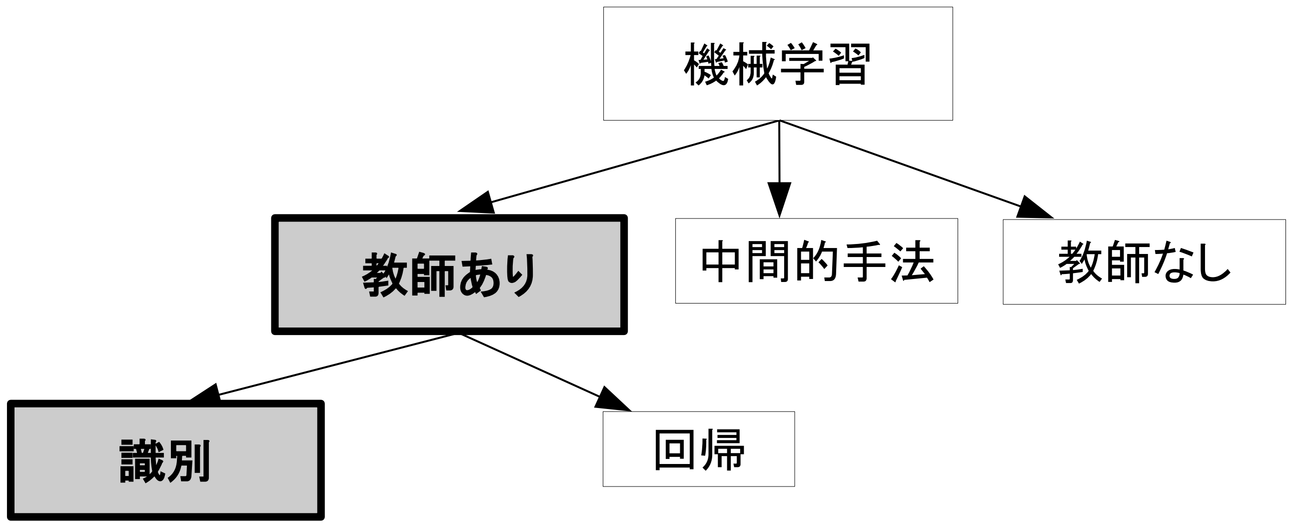
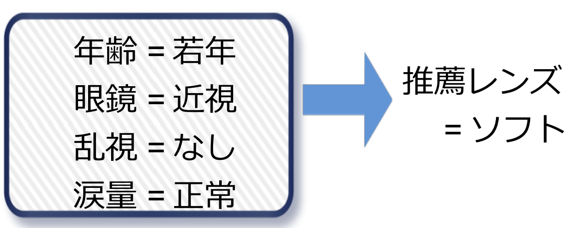
Section 2

- 識別1 (3,4章)



- カテゴリ特徴



3.　識別　－概念学習－

- 学習データ
  (テニスをする日；weather.nominal.arff)

Viewer

Relation: weather.symbolic

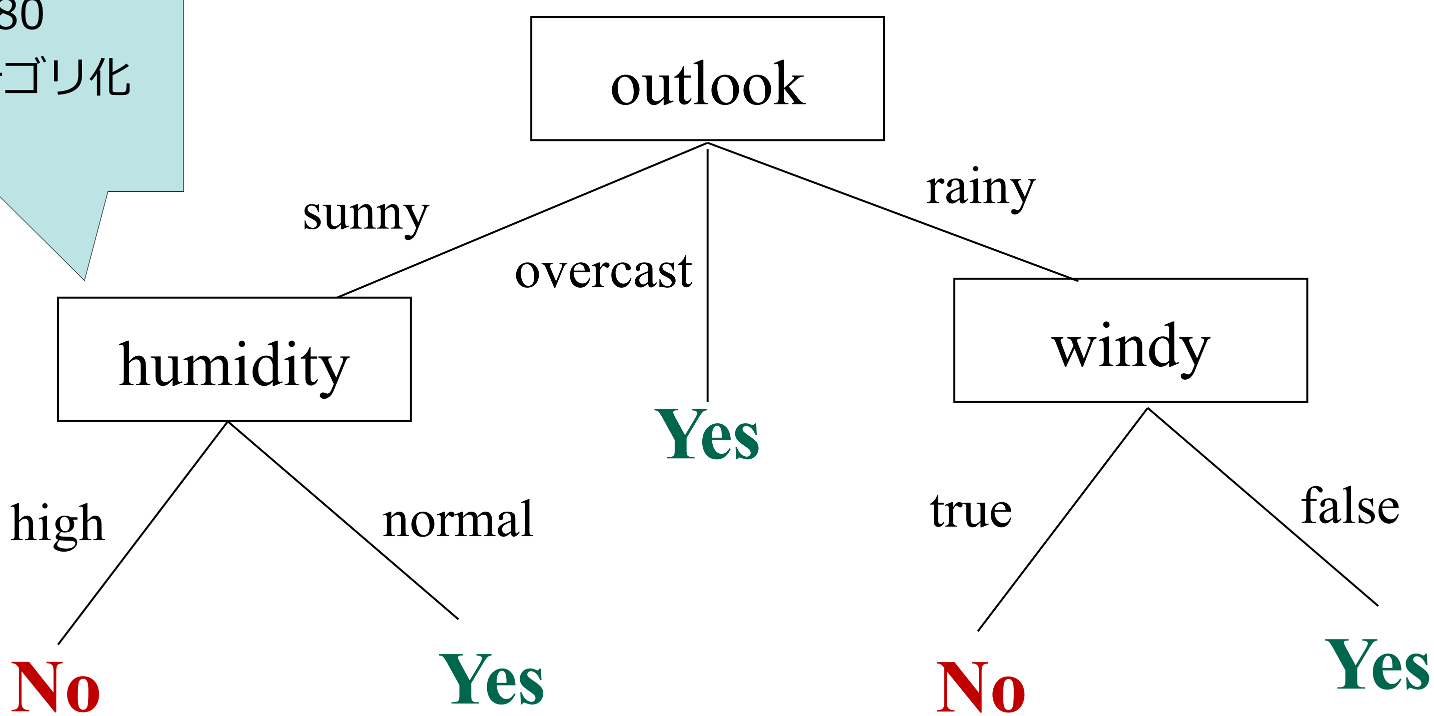
No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: **play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Add instance　Undo　OK　Cancel

3.4 決定木の学習

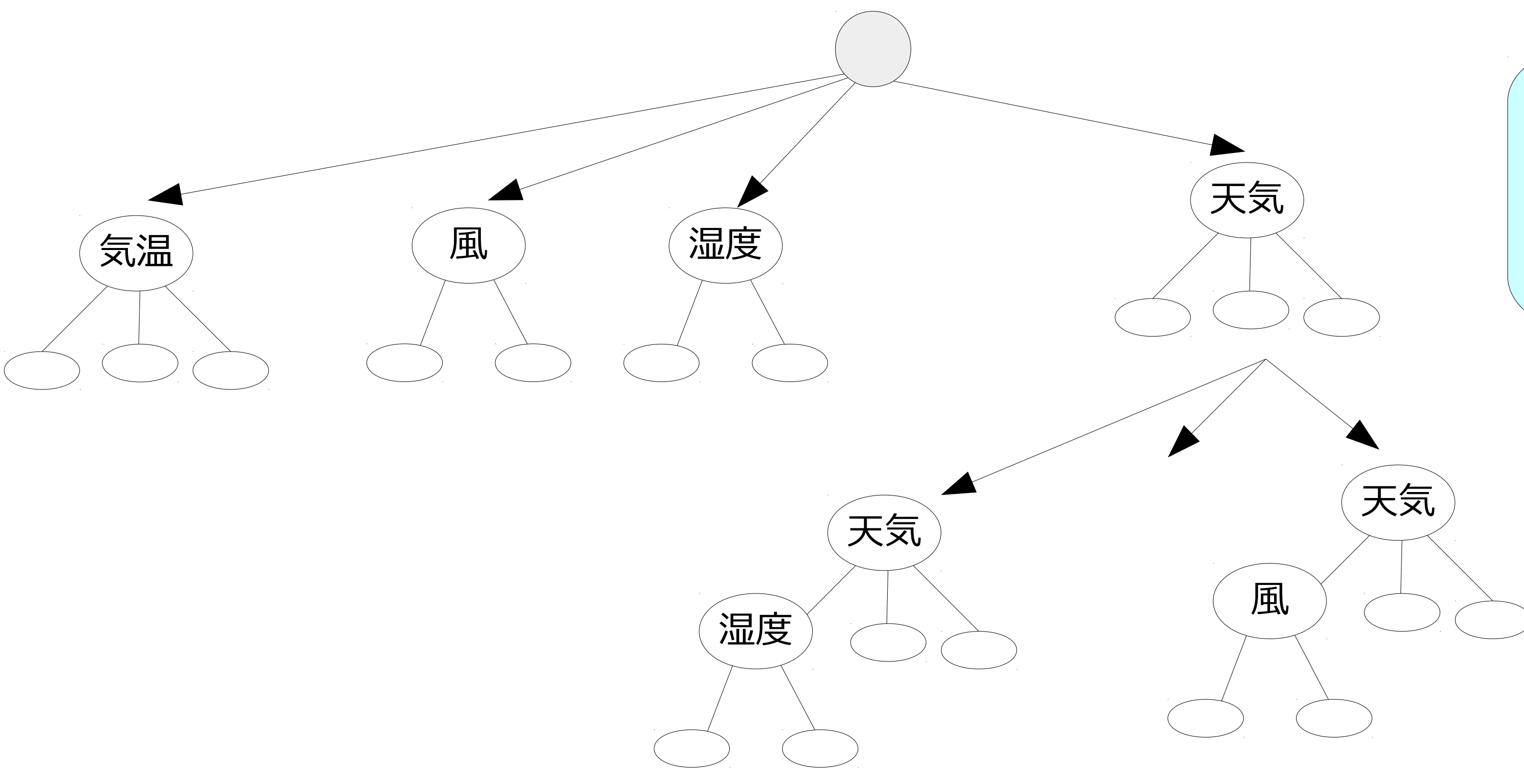
- 結果として得られる決定木

数値データの場合は、
humidity > 80
のようにカテゴリ化
する。



3.4 決定木の学習

- 決定木学習の考え方
  - 節はデータを分割する条件を持つ
    - できるだけ同一クラスのデータが偏るように
  - 分割後のデータ集合に対して、同様の操作を行う
  - 全ての葉が単一クラスの集合になれば終了

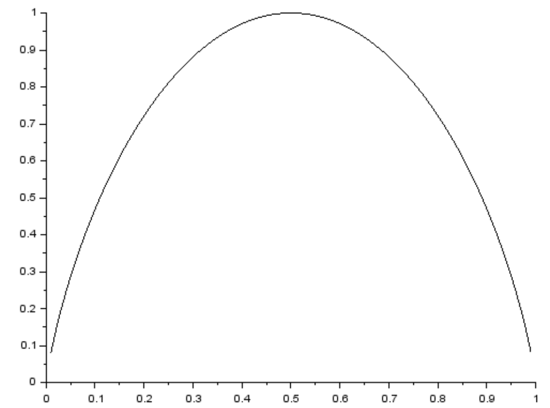


この手順に従うと、一般には小さな木ができる

バイアス
複雑な説明よりも単純な説明の方が汎用性が高い

属性の分類能力 (1/2)

- 分類能力の高い属性を決定する方法
  - その属性を使った分類を行うことによって、なるべくきれいにクラスが分かれるように
  - エントロピー
    - データ集合 S の乱雑さを表現
    - 正例の割合：p^+ ，　負例の割合：p^-
    - エントロピーの定義



$$Entropy(S) = -p^+ \log p^+ - p^- \log p^-$$

属性の分類能力 (2/2)

- 情報獲得量
  - 属性 A を用いた分類後のエントロピーの減少量
  - 値 v を取る訓練例の集合：Sv
  - Sv の要素数：|Sv|
  - 情報獲得量の定義

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy(Sv)$$

計算例

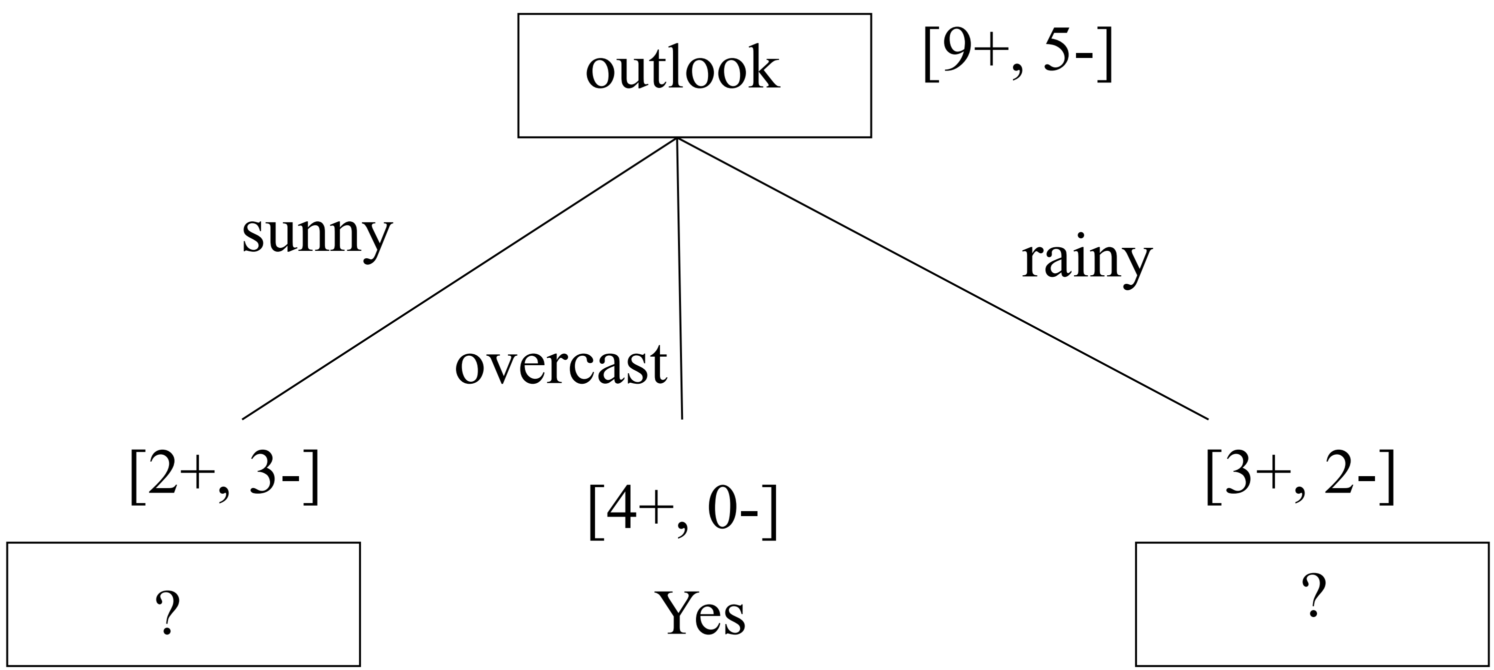
- 情報獲得量

  Gain(S, outlook)=0.246

  Gain(S, humidity)=0.151

  Gain(S, windy)=0.048

  Gain(S, temperature)=0.029



バイアスの検討

なぜ単純な木の方がよいか

- オッカムの剃刀
  「データに適合する最も単純な仮説を選べ」
  - 複雑な仮説
    → 表現能力が高い
    → 偶然にデータを説明できるかもしれない
  - 単純な仮説
    → 表現能力が低い
    → 偶然にデータを説明できる確率は低い
    → でも説明できた！
    → **必然**

連続値属性の扱い

- 連続値 A を持つ属性から真偽値 (A < c?) を値とするノードを作成

→ c をどうやって決めるか

気温	40	48	60	72	80	90
playTennis	No	No	Yes	Yes	Yes	No

c=(48+60)/2
=54

c=(80+90)/2
=85

情報獲得量の多い方

4.1 統計的識別とは

max $f(x)$: $f(x)$ の最大値
argmax $f(x)$: $f(x)$ が最大となる x

- 最大事後確率則による識別

$$C_{MAP} = \arg\max_i P(\omega_i|\boldsymbol{x})$$

$\boldsymbol{x}$ ：特徴ベクトル

$\omega_i \quad (1 \leq i \leq c)$ ：クラス

  - データから直接的にこの確率を求めるのは難しい

- ベイズの定理 $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

$$\begin{aligned} C_{MAP} &= \arg\max_i P(\omega_i|\boldsymbol{x}) \\ &= \arg\max_i \frac{P(\boldsymbol{x}|\omega_i)P(\omega_i)}{P(\boldsymbol{x})} \\ &= \arg\max_i P(\boldsymbol{x}|\omega_i)P(\omega_i) \end{aligned}$$

4.1 統計的識別とは

- ベイズ統計とは
  - 結果から原因を求める
  - ベイズ識別
    - 観測結果 $\boldsymbol{x}$ から、それが生じた原因 ω_i を求める
    - 通常、確率が与えられるのは原因→結果（尤度）
    - ベイズ識別では、事前分布 $P(\omega_i)$ が、観測によって事後分布 $P(\omega_i \mid \boldsymbol{x})$ に変化したと考えることができる

4.1 統計的識別とは

- 事前確率 $P(\omega_i)$
  - 特徴ベクトルを観測する前の、各クラスの起こりやすさ
- 事前確率の最尤推定

$$P(\omega_i) = \frac{n_i}{N}$$

N: 全データ数、n_i: クラス ω_i のデータ数

4.1 統計的識別とは

- 尤度 $P(\boldsymbol{x}|\omega_i)$
  - 特定のクラスから、ある特徴ベクトルが出現する尤もらしさ
- d 次元ベクトルの場合の最尤推定
  - 値の組合せがデータ中に出現しないもの多数

Viewer

Relation: weather.symbolic

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

Weka の weather.nominal データ

3×3×2×2=36 種類の組合せ

4.2.2 ナイーブベイス識別

- ナイーブベイズの近似
  - 全ての特徴が独立であると仮定

$$P(\boldsymbol{x}|\omega_i) = P(x_1, \ldots, x_d|\omega_i)$$

$$\approx \prod_{j=1}^{d} P(x_j|\omega_i)$$

$$C_{NB} = \arg\max_i P(\omega_i) \prod_{j=1}^{d} P(x_j|\omega_i)$$

4.2.2 ナイーブベイス識別

- 尤度の最尤推定

$$P(x_j|\omega_i) = \frac{n_{ij}}{n_i}$$

n_{ij}: クラス ω_i のデータのうち、
j 次元目の値が x_j の個数

ゼロ頻度問題

- 確率の m 推定

$$P(x_j|\omega_i) = \frac{n_{ij} + mp}{n_i + m}$$

p: 事前に見積もった各特徴値の割合
m: 事前に用意する標本数

- ラプラス推定
  - m: 特徴値の種類数、 p: 等確率　とすると、 $mp=1$

Section2 のまとめ

- 決定木
  - カテゴリデータの学習に適する
  - 学習結果の解釈が可能
- 統計的識別
  - 識別結果を確率付きで出力することができる