決定木　その3
属性の選び方

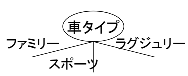
慶應義塾大学理工学部
櫻井彰人

属性の選び方

- 属性タイプによって異なる
  - 名義変数
  - 順序変数
  - 数値変数
- いくつに分割するかによって異なる
  - 2分割
  - 多分割

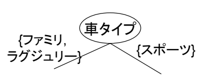
名義変数による分割

- 多分割: 当該変数の変数値の「異なり数」分、分割する.



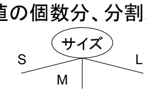
- 2分割: 変数値を2個に分割する.
  最適な分割を求める必要あり.

または

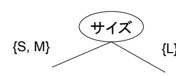


順序変数に基づく分割

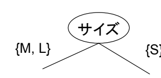
- 多分割: 異なる値の個数分、分割.



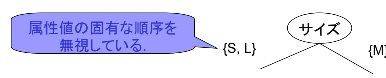
- 2分割: 2つの部分集合に分割.
  最適分割を見つける必要あり.



OR



- この分割は?



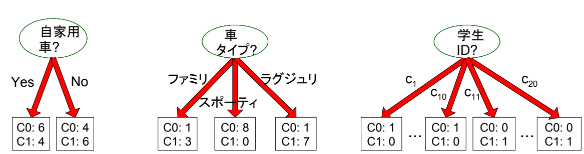
数値変数に基づく分割

- 身長、体重、血圧、コレステロール値、、、、
  - 1単位ずつ分けると分けすぎ。
  - 離散化: いくつかの境(閾値ともいう)を設けて、いくつかに分ける。
- いくつかの方法がある
  - 離散化 して順序属性として扱う
    - 静的 – 最初に一回だけ離散化
    - 動的 – 等幅区間、等頻度区間(パーセンタイル)、クラスタリング
  - 2値判別: (A < v) または (A ≥ v)
    - すべての可能な分割を考え、ベストなものを見出す
    - 計算が一層必要となることも

やはり、分割の問題(分割の良さを比較する問題)だ！

最良な分割はどうやって見つける？

分割前: C0 (クラス0)に 10 データ,
C1 (クラス1)に 10 データ



どの条件が最適か?

“最良”の属性の選択

- 最良の分類木とは何だろうか？
  - 「正しい木である」こと
  - それはそうだが、我々は、出来上がった木が「正しい」かどうか分らない
    - 正解を知っている「先生」がいないのだ。
  - そこで、こう考えた。「これから来るデータを正しく分類できる」ならどうだろうか？
    - 「分類木」が間違っていても、答え（新しいデータの分類結果）が正しければよい。
  - しかし、「これから来るデータ」は、入手できないよね。
  - そこで、こう考えた。与えられたデータを、訓練データとテストデータに分けよう。
    - 訓練データで分類木を作り、テストデータでテストすればよい。
  - これはすごくいい考えだ。けれども、これでは、訓練データを使って一個分類木を作って、テストデータで調べるだけであって（点数が分る。評価するという）、それがよいものかどうか分らない。
  - そこで、こう考えた。最もよい分類木を作るためには、色々な訓練データを作ってみればよい。ランダムに訓練データを作って、テストデータでテストし、テスト結果が最良のものを選べばよい。これはよい考えだ！

“最良”の属性の選択

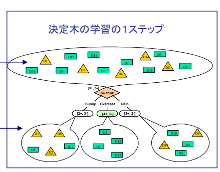
- 最良の分類木とは何だろうか？
  - いや、そうでもない。それなら、最初から、全データを正しく分類する分類木を作れば、よいでしょう。どんなテストデータに対しても正しく分類する。
  - これは困った。だめか。やっぱり、テストデータを使って最も良いものを選んではだめなんだ。
  - そこで、こう考えた。最もよいものを選ぶのはやめよう。身の回りで一番よいもので我慢しよう。
  - (1) その方法の一つとして、訓練データは固定して、そのもとで、いくつか決定木を作り、テストデータで評価して、一番よいものを使おう。 ← 例：pruning法
  - (2) もう一つは、昔の哲学者の言葉に従うことだ。オッカムが言うには「データを説明する同じような仮説が複数あるときには、一番短い仮説をとれ」
  - 実は方法(1)の場合も、「短い」決定木が選ばれることが知られている。

“最良”の属性の選択

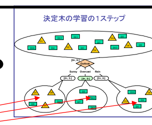
- 最良の分類木とは何だろうか？
  - 間違いが同じ程度なら、一番小さい分類木がよいと、いえる
    - 間違いが同程度でないときどうするかは、少々難しい問題ゆえ、ここでは省略。
- そうであれば、「最良の属性」とは、その属性を選んだら分類木が小さくなりそうな属性であろう。
- となると、「最良の属性」とは、その属性を選んだら、（訓練データが分割されるわけだが）分割後の訓練データが、各分割内で、そろっているような属性だ。

こちらが、ごちゃごちゃでも

こちらが、すっきりするような属性を選ぶべきだ



最良な属性、最良な分割は？



つまり

- 「決定！」に近くなる方がよい、すなわち:
  - 新ノード内のクラス分布が　となる 分割がベター
  - どこかのクラスが圧倒的な多数となる（これが　）ということは、それだ！といっても間違いが少ないから
- そのためには、（ノードの）　さの物差しが必要:

C0: 5
C1: 5

非-　純度が低い

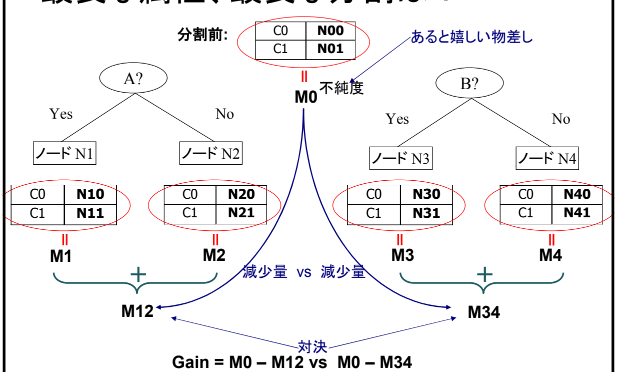
不純度が高い

C0: 9
C1: 1

　純度が高い

不純度が低い

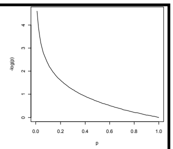
最良な属性、最良な分割は？



不純度のものさし

- エントロピー
- ジニ・インデックス Gini Index
- 誤分類率

エントロピー



- ____とも呼ばれる。式で書くと

$$H(p_1,\ldots,p_m) = -p_1\log_2 p_1 - \ldots - p_m\log_2 p_m$$
$$= p_1(-\log_2 p_1) + \ldots + p_m(-\log_2 p_m)$$

- （比較のために）サイコロの出る目の平均

$$+\,p_1 * 1 + p_2 * 2 + \ldots + p_6 * 6$$

- つまり、平均情報量が情報量の平均だとすると $-\log_2 p_i$ が「____」ということになる

負の符号「－」がついているのは、p<1 故 log g <0 となるが、負の数はいろいろと不便なため、符号反転しているから。

情報量

ある事象の情報量は、
その事象が起こったということを（他の皆が知らないときに）
知ることの価値

「事象」として「コインの表が出ること」（確率1/2））としよう。
「表が出たこと」を知る価値を a としよう。
「コイン1が表」「コイン2が表」という2つの情報を知る価値は a + a = 2a だろう（一つずつ聞く場合を考えればよい）。
「コイン1が表」「コイン2が表」の二つの事象が起こる確率は ½ * ½ =1/4 。

「事象」として「サイコロの1が出ること」（確率1/6））としよう。
「1が出たこと」を知る価値を b としよう。
「サイコロ1が1」「サイコロ2が1」という2つの情報を知る価値は b + b = 2b だろう（一つずつ聞く場合を考えればよい）。
「コイン1が表」「コイン2が表」の二つの事象が起こる確率は 1/6 * 1/6 =1/36 。
つまり、事象が起こる確率が2乗になると、価値は2倍になる

情報量を表す関数

事象が起こる確率が2乗になると、価値は2倍になる

事象が起こる確率 p が p^2 になると、価値 v は 2v になる

事象が起こる確率 p が p^2 になると、価値 v(p) は $v(p^2) = 2v(p)$ になる

上記のような関数は log しかないことが示せる（底は決まらない。何でもよい）

そこで、底を2とし価値が正になるように符合反転すると（底を1/2 にしたのと同じ）、生起確率 p の事象が生起したことを知るという情報の価値は、-log p とすればよいことが分る。

$$\text{情報量}(p) = -\log_2 p$$

不公平かもしれないコイン

- 表が出る確率 p, 裏が出る確率が 1-p であるコインのコイン投げを考える。
- このコインを1回投げたときに出た「表・裏」を知る情報の価値はどのくらいであろうか？
- 「表が出る」という情報の価値は、____「裏が出る」という情報の価値は、____である。
- 表が出る確率は p, 裏が出る確率は 1-p であるので、この確率に基づく（情報価値の）平均値を考えよう

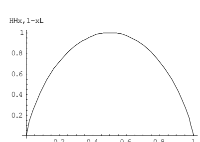
$$H(p,1-p) = p(-\log_2 p) + (1-p)(-\log_2(1-p))$$
$$= -p\log_2 p - (1-p)\log_2(1-p)$$

不公平かもしれないサイコロ

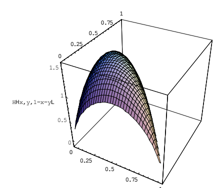
- 「目iが出る」確率 p_i であるサイコロを考える。
- このサイコロを1回投げたときに出た目を知る情報の価値はどのくらいであろうか？
- 「目iが出る」という情報の価値は、-log p_iである。
- この確率に基づく（情報価値の）平均値を考えよう

$$H(p_1,p_2,\ldots,p_6)$$
$$= p_1(-\log_2 p_1) + p_2(-\log_2 p_2) + \cdots + p_6(-\log_2 p_6)$$
$$= -p_1\log_2 p_1 - p_2\log_2 p_2 - \cdots - p_6\log_2 p_6$$

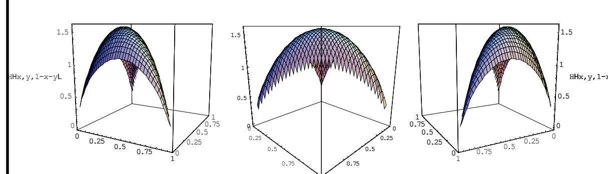
グラフ



H(x,1-x)



H(x, y, 1-x-y)



符号理論から

- 確率変数 X の独立サンプルを観測しているとする
- X は4個の値をとる

P(X=A) = 1/4	P(X=B) = 1/4	P(X=C) = 1/4	P(X=D) = 1/4

- 従って、例えば、: BAACBADCDADDDA…
- シリアルリンク（1本の信号線）で2進符号を送る場合、個々の観測を2ビットに符号化できる (e.g. A=00, B=01, C=10, D = 11)

0100001001001110110011111100…

より少ないビット数で

- もし、誰かが、実は、等確率ではないのだヨと教えてくれたら

P(X=A) = 1/2	P(X=B) = 1/4	P(X=C) = 1/8	P(X=D) = 1/8

- 可能なのは…

… 平均では、1観測あたり 1.75ビットとなるような符号の作り方.

A	0
B	10
C	110
D	111

最短な符号の場合

- 確率変数 X は *m* 個の値をとるとする…

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

- X の独立試行から得られる値（の観測）の列を送信する場合、1観測（記号）あたりのビット数（の期待値）を最小化する場合、その値は何か？それは、実は

平均符合長

符合の長さ。小数であるが、ご容赦

$$H(p_1,\dots,p_m) = -p_1\log_2 p_1 - \dots - p_m \log_2 p_m$$
$$= p_1(-\log_2 p_1) + \dots + p_m(-\log_2 p_m)$$

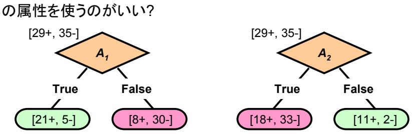
- H(X) : X の
- 下限であること（これ以上短くならないこと）は比較的やさしい。難しいのは、いくらでもこれに近づけることができるということの証明。これを見出したのは Shannon である.

情報量増分

- 定義
  - 属性 A に関する*D* の情報量増分 は, A を用いた分割によるエントロピー減少分の期待値:

$$Gain(D,A) \equiv H(D) - \sum_{v \in values(A)} \left[\frac{|D_v|}{|D|}\cdot H(D_v)\right] = \frac{1}{|D|}\left(|D|\cdot H(D) - \sum_{v \in values(A)} |D_v|\cdot H(D_v)\right)$$

  - 但し D_v は $\{x \in D \mid x.(A) = v\}$, 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
  - 補足: *A* による分割によって生じる部分集合 D_v の大きさに従ってエントロピーの大きさを調整
    - エントロピー値は、「集合の要素一個あたり」の情報量となっているため
- どちらの属性を使うのがいい?



GINI に基づく分割基準

- これまで説明してきた分割基準はエントロピーであった:

$$Entropy(t) = -\sum_j p(j \mid t)\log p(j \mid t)$$

(注: *p(j | t)* はノード t におけるクラス j データの相対頻度).

- 別法に GINI インデックスを用いるものがある:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

- 両者とも:
  - 最大値 ($\log n_c$ または $1 - 1/n_c$) が得られるのは, 当該データがどのクラスにも等分に分配されているときである. 「等分である」ということは何の面白さもない. しかし、皆が勝負の行方が分らないとき、自分だけどちらが勝ったかを知らせる情報は、非常に価値がある。
  - 最小値 (0.0) が得られるのは, すべてのデータが同一のクラスに属するとき. 初めからどちらが勝つかは皆が知っているのだから、勝負の結果を知らせる情報は平均的には、全く価値がない。弱い方が勝った場合にはその情報は価値がものすごくあるが、殆ど起こらないので、平均をとってしまうと0に近くなってしまう