T が対象の属性を含まない場合、Step1と同様に葉とするが、そのクラスは事例からは決められない

ものによって処理が変わる

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	低い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No

湿度:「高い」で分割後

天気	風速	湿度	花見
晴れ	強い	高い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No



例えば親ノードの多数決でクラスを決める

決定木学習アルゴリズム : Step3

T が複数のクラスを有する場合には、ある属性を選択して、その属性が持ちうる属性値により、事例 T を分割する。

そして分割された事例集合の各々についてStep1~ Step3を再帰的に実行

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No

YesとNoの2種類のクラスがあるので分割させる必要がある

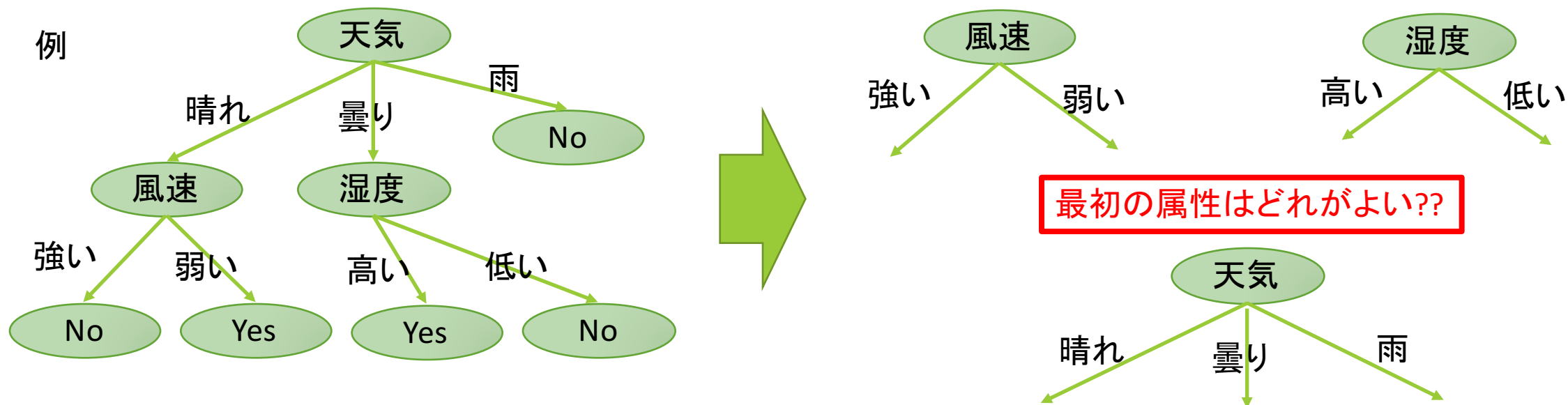
今日紹介する学習アルゴリズム

- 決定木学習で、分岐させる属性の決定方法が問題になる
- 生成された決定木の性能に大きな影響を与える
- 代表的な決定木作成アルゴリズム
 - ID3アルゴリズム
 - CARTアルゴリズム

3.分割の良さを表す指標

分割の良さを表す指標

- 実際のどのようにして属性を決め、分割していくのがよいのか?
- 例では最初の属性は”天気”であったが、実際はもっと良い属性があるかもしれない



分割の良さを表す指標

- 情報ゲイン ← こっちを扱う
- ジニ不純度(* appendix)

情報ゲイン

- 事例 T の分割前と分割後のエントロピーの差を利用
- 事例 T を $\{B_1, B_2, \dots, B_N\}$ に分割するときの情報ゲイン $G(T)$

エントロピー

$$E(T) = - \sum_{c \in C} p(c) \log p(c) \cdots (1)$$

$p(c)$: 全体のクラス C に対するクラス c の割合

情報ゲイン

$$G(T) = E(T) - \sum_{b \in B} p(B_b) E(B_b) \cdots (2)$$

情報ゲインが大きい方が良い分割になる

例)「天気」の情報ゲイン

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No

T

$$E(T) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.972$$

晴れ	強い	高い	No
晴れ	弱い	高い	Yes

$$E(B_{\text{晴れ}}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

曇り	弱い	高い	Yes
曇り	強い	低い	No

$$E(B_{\text{曇り}}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

雨	弱い	高い	No
---	----	----	----

$$E(B_{\text{雨}}) = -\frac{1}{1}\log\frac{1}{1} = 0$$

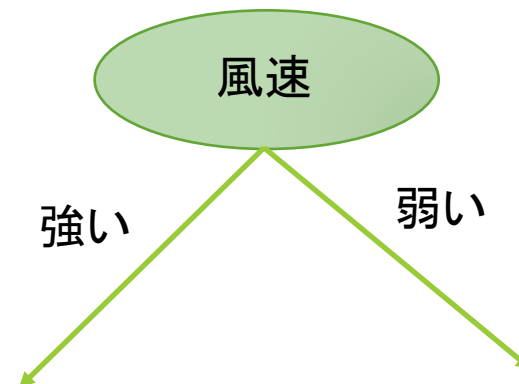
$G(T)$

$$\begin{aligned}
 &= E(T) - \sum_{b \in B} p(B_b)E(B_b) \\
 &= 0.972 - \left(\frac{2}{5} \cdot 1\right) + \left(\frac{2}{5} \cdot 1\right) + \left(\frac{1}{5} \cdot 0\right) \\
 &= 0.172
 \end{aligned}$$

風速、湿度についても
情報ゲインを算出
最大値が属性となる

情報ゲイン

- 天気で分割した時
 - 情報ゲイン: 0.172
- 風速で分割した時
 - 情報ゲイン: **0.423**
- 湿度で分割した時
 - 情報ゲイン: 0.172



最初の属性は**“風速”**が最適となる

4.ID3アルゴリズム

ID3アルゴリズム

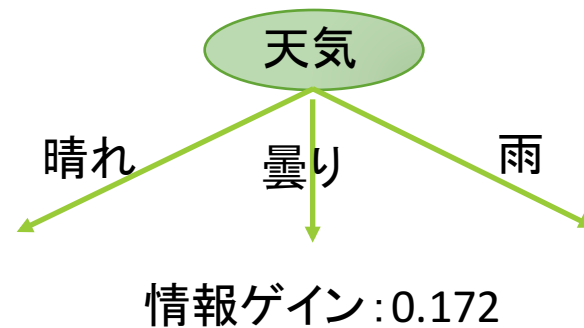
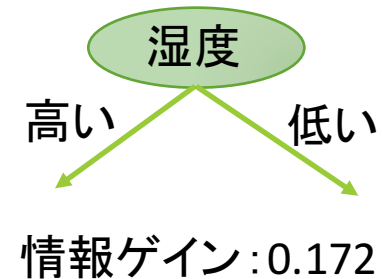
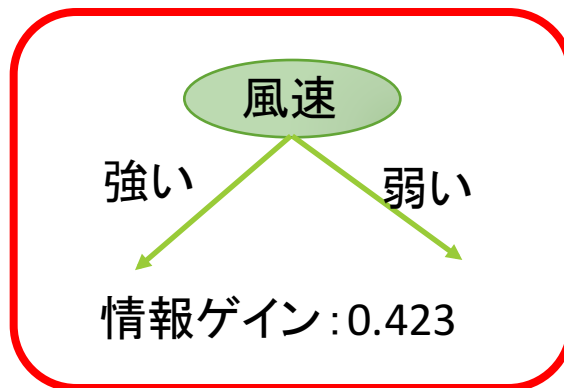
- Iterative Dichotomiser 3 Algorithm のこと
- 属性値の種類が3種類以上あっても問題がない
- 属性値はカテゴリである必要がある

IDアルゴリズム

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No



情報ゲイン最大

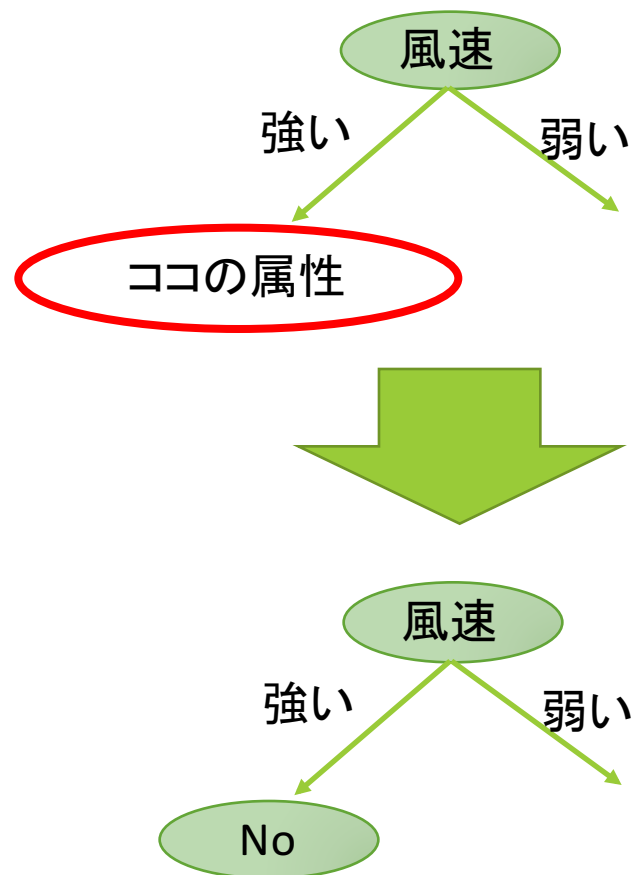


IDアルゴリズム

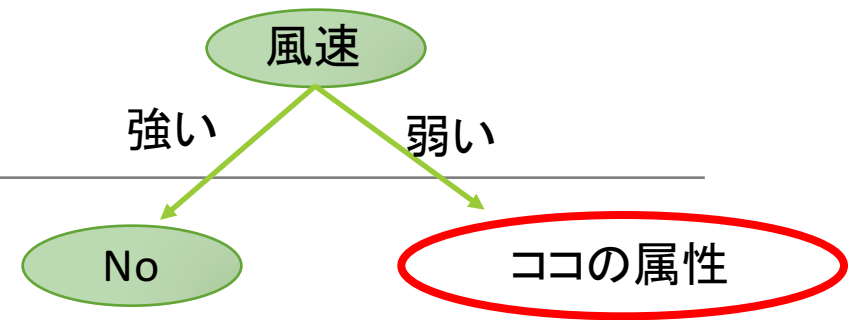
- ・風速:「強い」で分割後

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	強い	低い	No

クラスの色がNoになったので
こちら側の分岐は終了

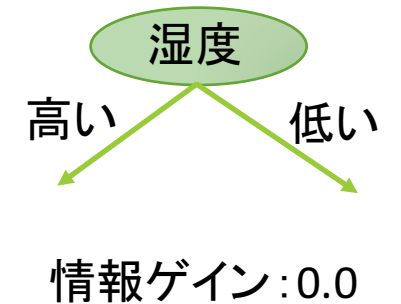
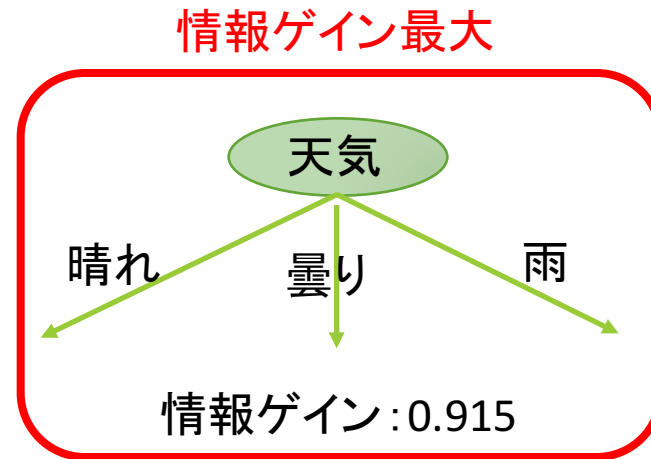


IDアルゴリズム



・風速:「弱い」で分割後

天気	風速	湿度	花見
曇り	弱い	高い	Yes
晴れ	弱い	高い	Yes
雨	弱い	高い	No

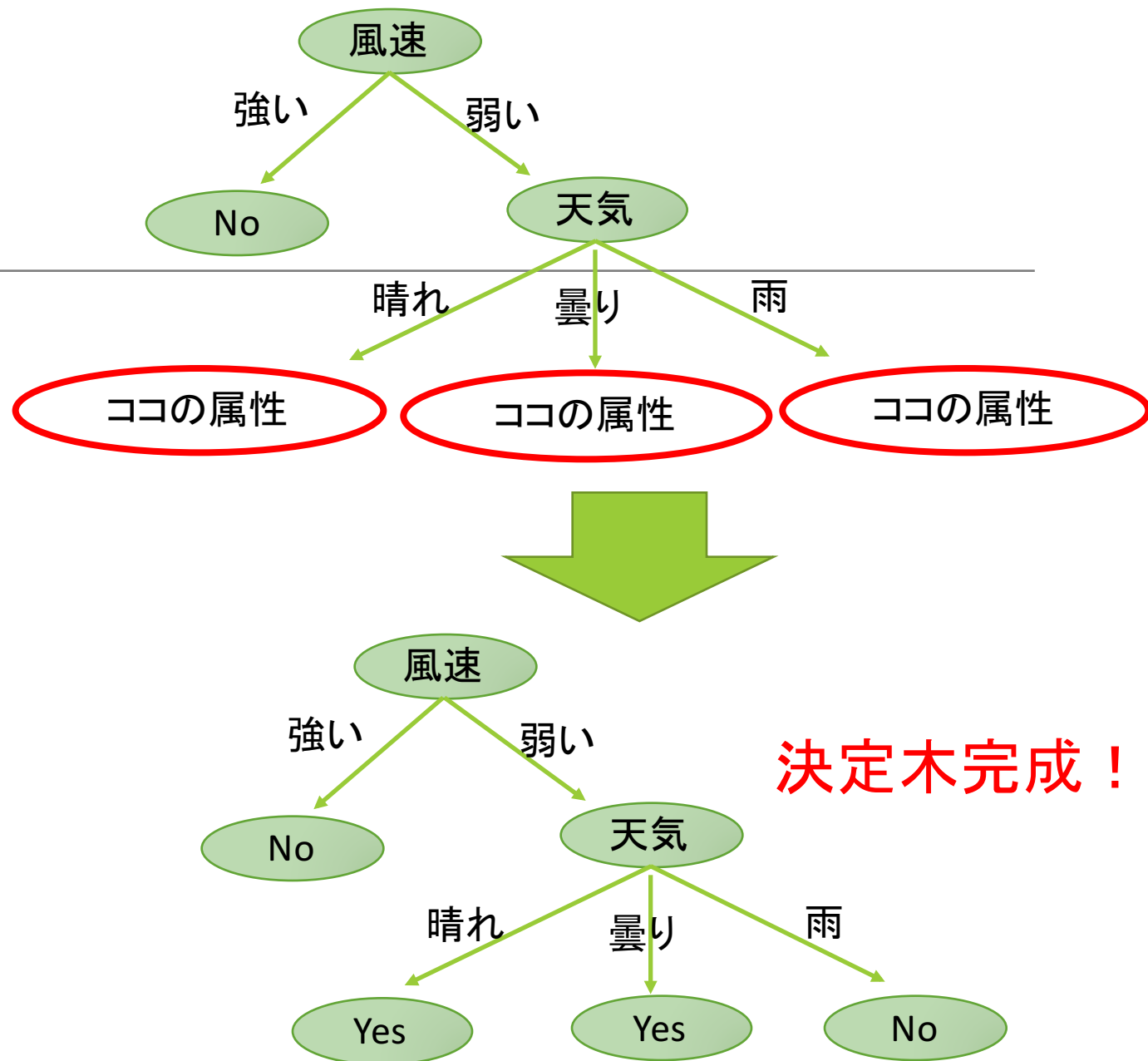


IDアルゴリズム

・風速:「弱い」で分割後

天気	風速	湿度	花見
曇り	弱い	高い	Yes
晴れ	弱い	高い	Yes
雨	弱い	高い	No

それぞれの属性に対して
クラスが同一になるので
終了



5. CARTアルゴリズム

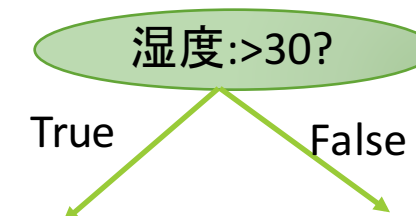
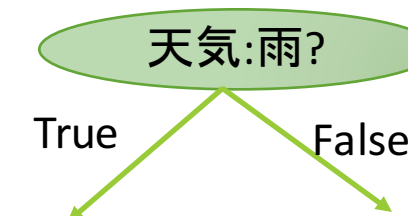
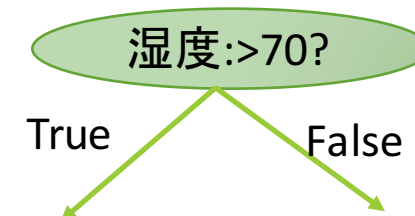
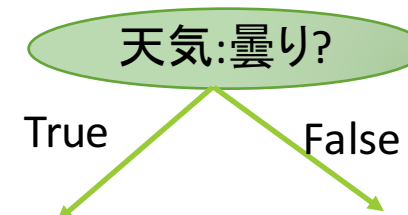
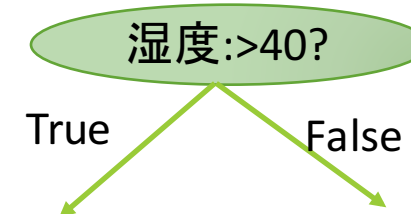
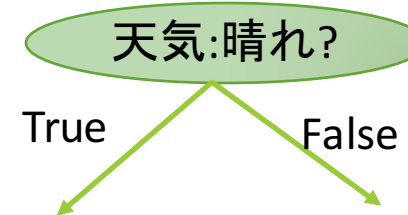
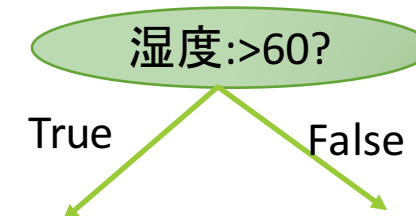
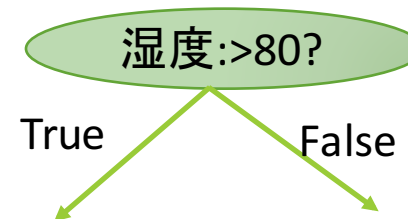
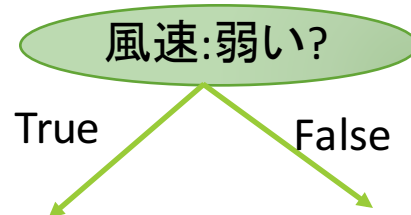
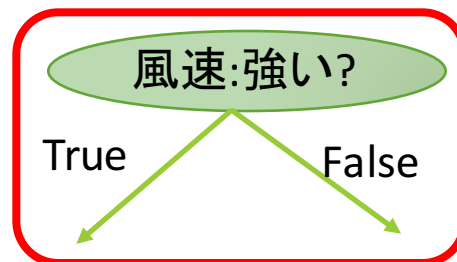
CARTアルゴリズム

- Classification And Regression Tree Algorithmのこと
- 2分岐でしか考えていない
- 属性値はカテゴリでも、数値でも対応

CARTアルゴリズム

天気	風速	湿度	花見
晴れ	強い	60	No
曇り	弱い	40	Yes
曇り	強い	70	No
晴れ	弱い	30	Yes
雨	弱い	80	No

情報ゲイン最大



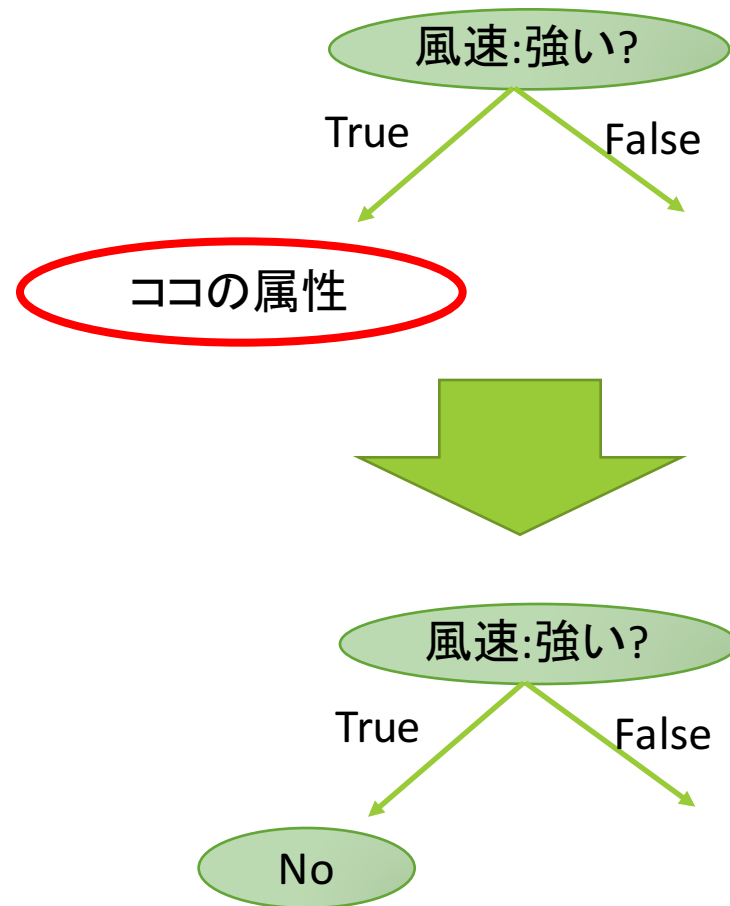
この中から、情報ゲインが最も高いものを最初の属性にする。

CARTアルゴリズム

- ・風速:「強い?」 True で分割後

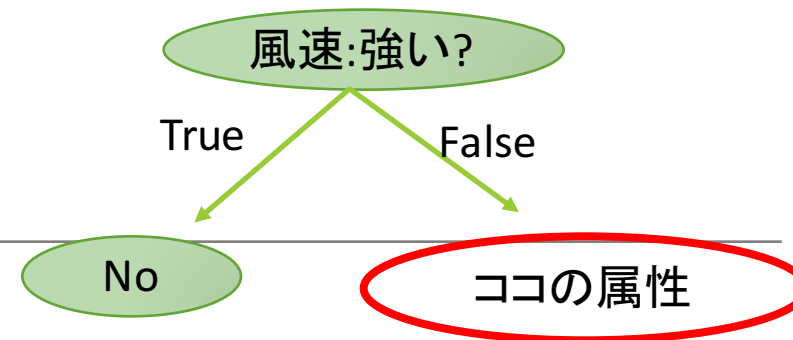
天気	風速	湿度	花見
晴れ	強い	60	No
曇り	強い	70	No

クラスの色がすべてNo になったので
こちら側は終了

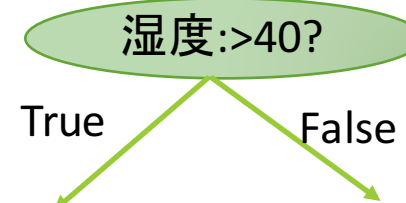
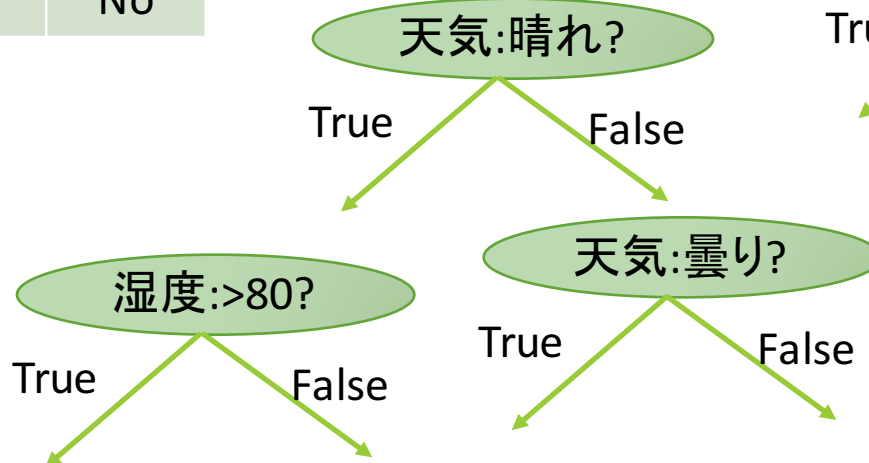


CARTアルゴリズム

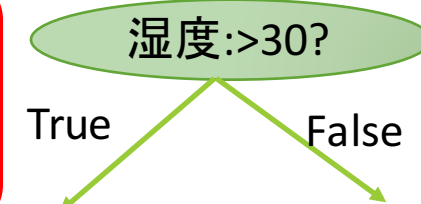
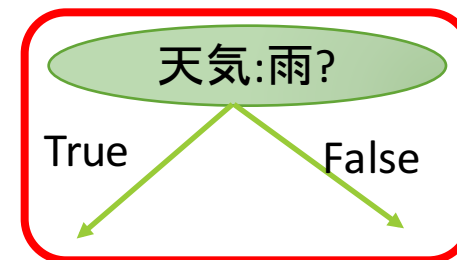
- 風速:「強い?」 True で分割後



天気	風速	湿度	花見
曇り	弱い	40	Yes
晴れ	弱い	30	Yes
雨	弱い	80	No



情報ゲイン最大

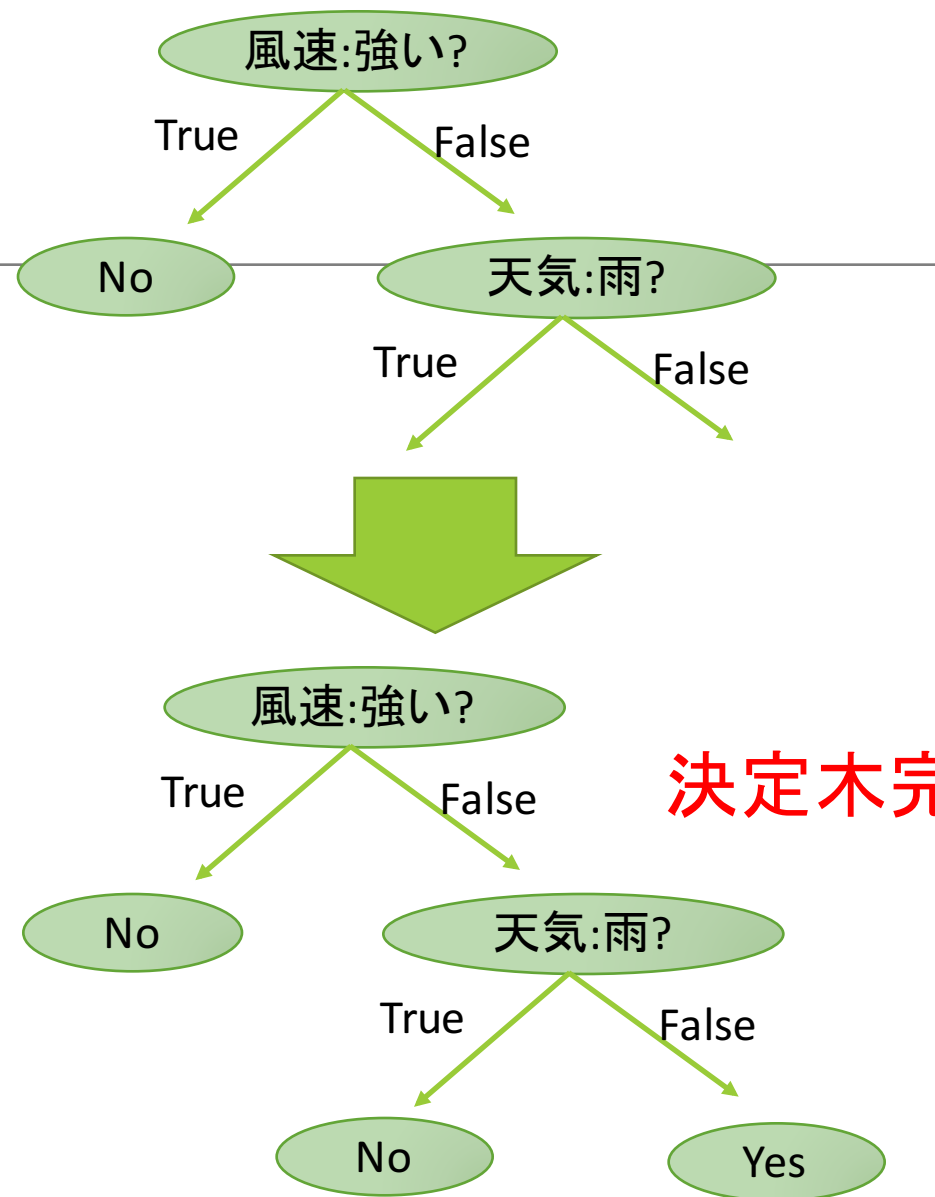


CARTアルゴリズム

- 風速:「強い?」 True で分割後

天気	風速	湿度	花見
曇り	弱い	40	Yes
晴れ	弱い	30	Yes
雨	弱い	80	No

クラスの値が
雨であった時:すべてYes
雨でなかった時:すべてNo
になったので終了

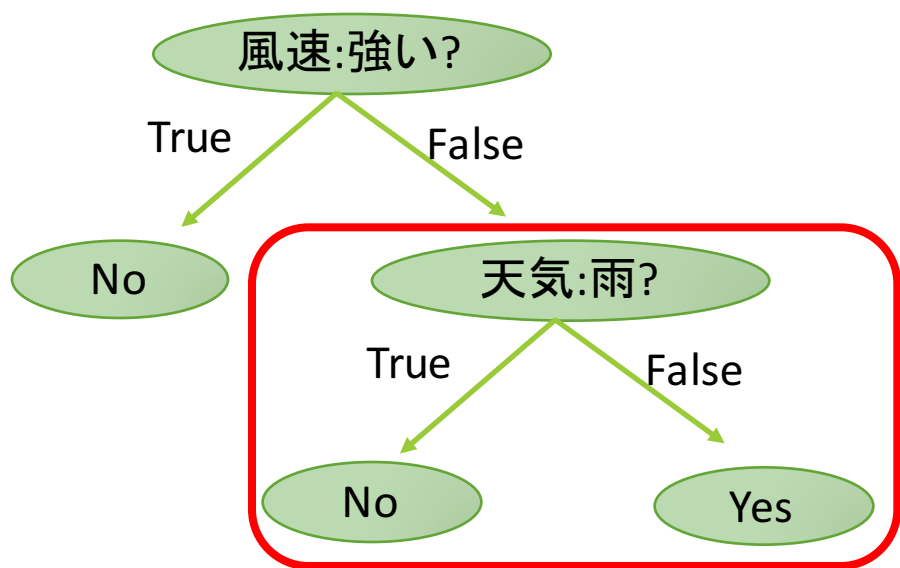


6. 枝刈り

枝刈り

- 決定木学習の問題の中に過学習がある
- 紹介したアルゴリズム(ID3,CART)はクラスを唯一にするまで分割する
- 最後まで分割を行えば、少しずつエントロピーは減少するが学習データに対して厳密になりすぎ汎化性が失われる可能性がある
- 枝刈りでは親ノードを共有するノード群について、合成した時(平均値)のエントロピーの上昇が指定の未満かチェックする

枝刈り



閾値: 1.0

親ノード(分割前)のエントロピー

天気	風速	湿度	花見
曇り	弱い	40	Yes
晴れ	弱い	30	Yes
雨	弱い	80	No

$$-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

子ノード(分割後)のエントロピー

天気	風速	湿度	花見
曇り	弱い	40	Yes
晴れ	弱い	30	Yes

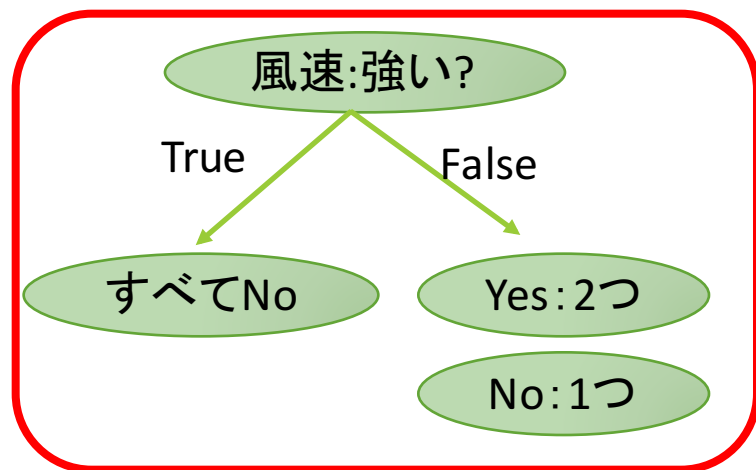
$$-\frac{2}{2}\log\frac{2}{2} = 0.0$$

天気	風速	湿度	花見
雨	弱い	80	No

$$-\frac{1}{1}\log\frac{1}{1} = 0.0$$

$$0.918 - \frac{(0.0 + 0.0)}{2.0} = 0.918 < 1.0 \quad \text{枝刈り確定!!}$$

枝刈り



閾値: 1.0

親ノード(分割前)のエントロピー

天気	風速	湿度	花見
晴れ	強い	60	No
曇り	弱い	40	Yes
曇り	強い	70	No
晴れ	弱い	30	Yes
雨	弱い	80	No

子ノード(分割後)のエントロピー

天気	風速	湿度	花見
曇り	弱い	40	Yes
晴れ	弱い	30	Yes
雨	弱い	80	No

$$-\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.970$$

天気	風速	湿度	花見
晴れ	強い	60	No
曇り	強い	70	No

$$-\frac{2}{2}\log\frac{2}{2} = 0.0$$

$$-\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.918$$

$$0.970 - \frac{(0.918 + 0.0)}{2.0} = 0.511 < 1.0 \text{ 枝刈り確定!!}$$

枝刈り

- 閾値を1.0に設定した時、分割しないほうがいいらしいです

Yes: 2つ, No: 3つ

ご清聴ありがとうございました。

appendix

ジニ不純度

- 集合をある属性で分割し、属性値の中でランダムにひとつを当てはめる場合の期待誤差率
- 値が高いほど集合の要素はバラついている

$$GINI = \sum_{c \in C} \{p(c) \cdot p(\bar{c})\} \cdots (3)$$

$c \in C$: ある属性で分割した時のクラスの集合

$p(c)$: クラス集合 C 内の要素 c が選ばれる確率

$p(\bar{c})$: 要素 c に対して誤った属性の値あてはめられる場合

例)「天気」のジニ不純度

天気	風速	湿度	花見
晴れ	強い	高い	No
曇り	弱い	高い	Yes
曇り	強い	低い	No
晴れ	弱い	高い	Yes
雨	弱い	高い	No



属性: 天気
ランダムに選んだ属性値: 晴れ

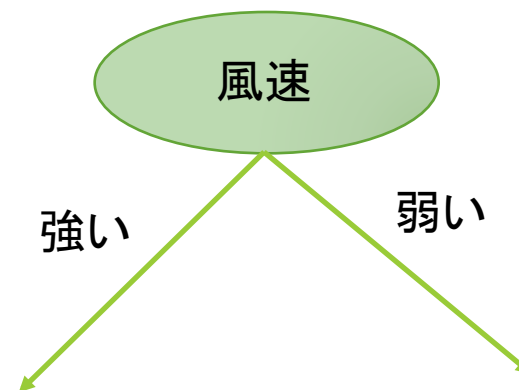
天気	風速	湿度	花見
晴れ	強い	高い	No
晴れ	弱い	高い	Yes



$$\begin{aligned} GINI &= \sum_{c \in C} \{p(c) \cdot p(\bar{c})\} \\ &= \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) \\ &= 0.5 \end{aligned}$$

ジニ不純度

- 天気で分割した時
 - ランダムな属性値「晴れ」
 - ジニ不純度: 0.5
- 風速で分割した時
 - ランダムな属性値「弱い」
 - ジニ不純度: **0.44**
- 湿度で分割した時
 - ランダムな属性値「高い」
 - ジニ不純度: 0.5



最初の属性は**“風速”**が最適となる