情報意味論 (3) 決定木

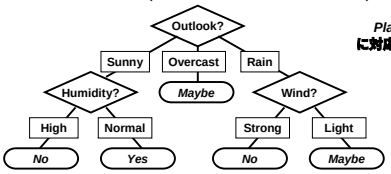
櫻井彰人
慶應義塾大学理工学部

本日の目標

- 決定木とは
- ID3 学習アルゴリズム
- エントロピー、情報量のゲイン
- 過剰適合 (overfitting) , 過学習 (overlearning)

決定木 Decision Trees

- 分類器 Classifiers
  - 事例 (ラベルのついていないもの): 属性 attribute (または特徴 feature) のベクトル
- 内節 Internal Nodes: 属性値のテスト
  - 典型的: 等しいかどうかのテスト (e.g., "Wind = ?")
  - その他 不等式や様々なテストが可能
- 枝 Branches: 属性値 (テストが等式以外のときはテストの結果)
  - 一対一対応 (e.g., "Wind = Strong", "Wind = Light")
- 葉 Leaves: 割当てた分類結果 (分類クラスのラベル Class Labels)



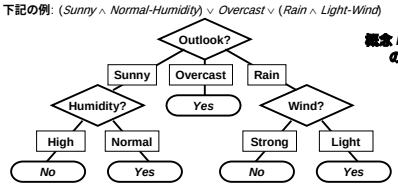
PlayTennis に対応する決定木

決定木

- 決定木が表す意味
  - 各節 (葉以外) は、一つの属性をテスト
  - その各枝は属性値に対応
  - 各葉は一つの類に対応
- 各テストの表現
  - and, or, exclusive or,,,
    - $(A \wedge B) \vee (C \wedge \neg D \wedge E)$
  - *M* of *N*

ブール関数を用いた決定木

- ブール関数を用いた決定木
  - 表現力: 任意のブール関数 (リテラルは属性変数のテスト)
  - なぜ?
    - 復習: 任意の論理式は Disjunctive Normal Form (DNF) でかける
    - 下記の例: (*Sunny* ∧ *Normal-Humidity*) ∨ *Overcast* ∨ (*Rain* ∧ *Light-Wind*)



概念 *PlayTennis* の決定木

- テストに使える他のブール関数
  - ∧, ∨, ⊕ (XOR)
  - $(A \wedge B) \vee (C \wedge \neg D \wedge E)$
  - *m*-of-*n*

他の例: 帝王切開のリスク予測

- 1000 人の女性患者記録から学習
- 負例が帝王切開に対応
  - Prior distribution: [833+, 167-] 0.83+, 0.17-
  - *Fetal-Presentation* = 1: [822+, 167-] 0.88+, 0.12-
    - *Previous-C-Section* = 0: [767+, 81-] 0.90+, 0.10-
      - ***Primiparous* = 0: [399+, 13-]** **0.97+, 0.03-**
      - ***Primiparous* = 1: [368+, 68-]** **0.84+, 0.16-**
        - ***Fetal-Distress* = 0: [334+, 47-]** **0.88+, 0.12-**
          - ***Birth-Weight* < 3349** **0.95+, 0.05-**
          - ***Birth-Weight* ≥ 3347** **0.78+, 0.22-**
        - ***Fetal-Distress* = 1: [34+, 21-]** **0.62+, 0.38-**
    - *Previous-C-Section* = 1: [55+, 35-] 0.61+, 0.39-
  - *Fetal-Presentation* = 2: [3+, 29-] 0.11+, 0.89-
  - *Fetal-Presentation* = 3: [8+, 22-] 0.27+, 0.73-

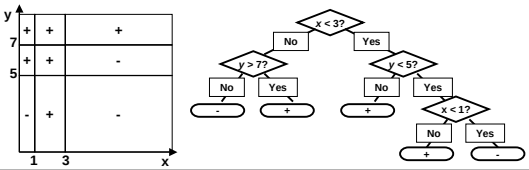
C-section: **帝王切開**, Fetal_Presentation: **胎児の胎位**,
Primiparous: **初産**

どんな時、決定木を用いるか

- 事例が属性 - 属性値ペアで表現される
- 目標関数が離散値をとる（分類問題）
- 選言を含む仮説が必要
- ノイズが入っている可能性がある
- 例（実際に Mitchell が適用した）
  - 機器故障診断、病名診断
  - 与信リスクの分析
    - クレジットカード, ローン
    - 保険
    - 消費者による不正行為
    - 従業員の不正行為

決定木と判別境界

- 事例は，多くの場合，離散属性値で表現される
  - 典型的な型
    - 名義・名辞 nominal ({red, yellow, green})
    - 離散化・量子化 quantized ({low, medium, high})
  - 数値の取り扱い
    - 離散化 discretization, ベクトル量子化 vector quantization
    - 閾値を用いて分割する
- 例: 軸並行な方形によって事例空間を分割する

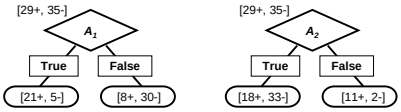


決定木の学習: トップダウン帰納 (*ID3*)

- アルゴリズム *Build-DT* (*事例 Examples*, *属性 Attributes*)

```
IF 全事例が同一ラベルをもつ THEN RETURN (label を付した葉節)
ELSE
    IF 属性値が空集合 THEN RETURN (多数派 label を付した葉節)
    ELSE
        最良属性 A を根節として選ぶ
        FOR A のそれぞれの値 v
            条件 A = v に対応して，根節から枝を作成する
            IF {x ∈ Examples: x.A = v} = Ø THEN RETURN (多数派 label を付した葉節)
            ELSE Build-DT ({x ∈ Examples: x.A = v}, Attributes – {A})
```

- どの属性が最良か?



適用範囲を広げる

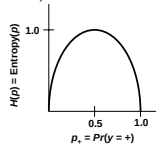
- これまでのアルゴリズムでの仮定
  - 離散 *出力*
    - 実数値出力も可能
    - Regression trees [Breiman *et al*, 1984]
  - 離散 *入力*
    - 量子化の方法あり
    - 内部の等式テストの変わりに *不等式* を使用する (以前の方形の例)
- 規模の拡大
  - 大規模データベース (VLDB) からの知識発見やデータマイニング (KDD) では重要
  - よい話: 多くの *事例* を対象とするよいアルゴリズムあり
  - 悪い話: あまりに多い *属性* を扱うのは難しい
- あると助かる他の耐性
  - ノイズのあるデータ (分類ノイズ classification noise ≡ ラベルの間違い; 属性ノイズ attribute noise ≡ 不正確または低精度のデータ)
  - 欠測値

"最良"の属性の選択

- 目的
  - できるだけ小さい決定木を作る (オッカムの剃刀)
  - 条件: 訓練データのラベルと consistent
- 障害
  - *最小の* consistent な仮説 (i.e., 決定木) を見出すことは NP-hard
  - *Build-DT* (再帰的なアルゴリズム)では
    - 単純な木を作るのに グリーディな発見的な探索 greedy heuristic search
    - 最適性の保証はできない
- そこで:
  - 要請: できるだけ同一のラベルをもつ集合に、事例を分割するような属性
  - その結果: 葉節(ラベルが同一)に近くなる
  - もっともよく使われるヒューリスティック
    - J. R. Quinlan が提案
    - 情報増分 information gain に基づく
    - *ID3* アルゴリズムで使用される

エントロピー: 直感的説明

- 不確実さ・不明瞭さの尺度
  - はかる量
    - 純粋さ purity: 事例集合が、ただ一つのラベルをもつ状態に、どれだけ近いか
    - 不純さ impurity (乱雑さ disorder): ラベルがまったく分からない状態にどれだけ近いか
  - 尺度: エントロピー
    - 比例する対象: 不純さ impurity, 不確実さ uncertainty, 不規則さ irregularity, 驚き surprise
    - 反比例する対象: 純粋さ purity, 確かさ certainty, 規則性 regularity, 冗長さ redundancy
- 例
  - 簡単のため, H = {0, 1}, かつ $Pr(y)$ に従って分布すると仮定
    - 離散的なクラスラベル (2個より多い) を持ちうる
    - 連続確率変数: 微分エントロピー differential entropy (和を積分にしただけ)
  - y に関して最も純粋: どちらか
    - $Pr(y = 0) = 1$, $Pr(y = 1) = 0$
    - $Pr(y = 1) = 1$, $Pr(y = 0) = 0$
  - 純粋さが最も少ない確率分布は?
    - $Pr(y = 0) = 0.5$, $Pr(y = 1) = 0.5$
    - 最大: 不純さ/不確実さ/不規則性/驚き
    - エントロピーの性質: 凹関数 ("上向きに凸")



エントロピー: 情報理論的定義

- **材料**
  - *D*: 事例の集合 {<x_1, $c(x_1)$>, <x_2, $c(x_2)$>, …, <x_m, $c(x_m)$>}
  - $p_+ = Pr(c(x) = +)$, $p_- = Pr(c(x) = -)$
- **定義**
  - *H* は確率密度関数 *p* 上で定義する
  - *D* の事例に対して, その + と – ラベルの頻度を p_+ と p_- で表す
  - *D* の *c* に対するエントロピー は:

$$H(D) \equiv -p_+ \log_b(p_+) - p_- \log_b(p_-)$$

- **単位は**?
  - 対数の底による (*b* = 2 なら bits, *b* = *e* なら nats, 等.)
  - ビットは, 最悪の場合 ($p_+ = 0.5$) の一事例を符号化するのに必要とされる
  - 不確実さが小さければ (e.g., $p_+ = 0.8$), 1ビットより小さくて十分

情報増分: 情報理論的定義

- **属性値に基づく分割**
  - 復習: *D* の分割 partition は, 和集合が *D* となるような排他的部分集合の集合
  - 目標: 属性 A の属性値に基づき分割することにより, 減少する不確実性の尺度
- **定義**
  - 属性 A に関する *D* の情報増分 information gain of *D* relative to attribute A は, A の分割によるエントロピー減少分の期待値:

$$Gain(D, A) \equiv -H(D) - \sum_{v \in values(A)} \left[\frac{|D_v|}{|D|} \bullet H(D_v) \right]$$

  - 但し D_v は $\{x \in D: x.A = v\}$, すなわち, *D* 中の事例で属性 *A* の値が *v* であるものの集合
  - アイデア: *A* の分割; 部分集合 D_v の大きさに従ってエントロピーの大きさを調整
- **どちらの属性が最良**?

例

- **概念** *PlayTennis* **用の訓練事例**

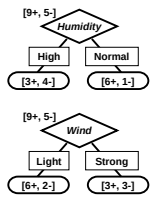
Day	Outlook	Temperature	Humidity	Wind	*PlayTennis*?
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal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

- *ID3* ≡ *Build-DT* **但し** *Gain*(•) **を使用**
- **では** *ID3* **はどう決定木を作るのか**?

ID3 による *PlayTennis* 決定木作成 [1]

- **根節の属性を選ぶ**

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



- **事前 (無条件) 分布**: 9+, 5-
  - *H*(*D*) = -(9/14) lg (9/14) - (5/14) lg (5/14) *bits* = 0.94 bits
  - *H*(*D*, *Humidity* = High) = -(3/7) lg (3/7) - (4/7) lg (4/7) = 0.985 bits
  - *H*(*D*, *Humidity* = Normal) = -(6/7) lg (6/7) - (1/7) lg (1/7) = 0.592 bits
  - *Gain*(*D*, *Humidity*) = 0.94 - (7/14) * 0.985 + (7/14) * 0.592 = 0.151 bits
  - **同様に**, *Gain* (*D*, *Wind*) = 0.94 - (8/14) * 0.811 + (6/14) * 1.0 = 0.048 bits

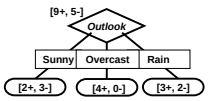
$$Gain(D, A) \equiv -H(D) - \sum_{v \in values(A)} \left[\frac{|D_v|}{|D|} \bullet H(D_v) \right]$$

ID3 による *PlayTennis* 決定木作成 [2]

- **根節の属性を選ぶ**

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

  - *Gain*(*D*, *Humidity*) = 0.151 bits
  - *Gain*(*D*, *Wind*) = 0.048 bits
  - *Gain*(*D*, *Temperature*) = 0.029 bits
  - *Gain*(*D*, *Outlook*) = 0.246 bits
- **次の属性を選ぶ (部分木の根節)**
  - **どの事例もどこかの葉に含まれるか純粋度** = 100% **になるまで続ける**
  - **ところで**, **純粋度** = 100% **とは**?
  - *Gain*(*D*, *A*) < 0 **となりうるか**?



ID3 による *PlayTennis* 決定木作成 [3]

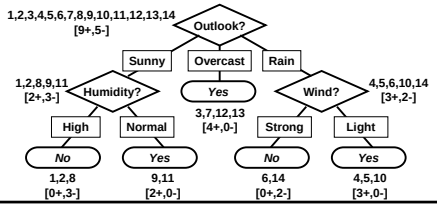
- **次の属性の選択 (部分木の根節)**

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

  - **約束**: log (0/a) = 0
  - $Gain(D_{Sunny}, Humidity)$ = 0.97 - (3/5) * 0 - (2/5) * 0 = 0.97 bits
  - $Gain(D_{Sunny}, Wind)$ = 0.97 - (2/5) * 1 - (3/5) * 0.92 = 0.02 bits
  - $Gain(D_{Sunny}, Temperature)$ = 0.57 bits
- **トップダウンの帰納**
  - **離散値属性については**, $O(n)$ **回分割をすれば終了**
  - **木のレベルそれぞれで**, **訓練データを一回なめる** (**なぜ**?)

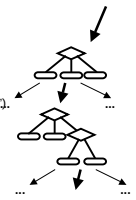
ID3 による *PlayTennis* 決定木作成 [4]

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



ID3 による仮説空間探索

- **探索問題**
  - **探索の対象は *決定木の空間*, 離散関数をすべて表現可能**:
    - Pros: **表現力**; **柔軟性**
    - Cons: **計算量**; **巨大**, **意味の分からない木** (**できたら次回**) **も含む**
  - **目的**: **もっともよい決定木を見出す** (**最小な** consistent **な木**)
  - **障害**: **この木を見出す問題は** NP-hard
  - Tradeoff
    - heuristic **の使用(探索の案内役としての目の子**)
    - **グリーディ** greedy **アルゴリズムの使用**
    - **その一つとして、バックトラックなしの山登り** hill-climbing (gradient "descent").
- **統計的学習**
  - **事例の部分集合** D_v **の統計的な量** p_+, p_- **に基づく決定**
  - ID3 **では**, **全てのデータを使用**
  - **ノイズのあるデータに対してロバスト**



ID3 の帰納バイアス

- **ヒューリスティック**：**探索** :: **帰納バイアス**：**帰納一般化**
  - H **は** X **の冪集合 (全部分集合の集合)**
  - ⇒ **バイアスなし? いや、そうではない**...
    - **短い木を選好 (終了条件から)**
    - **情報増分が高い属性を根節に近いところにおく**という選好
    - *Gain*(•): *ID3* **の帰納バイアスを示すヒューリスティック**
  - *ID3* **の帰納バイアス**
    - **ある仮説への選好がヒューリスティック関数に込められている**
    - **比較せよ**: **仮説空間** H **への制限(命題論理の正規形の議論**: k-CNF, etc.)
- **短い木を好むこと**
  - **データに適合する木の中で最短のものを選ぶ**
  - **オッカムの剃刀バイアス**: **観測を説明する最短の仮説**

術語

- **決定木** Decision Trees (DTs)
  - Boolean **決定木**: **目標概念が2値** (i.e., Boolean-valued)
  - **決定木の作り方**
    - **量子化**: **ベクトル量子化等** (**入力をいくつかの部分空間に分割**)
    - **離散化**: **連続値入力を離散値に** (e.g.., by **ベクトル量子化等**)
- **エントロピーと情報増分**
  - **未知の概念** c **に対するデータ集合** D **の** **エントロピー** $H(D)$
  - **属性** A **でデータ集合を分割することによる情報増分** $Gain(D, A)$
  - **不純さ**, **不確実さ**, **不規則性**, **驚き** versus **純粋さ**, **確実さ**, **規則性**, **冗長性**
- **発見的** Heuristic **探索**
  - **アルゴリズム** *Build-DT*: **グリーディ** greedy **探索** (**バックトラックなしの山登り**)
  - *ID3* : **ヒューリスティックな** *Gain*(•) **を用いた** *Build-DT*
  - **ヒューリスティック**：**探索** :: **帰納バイアス**：**帰納的一般化**

まとめ

- **決定木** Decision Trees (DTs)
  - Boolean ($c(x) \in \{+, -\}$) **でもよいし複数クラスでもよい**
  - **決定木モデルを使ってよいとき悪いとき**
- **アルゴリズム** *Build-DT*: **トップダウン**
  - **最良の属性を選んでは、分割**
  - **再帰的なアルゴリズム**
- **エントロピーと情報増分**
  - **ゴール**: **候補属性** A **に基づき分割したときに得られる、除去される不確実さの尺度**
    - **情報増分の計算 (エントロピーの変化)**
    - **情報増分を用いて木を構成**
  - *ID3* ≡ *Build-DT* **と** *Gain*(•)
- ID3: **仮説空間探索 (決定木空間中)**
- **発見的探索** Heuristic Search **と帰納バイアス**