情報と意味（4補足）
適切な一般さ

櫻井彰人
慶應義塾大学

今日の話題

- オッカムの剃刀（概念版）
  - 最も単純な規則を選べ
- Bayes の方法
  - 事後確率最大の仮説を選べ
  - 事前確率と事後確率の関係
- universal prior とデータ圧縮による予測
- universal probability による MDL
  - MDL はオッカムの剃刀の数学的定式化

Induction

- OED (Oxford English Dictionary) によれば
  - the process of inferring a general law or printciple from the observations of particular instances
  - これは、inductive inference のこととする
  - inductive reasoning は: the process of reassigning a probability (or credibility) to a law or proposition from the observation of particular events

エピクロスの多説明原理

- ギリシャの哲学者 Epicurus
  - If more than one theory is consistent with the observations, keep all theories (Principle of Multiple Explanations).
- その一つの理由: 一つを他から選び出す理由がない

Occam の剃刀

- 人口に膾炙しているのは
  - Entities should not be multiplied beyond necessity.
- Bertrand Russell によれば
  - It is vain to do with more what can be done with fewer.
- 最も普通の解釈
  - Among the theories that are consistent with the observed phenomena, one should select the simplest theory.

Isaac Newton の言葉

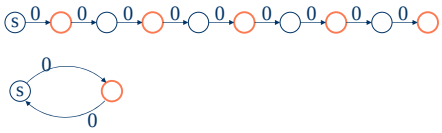
- We are to admit no more causes of natural things than such as are both true and sufficient to explain the appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.

有限状態オートマトンの例

- 有限状態オートマトンA
  - Aは初期状態と受理状態とを含む有限個の状態を持つ。入力文字を読み、現在状態と該文字のみにより次の状態を決める。
  - Aは入力文字列を受理と拒絶に分ける。例:
    - 受理: 0, 000, 00000, 0000000;
    - 拒絶: ε, 00, 0000, 000000.
  - これに対応する2つのオートマトンを考える

2つのオートマトン

- 次のどちらのオートマトンがよいだろう



あまりに単純に用いると、、

Once upon a time, there was a little girl named Emma. Emma had never eaten a banana, nor had she ever been on a train. One day she had to journey from New York to Pittsburgh by train. To relieve Emma's anxiety, her mother gave her a large bag of bananas. At Emma's first bite, the train plunged into a tunnel. At the second bite, the train broke into daylight again. At the third bite, Lo! into a tunnel; the fourth bite, La! into daylight again. And so on all the way to Pittsburgh. Emma, being a bright little girl, told her grandpa at the station, "Every odd bite of a banana makes you blind; every even bite puts things right again."
[N.R. Hanson, Perception and Discovery, 1969, Freeman and Cooper, p.359]
M.Li and P.Vitanyi, An Introduction to Kolmogorov Complexity and Its Applications, 1997, Springer-Verlag, p.318.

もう一つの例

- 例えば次のような規則としてみよう
  - 観測データに最もよく合致する仮説をとれ
  - もし複数あるなら最も単純な仮説をとれ
- 次の例ではどうなる?
  - 白球と黒球がたくさん入った壷(中は見えない)がある。ランダムに1個取り出し色を記録し戻す、ことを n 回繰返し白球の割合を推定する。m 個あった時、m/n と推定する。
  - 正解1/3が得られる確率は0か、0に漸近する。

単純さとは

- 簡単には定義できそうもない
  - 1/4 は 1/10 より単純か?
  - 1/3 は 2/3 より単純か?
    - 1/3が白球なら2/3は赤球
  - $x^{100}+1$ は $13x^{17}+5x^3+7x+11$ より単純か?

Bayes の考え

- 確率的な観点から:
  - Bayes の法則: 仮説 H が真である確率は、観察者の初期の信念(belief)(prior probability: 事前確率)と H が与えられた時の観測データ D の条件付確率との積である

Bayes の方法の役割

- 実際的な学習アルゴリズムの基礎
  - Naïve Bayes 学習
  - Bayesian belief network
  - 事前知識を組み込む
- 概念的な枠組みの提供
  - 他の学習アルゴリズムのよき基準
  - オッカムの剃刀の洞察

Bayes の定理

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

- $P(H)$ = 仮説 H の事前確率
  - a priori, initial, or prior.
- $P(D)$ = 訓練例 D の事前確率
- $P(H|D)$ = D が与えられた時のH の確率
  - 事後確率: final, inferred, or posterior.
- $P(D|H)$ = H が与えられた時のDの確率

Bayes の定理（複数の仮説）

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{\sum_i P(D \mid H_i)P(H_i)}$$

- 仮説は網羅的で排他的であるとする。
  - $\sum_i P(H_i)=1$
- 仮説 H に対して $P(D|H)$ が計算できるとする。

仮説の選択（事後確率最大）

- 一般には、観測データのもと、最も確からしい仮説を選びたい
  - Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)P(h) \end{aligned}$$

仮説の選択（最尤仮説）

- もし、$P(h_i)=P(h_j)$ であれば、簡単化できて、最尤仮説(maximum likelihood hypothesis):

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

確率の基本公式

- 積事象: $P(A \wedge B)=P(B|A)P(A)=P(A|B)P(B)$.
- 和事象: $P(A \vee B)=P(A)+P(B)-P(A \wedge B)$.
- 事象 $A_1, A_2, \ldots, A_2$ が排他的かつ網羅的であれば、$\sum_{i=1}^{n} P(A_i)=1$であって $P(B)=\sum_{i=1}^{n} P(B|A_i)P(A_i)$

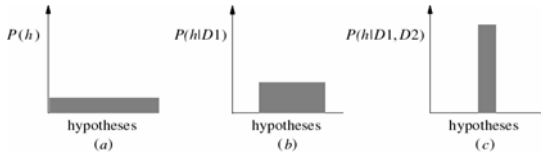
概念学習(FindS)との関係

- 通常の概念学習の課題を考える
  - 事例集合 X, 仮説空間 H, 訓練事例 D
  - FindS は $VS_{H,D}$ から最も特殊な仮説を出力
- Bayes 規則が選ぶ MAP仮説は？
- FindS は MAP 仮説を出力するか？

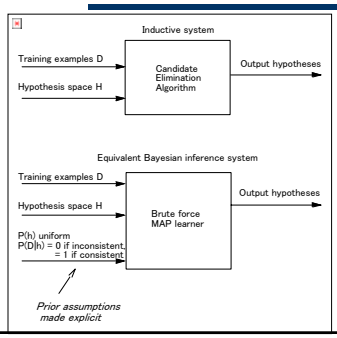
概念学習との関係

- 事例集合 $\{ x_1,\ldots,x_m \}$ を固定
- 訓練集合 D は $\{c(x_1),\ldots,c(x_m)\}$
- 次のような h を選ぶ
  - $P(D|h)=1$, h がDと整合していれば
  - $P(D|h)=0$, そうでなければ
- P(h)=1/|H|, すなわち一様分布とする
  $P(h|D)=1/|VS_{H,D}|$, h がDと整合していれば
  $P(h|D)= 0$, そうでなければ

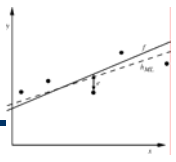
事後確率の変化



概念学習と等価なMAP学習



実数値関数の学習



- 実数値関数の目標 f を考える
- 訓練例 $<x_i, d_i>$. 但し、d_i はノイズを含み、
  - $d_i=f(x_i)+e_i$
  - e_i は平均0のガウス分布を持つ確率変数
- この時、h_{ML} は次の誤差平方を最小化する

$$h_{ML} = \arg\min_{h\in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

$$h_{ML} = \arg\max_{h\in H} p(D\,|\,h)$$

$$= \arg\max_{h\in H} \prod_{i=1}^{m} p(d_i\,|\,h)$$

$$= \arg\max_{h\in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}$$

自然対数をとったものを最大化する

$$h_{ML} = \arg\max_{h\in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

$$= \arg\max_{h\in H} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

$$= \arg\max_{h\in H} \sum_{i=1}^{m} -(d_i - h(x_i))^2$$

$$= \arg\min_{h\in H} (d_i - h(x_i))^2$$

未知事例の最もありうる分類

- これまで、事例 D のもとでの最もありうる仮説を求めてきた（例：h_{MAP}）。
- 未知事例の最もありうる（確率が高い）分類はどうなるのであろうか？
  - $h_{MAP}(x)$ は最もありうる分類ではない！
    - 次の例で、x のもっともありうる類別は？
    - 3仮説: $P(h_1|D)=0.4$, $P(h_2|D)=0.3$, $P(h_3|D)=0.3$
    - 新事例: $h_1(x)=+$, $h_2(x)=-$, $h_3(x)=-$

Bayes 最適な分類

- Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D)$$

- 例 $P(h_1|D)=0.4,\ P(-|h_1)=0,\ P(+|h_1)=1$
  $P(h_2|D)=0.3,\ P(-|h_2)=1,\ P(+|h_2)=0$
  $P(h_3|D)=0.3,\ P(-|h_3)=1,\ P(+|h_3)=0$

このとき $\sum_{hi \in H} P(+|h_i)\ P(h_i|D) = 0.4$
$\sum_{hi \in H} P(-|h_i)\ P(h_i|D) = 0.6$

そして $\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) = -$

事前確率分布

- 現実世界では、事前確率分布は、未知か、計算不能か、存在しないと思われる
  - 例えば、文書における単語の生起頻度の事前分布はあるのだろうか？　年齢、社会的背景、人口分布で大きく異なりうる
- そこで、（真の事前確率分布でなくとも）結果がほぼ同じになるような事前分布はないだろうか？

もとをたどると

- D.Hume (1716-1776):
  - 真の帰納は不可能である。なぜなら、我々は、既知のデータと既知の方法によってのみ結論に達しうるからである。
  - そこで多くの哲学者は確率的方法を模索した

Solomonoff

- Solomonoff [1964, 1978]:
  - 全ての帰納的推論は、2進数列を外挿する形に帰着できる
    - 外挿するには、既知の数列に基づく仮説が必要
    - 外挿するときに使われる方法には単純性（simplicity）と中立性（indifference）とがある
    - 単純性：最も単純な仮説が最も信頼できる
    - 中立性：そうでない証拠がないときには、複数の仮説はどれも同様に信頼できる

2進列の外挿と帰納推論

- B={0,1}（有限集合ならなんでもよいが）
  - B^*=0,1の有限列全ての集合
  - B^∞=0,1の片無限列全ての集合
- 現象や概念はB^∞上の確率分布 μ
  - （形式的な）知識はB^*の要素
  - 実験の方法もB^*の要素
  - 結果もB^*の要素
  - こういったものが無限に確率的に生起する

確率測度μ

- $\mu(x)$: 文字列 x を先頭とする全文字列の測度
  - $\mu(y|x)= \mu(xy) / \mu(x)$
- semimeasure:
  - μ(空文字列) ≤ 1
  - $\mu(x) \geq \sum_{a \in \mathrm{B}} \mu(xa)$
- **M**: a universal enumerable continuous semimeasure
  - 全enumerable continuous semimeasures の集合Mの要素μ_0 であって、$\forall \mu \in M\ \exists c>0\ \forall x \in \mathrm{B}^*\ \mu_0(x) > c\mu(x)$

Mの性質

- $S_n=\sum_{l(x)=n-1} \mu(x)(\mathbf{M}(0|x) - \mu(0|x))^2$ とおく
  - μ は recursive semimeasure
- $\sum_n S_n \leq (1/2)\, K(\mu) \ln 2$
  - K は Kolomogorov complexity. μ（の符号化）を出力する self-delimiting program の最短長
- $\mathbf{M}(y|x)/\mu(y|x) \to 1$ (μ確率1で)
  - y はある固定した長さ、x の長さが増大するとき
  - μ は positive recursive measure

Mの性質

- $\lim_{l(x)\to\infty} \log \mu(y|x) - \log \mathbf{M}(y|x) = 0$ であるので、$-\log \mu(y|x)$ を最小化することと $-\log \mathbf{M}(y|x)$ を最小化することとは同一
  - 条件がついてはいる

要は

- どんな（普通の）prior μ に対しても、
  - $\log \mathbf{M}(x) - \log \mu(x)$ はある定数以下
    - これは定義。こういうMが存在する（これは定理）ということが大切
  - 0と1の有限列 x の後0 がくる確率の予測の **M** と μ との差（2乗の期待値）は、x の長さ n に対し $1/n$ より速く減少する。

Mによる予測

- **M**による予測は、全てのsemimeasure による予測の重みつき線型和
- 最短長プログラムによる予測はdominant か？
  - 実は No.
- しかし、殆どの場合は dominant
- **M**の性質から、殆どの場合、MDLによる予測が最良であることが分かる

データ圧縮による外挿

- ある μ による外挿を考える
- $-\log \mu(y|x) = -\log \mu(xy) + \log \mu(x)$ を最小にする y を求めたい (但し、$l(y)=n$)。
- 一方、$\lim_{l(x)\to\infty} -\log \mu(y|x) = Km(xy) - Km(x) + O(1) < \infty$
  - $Km(x)$ は x を先頭とする無限文字列を出力する monotone machine の最短記述長
  - monotone machine: work tape 以外に one-way read-only の input tape と one-way write-only のoutput tape を持つTM
  - $Km(x)$ と $-\log \mathbf{M}(x)$ とはμ-random 列に対しては定数の違いしかない

データ圧縮による外挿(2)

- すなわち: $\mu(y|x)$ が最大となるのは
  - $\mathbf{M}(y|x)$ が最大であって、その時に限る
  - y が x に関して最も圧縮された時であり、かつその時に限る(x からの固定長 y の外挿が)。
  - なお、xy が μ-random 列の prefix でなければならない、といった条件がある

例

- $\mathbf{M}(y|x)$ は x から y が簡単に説明できるほど大きな値となる。
  - $\mathbf{M}(1|0^n)=\Theta(1/n)$. すなわち、たくさん 0 が続いた後では、1 がでると説明するよりは 0 がでると説明するほうが簡単
  - Bernoulli過程$(p, 1–p)$ と仮定し、事前確率を p の一様分布として推定すると $p=(n+1)/(n+2)$. すなわち上記の値は $1/(n+2)$ となる

雑談

- 太陽が明朝昇らない確率は、10,000年来上りつづけた後であれば、約 1/3,650,000である。Laplace, A Philosophical Essay on Probabilities
- 「太陽に関する情報」が尽くされていれば、正しい(実世界ではこの前提が成立しないことが多い)

仮説の同定

- やはりBayesに基づく
  $\Pr(H|D) = \Pr(D|H)\ P(H) / \Pr(D)$
  - H: 仮説、P: 仮説の事前確率、D: 観測データ
  - DとPを固定して $\Pr(H|D)$ を最大化するHを求める
  - これは次のものの最小化と同じ
    $-\log \Pr(H|D) = -\log \Pr(D|H) -\log P(H) + \log \Pr(D)$,
    $-\log \Pr(D|H) -\log P(H)$

Shannon-Fano符号で解釈

- $-\log \Pr(D|H) -\log P(H)$ の各項は、
  - $-\log P(H)$: Shannon-Fano符号による、仮説 H の(期待)符号長(の下限)
  - $-\log \Pr(D|H)$: Shannon-Fano符号による、仮説 H の元での観測データ D の(期待)符号長(の下限)
- どうして、Shannon-Fano符号でなければいけないの? (Hが確率的でないときは?)

確率測度P

- $P(x)$: 自然数で定義され実数値をとる関数
- semimeasure:
  - $\sum_{x\in \mathsf{N}} P(x) \le 1$
- **m**: a universal enumerable discrete semimeasure
  - 全enumerable discrete semimeasures の集合Mの要素P_0であって、$\forall P\in M\ \exists c>0\ \forall x\in \mathsf{N}\ P_0(x) > cP(x)$

mの性質

- $\log \mathbf{m}(H) = -K(H) \pm O(1)$
  $\log \mathbf{m}(D|H) = -K(D|H) \pm O(1)$
  - K は Kolomogorov complexity. H（の符号化）を出力する self-delimiting program の最短長
- $K(D|H) + K(H) - \alpha(P,H)$
  $\leq -\log \Pr(D|H) - \log P(H) \leq K(D|H) + K(H)$
  - $\alpha(P,H) = K(\Pr(\cdot|H)) + K(P) \leq K(H) + O(1)$
  - これを FI (fundamental inequality) と呼ぶことにする

FIが成立するには

- H が P-randomである
  - もし P-random でない仮説が真の仮説であったとすると、$K()$ を用いては得られない
- D が $\Pr(\cdot|H)$-random である
  - もし ある仮説のもと D が $\Pr(\cdot|H)$-random でないとすると、仮説は正しくない可能性がある
- 全ての仮説が P-random になるよう P を定め、D が $\Pr(\cdot|H)$-random となるような H を選べばよい（MDLにもBayesにも反しない）

要は

- P として **m** を用いれば、MDL（計算機の世界）とBayes（確率の世界）とは等価になる
- $P(x) = \mathbf{m}(x) = 2^{-K(x)}$ とすればよい
  - K は Kolomogorov complexity. x（の符号化）を出力する self-delimiting program の最短長
  - すなわち、ある符号化方法（プログラム方法）による x の符号の最短の長さ
  - 確率的事象なら、通信路符号化と類似になる

Bayes と MDL

- Bayes:
  - $\Pr(D|H)\,P(H)$ を最大化するHを求めよ
- MDL原理（規準）:
  - $K(D|H) + K(H)$を最小とする仮説を求めよ
- 注:
  - P が分かっているときは Bayes
  - 確率的ではないとき、P が不明の時は MDL

Bayes と MDL

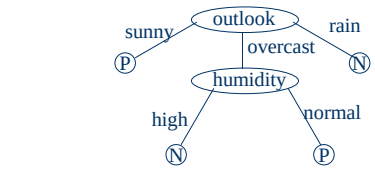
- Bayes:
  - 観察者の初期の信念（belief）（prior probability: 事前確率）と H が与えられた時の観測データ D の条件付確率との積を最大化するHを求めよ
- MDL原理（規準）:
  - 次の和を最小とする仮説を求めよ
    - 理論の記述長（ビット数）
    - 理論を用いて観測データを記述した時の、データの記述長（ビット数）

例

No.	Outlook	Temp	Humid	Windy	Class
1	overcast	hot	high	not	N
2	overcast	hot	high	very	N
3	overcast	hot	high	med	N
4	sunny	hot	high	not	P
5	sunny	hot	high	med	P
6	rain	mild	high	not	N
7	rain	mild	high	not	N
8	rain	hot	normal	not	P
9	rain	cool	normal	med	N
10	rain	hot	normal	very	N
11	sunny	cool	normal	very	P
12	sunny	cool	normal	med	P

No.	Outlook	Temp	Humid	Windy	Class
13	overcast	mild	high	not	N
14	overcast	mild	high	med	N
15	overcast	cool	normal	not	P
16	overcast	cool	normal	med	P
17	rain	mild	normal	not	N
18	rain	mild	normal	med	N
19	overcast	mild	normal	med	P
20	overcast	mild	normal	very	P
21	sunny	mild	high	very	P
22	sunny	mild	high	med	P
23	sunny	hot	normal	not	P
24	rain	mild	high	very	N

工夫なし: $24(3 \log_2 3 + 1 + 1) = 162.12$ bits
予め全ての属性値の組合せの数え上げを送っておけば、54 bits ですむ



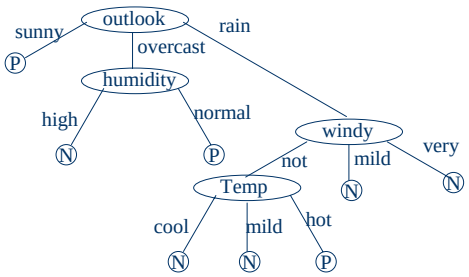
内部ノード　　属性値の個数は符号化せず

1 Outlook 0 P 1 Humidity 0 N 0 P 0 N

1　2　　1 1　1　　log3　　1 1　1 1 1 1 → 13.585bits

例外：1/54ケースなので、log54=5.75bits が簡単

合計：19.335 bits



1 Outlook 0 P 1 Humidity 0 N 0 P 1 Windy
1 Temp 0 N 0 N 0 P 0 N 0 N

1 · 20 + 2+log3 + log2+1=25.585 bits

まとめ

- オッカムの剃刀とBayes 最適化
  - 歴史的には二つの道
  - それぞれに欠点：
    - 「記述手段」に根ざす不定性
    - 「事前確率分布」
  - universal (semi-)measure による統一
    - TM による記述、$2^{-\text{記述長}} \approx$確率
  - 記述されるべきデータを含めて記述長最短の仮説を選択 ≈ 事後確率最大の仮説を選択

参考文献

- Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997.
- Ming Li and Paul Vitányi, An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag, 1997.

まとめ：MDL規準（principle）

- オッカムの剃刀：短い仮説を選べ
- MDL: 次を最小化する仮説 h を選べ

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D \mid h)$$

但し、$L_C(x)$ は符号化Cによるxの記述長

- 例：H=決定木、D=訓練例
  - $L_{C_1}(\mathrm{h})$: 木 h を表現するビット数
  - $L_{C_2}(D|h)$: h のもと D を表現するビット数
    h で表現できなかった例外のみを表現する
  - h_{MDL} は木の大きさと誤差とを勘案する

まとめ：MAPと比較すると

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D \mid h) P(h) \\ &= \arg\max_{h \in H} \log_2 P(D \mid h) + \log_2 P(h) \\ &= \arg\min_{h \in H} -\log_2 P(D \mid h) - \log_2 P(h) \qquad (1) \end{aligned}$$

- 情報理論によると
  - 確率 p の事象の最良な符号長は $-\log_2 p$ ビット
- (1)式は次のように解釈できる
  - $-\log_2 P(h)$: 最良符号のもとでの h の記述長
  - $-\log_2 P(D|h)$: 最良符号のもと、h が与えられたもとでの D の記述長

オッカムの剃刀（Ockham’s razor）

- OEDより
  - ***Occam’s****(also* ***Ockham’s)*** ***razor***, the leading principle of the nominalism of William of Occam (see Occamism), that for purposes of explanation things not known to exist should not, unless it is absolutely necessary, be postulated as existing; usually called the Law of Parcimony
- Modern approach では
  - The most likely hypothesis is the simplest one that is consistent with all observations.

オッカムの剃刀

- 平凡社世界大百科より（稲垣 良典）
  - 〈必要なしに実在を多数化してはならぬ〉という形で知られる〈思考節約の原理〉，いわゆる〈オッカムの剃刀(かみそり)〉は，ほんらい観察された事実，論理的自明性，神的啓示など〈十分な根拠〉なしにはいかなる命題も主張してはならないことを規定している。

William of Ockham

- Occam was a pupil of Duns Scotus, but rejected and opposed the Realism of his master, forming a new speculative sect who revived the tenets of Nominalism. He maintained that general ideas have no objective reality out of the mind, but are merely a product of abstraction. His teachings prepared the way for the overthrow of scholasticism.

実念論

- 平凡社世界大百科より（茅野 良男）
  - 実在論の最初の中世的形態。概念実在論ともいい，普遍的概念の実在性を主張する。類，種，種差などの〈普遍 universalia〉は，個物から独立し，〈個物に先立って ante rem〉実在すると説く，広義のプラトン主義的実念論が典型。個物のみ実在すると説く唯名論がこれに対立する。のち，普遍は〈個物において in re〉のみ実在すると説く，ゆるやかなアリストテレス的実念論が登場した。前者はエリウゲナ，アンセルムス，シャンポーのギヨームらが，後者はラ・ポレのジルベール，ソールズベリーのヨハネス，トマス・アクイナスらが代表。

唯名論

- 平凡社世界大百科より（茅野 良男）
  - 名目論ともいい，中世の実念論に対立する立場。個物のみ実在し，類・種などの普遍は実在せず，ただ人間の精神の中で〈個物の後にpost rem〉生じると説く。普遍は等しい個物に対する単なる〈声 vox〉ないし〈名 nomen〉であるか，個物に面して精神が懐胎し総括する〈概念 conceptus〉または概念の概念として精神により〈総括された記号 terminus conceptus〉とされる。ロスケリヌス，アベラール，オッカムが代表者。アベラールは普遍は名(概念)または〈言表 sermo〉であるとし，普遍の〈概念説 conceptualism〉と〈言表説 sermonism〉の発端となり，オッカムは普遍の〈記号説 terminism〉の先駆となる。唯名論は個体主義・感覚論への傾向をもち，近世の経験論を準備した。

予測精度の推定

- 検査集合（test set）を準備する
  - 学習後の性能を計測するための、例の集合
  - 通常は学習には使用しない
- しかし、例を集めるのが大変
- そこで、

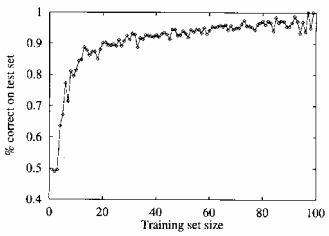
予測精度の推定（承前）

- 次の方法をとる
  1. 例を集める
  2. 訓練集合の大きさを決める（全例数より小）
  3. 全例を訓練集合と検査集合に分ける
     - ランダムにかつ排反集合となるように分ける
  4. 訓練集合を用いて学習．得られた仮説を H .
  5. 検査集合を用いて H の正答率を求める
  6. 上記 ③ を繰り返す
  7. 上記 ⑤ を繰り返す

学習曲線

- 訓練集合の大きさ vs 平均予測精度（推定）のグラフ

A learning curve for the decision tree algorithm on 100 randomly generated examples in the restaurant domain. The graph summarizes 20 trials. (Artificial Intelligence: A Modern Approach より Fig. 18.9)



情報理論の利用

- 「よい属性」を選ぶ基準の理論的基盤を作る
- よい属性：
  - その属性を用いて分類した時、他の属性での分類より、得られる情報利得の期待値が大
  - 答え（「待つ」と言うか言わないか）の予測精度が、その属性値を知ることにより向上するが、その値の大きいもの

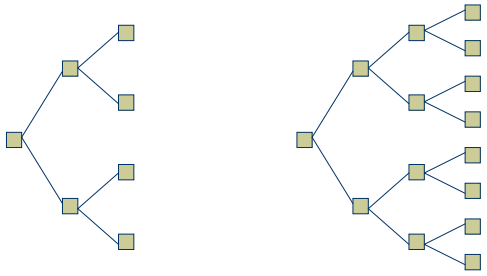
情報量（例による説明）

- 情報を売買する市場を考える．
  - そこでは、神様の集団がいて、将来発生する確率的事象の結果に関する情報を売ってくれる．
  - 2個等確率排反事象（公平なコイン投げ等）の結果に対し1ドルの値がついている．
  - 他の型の事象の結果は、これと等価な価格になっている

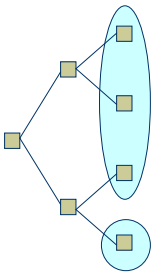
情報量（例による説明（承前））

- では、2^n 個の等確率排反事象の結果を教えてもらう対価はいくらであろうか？
  - n ドルである．何故か？
- では、一般に m 個の等確率排反事象の結果を教えてもらう対価はいくらであろうか？
  - $\log_2$ m ドルである．何故か？
- では、等確率でない、2個の排反事象の結果を教えてもらう対価はいくらであろうか？
  - 結果として発生した事象によって、買った情報の価値が異なってしまうので、長期間の平均を考えるべし

情報量（例による説明（承前））



情報量（例による説明（承前））



更に $\log_2 3$ ドル払えば完全に分かる

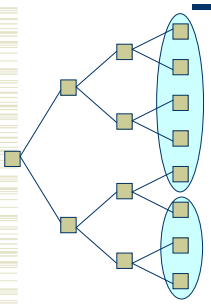
得た情報の価値が異なる

更に払う必要はない

長期間平均でみると
$(3/4) \log_2 3 + (1/4) 0$
ドル払えば完全に結果を知ることができることになる

従って、予め払うべき対価は、長期間この商売を続けるとして、
$2 - ((3/4) \log_2 3 + (1/4) 0)$
$= (3/4) \log_2 (4/3) + (1/4) \log_2 (4/1)$
ドルとなる

情報量（例による説明（承前））



この場合には、予め払うべき対価は、長期間この商売を続けるとして、
$3 - ((5/8) \log_2 5 + (3/8) \log_2 3)$
$= (5/8) \log_2 (8/5) + (3/8) \log_2 (8/3)$
ドルとなる

更に一般には、予め払うべき対価は、長期間この商売を続けるとして、
$p \log_2 (1/p) + (1-p) \log_2 (1/(1-p))$
$= -(p \log_2 p + (1-p) \log_2 (1-p))$
ドルとなる

情報量（例による説明（承前））

- では、等確率でない、m個の排反事象の結果を教えてもらう対価はいくらであろうか？
  - 長期間この取り引きを続けるとして、
    $I(p_1, p_2, .., p_m) = -\sum p_i \log_2 p_i$
    となる
  - この量は、結果を知る前の、結果に対する不確かさ（の期待値・平均値）を表していると考えられる.

情報量の単位ビット

- 「2個等確率排反事象（公平なコイン投げ等）の結果に対し1ドルの値がついている」としてきたが、正式には、これを1ビットというわけである

情報の説明（Modern Approach）

- 以下に示すが、不適当である（間違ってはいないが）

情報量の期待値（例による説明）

- コイン投げの裏表を当てることを考える
- もし結果が高精度で予測できるなら、真の結果を知ることの価値はあまり高くない
  - 1ドルを遣り取りするコイン投げの賭けに参加する
  - もし 0.99 の確率で表がでる細工がしてあり、それを既に知っているとする
  - 当然、表に賭ける. 期待利得は 0.98 ドルである
  - （トスの後で開く前に）胴元から「真の結果を教えてやる」といわれても、対価として 0.02 ドル以上払う気にはならない

情報量の期待値（承前）

- もしコインが公平であれば、結果は予測できず、真の結果を知ることの価値は高い
  - どちらに賭けでも，期待利得は 0 ドルである
  - （トスの後で開く前に）胴元から「真の結果を教えてやる」といわれれば、対価として 1ドル未満まで払う気になる
- 結果の予測が困難なほど，真の結果に関する情報の価値は高い（情報価値は受領人依存）

ノイズと過剰一致（overfitting）

- ノイズがある場合の一つの対処法は示した
  - 多数決をとるか、確率を返すか、賽を振る
- 例えば、重要な情報が欠落していても決定木は作成可能であるし、
- 不適切な属性を使って、見せ掛けの分類を行うことがありうる.

過剰一致の例

- 賽を一日一回振ってその特性を調べる、という実験を行った. その結果は、

例	曜日	月	賽の色	賽の目
X1	月曜	一月	青	5
X2	火曜	一月	赤	3
X3	水曜	一月	赤	5
X4	木曜	一月	青	3
X5	月曜	二月	赤	5
X6	火曜	二月	青	6
X7	水曜	二月	赤	1
X8	木曜	二月	青	4

過剰一致の例（承前）

- どんな決定木が作られるにせよ、まったくの偽の分類となる
- このように、データ中の意味の無い規則性を発見してしまうことを、過学習または過剰一致（overfitting）という
- これは、極めて一般的な問題であり、すべての学習アルゴリズムの課題である

過学習対策

- 統計的に有意な属性のみを用いる
  - χ^2 検定を用いる
  - 決定木では、「枝刈り（pruning）」と呼ばれる
- cross-validation（交差検定）:
  - 例を訓練集合と検査集合に分け、決定木の大きさを変えながら、学習と予測精度の推定を繰り返す。