情報意味論(2) 基礎概念他

櫻井彰人
慶應義塾大学理工学部

機械学習の材料

- 訓練データ・事例、学習データ・事例、
  - 事例＝instance＝sample
  - ある(一般には未知の)確率分布に従って生成される
  - 訓練データ＝独立に生成された事例の集合
- 仮説集合
- 希望の結果との差を示す数値
  - 誤差、誤り率、コスト
- 未知のデータ
  - 学習結果の能力を評価するデータ
  - 訓練データとは異なる

機械学習の手段

- 仮説集合
  - (機械学習の)答えの候補＝仮説
    - 一つの決定木 ＝ 一つの仮説
    - 作りうる決定木の集合 ＝ 仮説集合
- 学習過程
  - 仮説を一つとり、
  - 訓練データをうまく説明するかどうかを調べ
  - 満足が行く仮説であれば、それを答えとする。
  - 不満であれば、上記を繰り返す。

統計学であれば

- 仮説集合＝パラメータ付分布
  - 例: 正規分布(の集合)では、平均値と分散共分散行列がパラメータ
  - ノンパラメトリックは少々異なる
- ある評価関数
  - 例: サンプル集合を最も高い確率で生成する分布(を定めるパラメータ)
- その最良解を求めるもとのする。
  - 解は存在しないかもしれない
  - 計算手段は反復的になるかもしれない

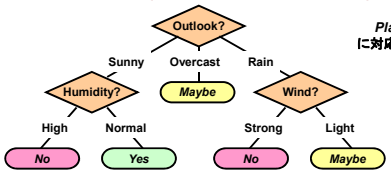
仮説集合の例:

決定木 Decision Trees

Day	Outlook	Temp	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Cloudy	Hot	High	Weak	Yes
D4	Rainy	Mild	High	Weak	Yes
D5	Rainy	Cool	Normal	Weak	Yes
D6	Rainy	Cool	Normal	Strong	No
D7	Cloudy	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rainy	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Cloudy	Mild	High	Strong	Yes
D13	Cloudy	Hot	Normal	Weak	Yes
D14	Rainy	Mild	High	Strong	No

Adapted from Mitchell, 1997

- 分類器 Classifiers である
  - 事例 : 属性 attribute (または特徴 feature) のベクトル＋ラベル
- 内節 Internal Nodes: 属性、または属性値のテスト
  - 典型的: 属性 or 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: 枝を選ぶ条件である属性値 (テスト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

学習過程＝仮説出力過程

- 一般に、学習過程は、仮説を出力する(仮説を仮説空間から選んでくる)過程である。
  - 一回だけ出力する
  - 予め分かる回数だけ出力する
  - 無限回出力する
- 仮説空間が有限な場合(仮説空間に含まれる仮説の個数が有限の場合)
  - 全部を試みて、最適なものを出力する
  - 一部を試みて、その中で、最適なものを出力する
    - 多くの場合、全部を調べる時間がない
- 仮説空間が無限の場合
  - 一部を試してみるしかない
  - 最良なものを見つける方法がない限り、候補を無限に出力し続ける
- いずれにせよ、探索順序が問題

仮説の選択順序

- 仮説を無限回出力する
  - 見かけ上は、一個の仮説を出力して終了
    - 指定した停止基準を満たした場合、終了とする
- なぜ無限回出力するか？
  - 最適解が、繰り返し計算の極限でしか求まらない
- 求める順序が問題
  - 段々「良く」なって行って欲しい
    - （何らかの基準に従い）好きなところで停止できる
  - 実際は、必ずしもそうではない。
- バイアス
  - どの順序をとるにせよ、一般には、見つけうる解（近似解）に片寄りが生じる。これを学習バイアスという。

7

決定木の学習

- 仮説集合は有限集合
  - 離散変数だけの時。ある葉に至る路上では、同じ属性は2度現れない。
  - 連続変数に対して。閾値は無限個ありうるが、異なる結果を出すものは有限個しかない。
- しかし、全数チェックはできない
  - 膨大すぎる。
- どうする？

8

ところで学習とは

9

Induction（帰納）

- 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

10

帰納とは（2）

- 帰納とは
  データに潜在する規則性を得ること
  物体の落下の実験データ → 万有引力の法則
  惑星の公転運動 → 楕円運動、面積速度一定、調和
- 帰納で得た規則の正しさはどう測るか

11

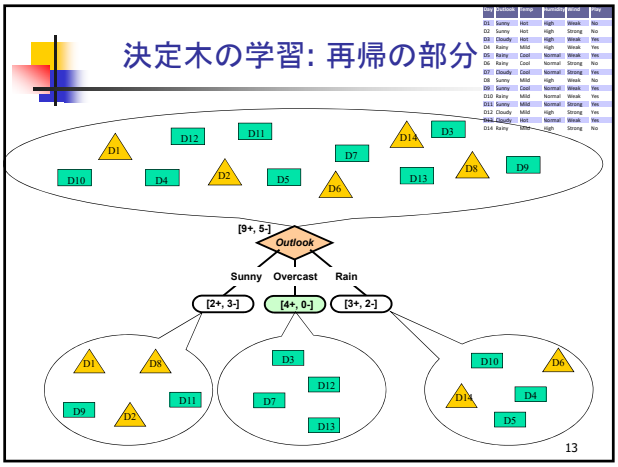
決定木の学習: トップダウン帰納 (*ID3*)

- アルゴリズム *Build-DT* (*Examples*, *Attributes*)
  - 部分木に再帰的に適用される
    - Examples: 事例の部分集合、Attributes: 属性の部分集合

IF *Examples* の *label* が同一 THEN RETURN (その *label* を付した葉節)
ELSE
  IF *Attributes* 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. 以下で作る木を子とする木を作り、値とする.
    FOR *A* のそれぞれの値 *v*
      条件 $A = v$ に対応した, 根節からの枝を作成する
      IF $\{x \in Examples \mid x.A = v\} = \emptyset$
        THEN *多数派 label* を付した葉節 を作成
        ELSE *Build-DT* ($\{x \in Examples \mid x.A = v\}$, Attributes – {A})

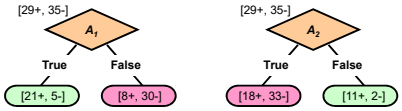
12

決定木の学習: 再帰の部分



決定木の学習: 最良属性とは？ (*ID3*)

- どの属性が最良か?
- 例えば、次の例では、どちらがよいか？



最良属性とは

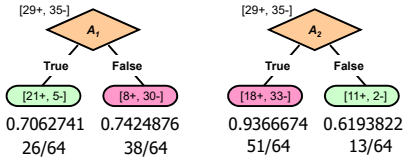
- その属性を選んだ結果、その先、木がより小さくなる方がよかろう
  - 木が小さい方がよさそうなのは、なぜか？
    - 詳細な説明は後日。ここでは、木が小さい⇒葉までの距離が短い⇒少ない属性で決定できる・決定までの判断数が少ない⇒本質に近そう

木の大きさ

＋の副合

最良属性とは: 計算例

属性を選ぶ前のエントロピーは $-\frac{29}{64}\log\frac{29}{64}-\frac{35}{64}\log\frac{35}{64}\approx 0.9936507$



0.7277758　　0.8722188

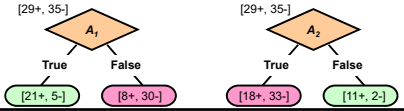
属性を選んだ後のエントロピー

情報量増分: 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 の属性値に基づく分割により削減される不確実性・不純性を計る
- 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?

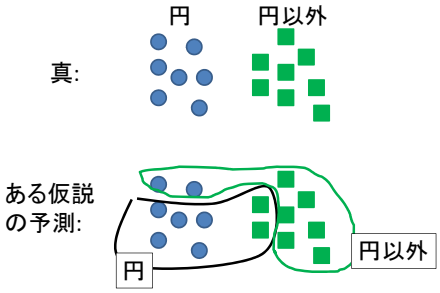


仮説の評価

- 仮説（モデルとも言います。文脈によります）の良さ・悪さを評価する。
  - 仮説は、使う
  - 何らかの意味で「精度」や「信頼性」の高い仮説を用いたい。
  - どんな評価方法がありうるか？

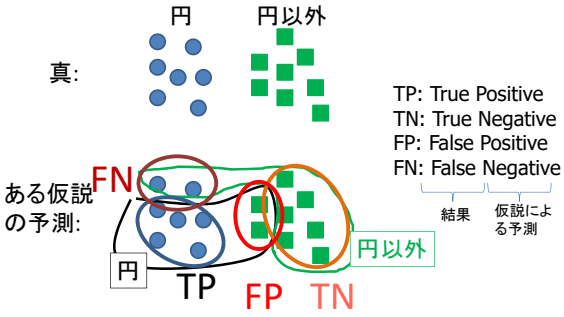
学習結果の評価

Precision と Recall の前に



19

TP, TN, FP, FN



20

Confusion matrix

		真値 P	真値 N
仮説の予測値	P	**TP** (True Positive)	**FP** (False Positive)
	N	**FN** (False Negative)	**TN** (True Negative)

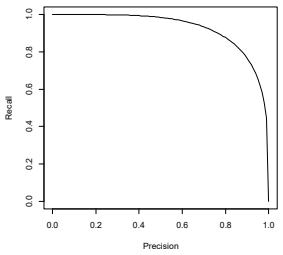
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

21

両者のTradeoff と F-measure



$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right)}$$

22

Confusion matrix

1-β=sensitivity =TPR= $\frac{TP}{FN+TP}$

		真値 P	真値 N
仮説の予測値	P	**TP** (True Positive)	**FP** (False Positive)
	N	**FN** (False Negative)	**TN** (True Negative)

$\alpha = \text{FPR} = \frac{FP}{FP+TN}$ 第一種の過誤

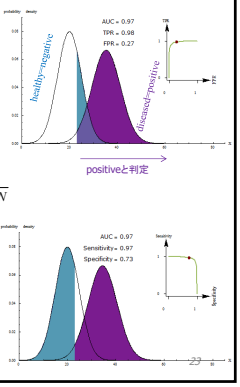
$\text{Precision} = \frac{TP}{TP+FP}$

1-α=specificity =TNR= $\frac{TN}{FP+TN}$

$\text{Recall} = \frac{TP}{TP+FN}$ 第二種の過誤

$\beta = \text{FNR} = \frac{FN}{FN+TP}$

$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$



帰無仮説は、「陽性でない」
（陽性であることを示したいから）

第一種の過誤＝棄却した(陽性だと言った)が、それは誤り
第二種の過誤＝受理した(陰性だと言った)が、それは誤り

23

ROC curve

- Receiver operating characteristics
  - "ROC"という用語はレーダが開発された当初、操作盤上にあったノブの名
    - http://www.math-koubou.jp/stata/files/r12/est006.pdf

24

ROC curve

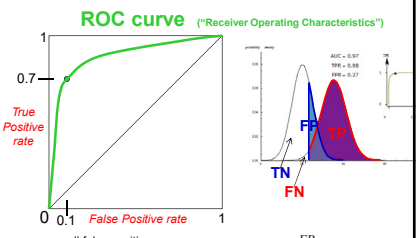
仮説の予測値 \ 真値	P	N
P	TP (True Positive)	FP (False Positive)
N	FN (False Negative)	TN (True Negative)

$$\frac{\#\text{ true positives}}{\#\text{ true positives} + \#\text{ false negatives}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\beta = \text{FNR} = \frac{FN}{FN+TP}$$

1-β=sensitivity



$$\frac{\#\text{ false positives}}{\#\text{ false positives} + \#\text{ true negatives}} \quad \alpha = \text{FPR} = \frac{FP}{FP+TN}$$

1-α=specificity

ROC Curves

- 閾値（後述）を変えながら、サンプルの正誤を数える
- area under the curve (AUC) が大きい方がよい
- 異なった学習手法の性能を比較するのに適している

25

訓練誤差と汎化誤差

- 訓練誤差・誤率: 訓練データ学習データに対する、仮説出力値の、真の出力値に対する誤差・誤率
  - 簡単に数えることができる。
- 未知データに対する誤差・誤率：（訓練データと同じ母集団から、同じ方法で抽出した）未知データに対する、仮説出力値の、真の出力値に対する誤差・誤り率。テストエラーともいう。

26

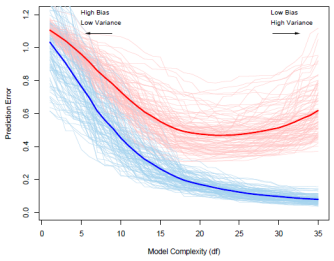


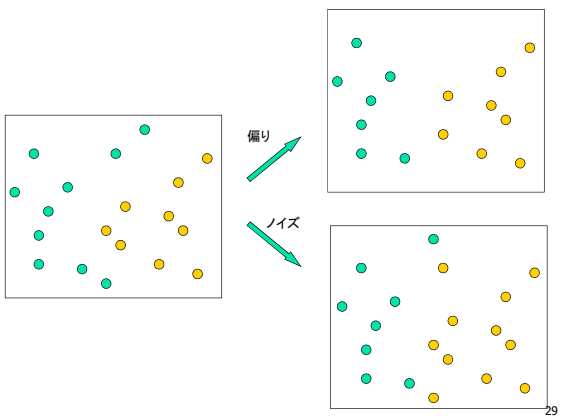
FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Elements of Statistical Learning 27

過学習

- over-learning とか over-training と呼ばれる
- 学習すべきでないものまで、学習してしまう
- 学習すべきでないもの
  - 学習データに含まれる偏り
    - 無限集合（真の概念が含む事例は無限個ある）の有限部分集合であるため、かならず、偏りがある。
  - 学習データに含まれる誤り
    - 現実データにはノイズがある。分類クラスにも属性値にもノイズは存在する。
- 学習してしまう
  - 学習能力が高いから
    - 調節可能なパラメータ数が多い

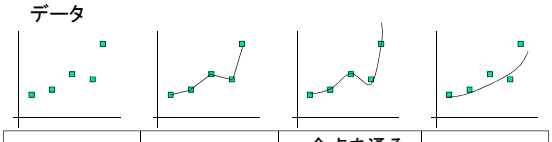
28



29

再掲: 関数近似の例（ノイズ）

データ



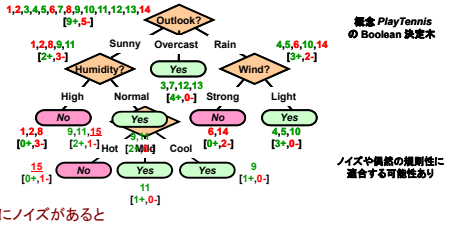
	区分線形	全点を通る4次多項式	2次多項式
パラメータ数	2×4+3=11	5	3+ノイズ

多分過学習？　　多分過学習

30

決定木における過学習: 例

- 既出例: 帰納した木



概念 *PlayTennis* の Boolean 決定木

ノイズや偶然の規則性に適合する可能性あり

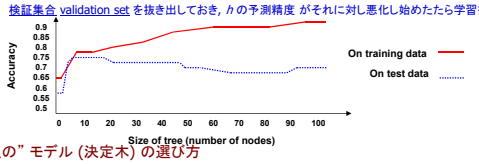
- 訓練事例にノイズがあると
  - 事例 15: <*Sunny, Hot, Normal, Strong, ->*
    - この例は実は noisy である. すなわち、正しいラベルは +
    - 以前に作成した木は、これを、誤分類する
  - 決定木はどのように更新されるべきか (incremental learning を考える)?
  - 新しい仮説 $h' = T$ の性能は $h = T$ より*悪く* なると予想される(*ノイズ*に騙されているから！)

帰納学習における過学習

- 定義
  - 仮説 h が訓練データ集合 D を過学習する (〜にoverfits する) というのは、もし他の仮説 h' で $error_D(h) < error_D(h')$ であるが $error_{test}(h) > error_{test}(h')$ となるものがあること
  - 原因: 訓練事例が少なすぎる (あまりにも少ないデータに基づく決断); *ノイズ*; 単なる偶然
- 過学習に対応するには?
  - 予防策
    - 過学習が発生する前に対応する
    - *重要な relevant* 属性(i.e., モデルにとって有用そうなもの)のみを用いる
      - *注意*: 鶏と卵の問題; 重要性 relevance を予測する尺度が必要
  - 回避策
    - 問題が起こりそうなときに、脇をすりぬける
    - テスト集合を確保しておき, 仮説 h がその上で悪くなりそうなときに、学習を停止する
  - 泳がせ策
    - 問題は発生するにまかせ, 発生を検出し, その後回復する
    - モデルを作ってみて, 過学習に寄与する要素を発見・除去する (刈る prune)

決定木学習: 過学習の予防と回避

- 過学習にどう立ち向かうか?
  - 予防策
    - *重要な* 属性を選択 (i.e., 決定木では有用)
    - *重要性* の予測: 属性を filter する, または 部分集合選択
  - 回避策
    - 検証集合 validation set を抜き出しておき, h の予測精度 がそれに対し悪化し始めたたら学習を停止



- "最良の" モデル (決定木) の選び方
  - 上述: 性能を測定するにあたって、訓練データとそれとは別の検証データを用いる
  - 別法: 最小記述長 Minimum Description Length (MDL):
    最小化せよ: *size*(h ≡ T) + *size* (*誤分類 misclassifications* (h ≡ T))

決定木学習: 過学習の予防と回避

- 基本的なアプローチが2つある
  - Pre-pruning (回避): *木を作成する途中で木の生長を止める*. 信頼性ある選択をするにたる十分なデータはないと判断されたとき
  - Post-pruning (回復): *木を一杯まで構築し節を削除する*. 削除するのは, 十分な証拠がないとみなされるもの
- 枝刈りすべき部分木を評価する方法
  - Cross-validation: 仮説の有用性を評価するために、予めデータをとりおく (Mitchell 第4章)
  - 統計的検定: 観測された規則性が偶然起こったものとして捨ててよいかどうかをテストする (Mitchell 第5章)
  - 最小記述長 Minimum Description Length (MDL)
    - 仮説 T の複雑度の増加分は, 単に(説明しようとしているデータの)*例外* を記憶するに必要な記述量より大きいか/小さいか?
    - Tradeoff: *モデル* を記述する versus *残余誤差* を記述する

学習とバイアス

- バイアス: 仮説間に順位があるとき、その順位
  - 同時に複数個の仮説をみたときの、選好順位
  - 一度に一個ずつ見るときの、探索順序
- データに適合する仮説は、一般に、多量にあるので、学習するにはバイアスが必要
  - 仮説を一個選択するのではなく、複数個の仮説を用いる場合でも、「データに適合する仮説をすべて用いる」のではない限り、バイアスが必要である。

学習:

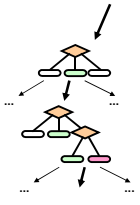
データ → 仮説

いや、しかし、データ以外の情報を使って、仮説を選択するのはまずいのではないか？

もし、バイアスが避けえないとしたら、どういうバイアスがよいのか？

ID3 による仮説空間探索

- 探索問題
  - 探索の対象は *決定木全部の空間*, すなわちブール関数をすべて表現可能な空間
    - Pros: 表現力; 柔軟性
    - Cons: 計算量; 巨大, 意味の分からない木も含む
  - 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
    - heuristics の使用(探索の案内役としての目の子)
    - 貪欲 greedy アルゴリズムの使用
    - すなわち、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_i 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

- 探索におけるヒューリスティックは帰納バイアスである
  - H は X の冪集合 (全部分集合の集合)
  - ⇒ 帰納バイアスなしと言ってよいか？ いや、そうではない…
    - 短い木への選好 (終了条件から) がある
    - 情報量増分が高い属性を根節に近いところにおくという選好がある
    - *Gain*(•): *ID3* の帰納バイアスを体現するヒューリスティック関数
  - *ID3* の帰納バイアス
    - ある仮説への選好をヒューリスティック関数に表現している
    - 比較してみる: 仮説空間 H を制限すること(命題論理の正規形に基づく制限: 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をとれ

学習時に用いる、データ以外の仮定。それにより、こちらの仮説がより良い、この仮説はとらない、この仮説はとるということが決まる
これがないと、データを説明する仮説が多数(無限に)あって、結論が得られない

いや、しかし、データ以外の情報を使って、仮説を選択するのはまずいのではないか？

もし、バイアスが避けえないとしたら、どういうバイアスがよいのか？

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.

オッカムの剃刀: ある選好バイアス

- 帰納バイアス2つ: 選好バイアス preference biases と言語バイアス language biases
  - 選好バイアス
    - *学習アルゴリズム*に(普通は暗黙的に)組み込まれている
    - 言い換えれば: 探索順序の規定
  - 言語バイアス
    - *知識 (仮説) の表現*に(普通は暗黙的に)組み込まれている
    - 言い換えれば: 探索空間の制限
    - *別名* 制限バイアス
- オッカムの剃刀 Occam's Razor: 賛成意見
  - 短い仮説の方が、長い仮説に比べ、個数が少ない
    - 例えば, ビット列で考えれば, 長さ n のものは $n+1$ のものに比べ半数, $n \geq 0$.
    - 短い仮説が、もしデータにぴったり合ったとしたら、偶然とは考え難い
      - 短い仮説は、個数が少ないので、説明できる現象の数が少ない
    - 長い仮説 (例: 200 個の節を持つ木,かつ $|D| = 100$) の場合には、偶然である可能性が高い
      - いずれかの木がデータにぴったり合う。どれに合うかは偶然であるが、どれかに合うこと自体は当然。
  - 得るものと捨てたもの
    - 他の条件が同一であれば, 複雑なモデルの汎化能力は単純なモデルほどではない
    - あとになってもっと柔軟な(微調整可能な)モデルが必要になることはないと仮定

オッカムの剃刀と決定木: 二つの問題

- オッカムの剃刀 Occam's Razor: 反対意見
  - 仮説空間 H に依存して *size*(h) が決まる。同じ h でも H が異なると *size*(h) が異なる。
  - 「小ささ」を選好することへの疑問: "少ない" ことは正当化にならない
- オッカムの剃刀 Occam's Razor は Well-Defined か?
  - 内部の知識表現 knowledge representation によってどの h が "短い" かがきまる --- 恣意的?
    - 例えば, テスト "(*Sunny* ∧ *Normal-Humidity*) ∨ *Overcast* ∨ (*Rain* ∧ *Light-Wind*)" は一個?
  - 答: 表現言語を固定; 十分長いところでは、長い仮説は、内部表現によらず、やっぱり長い
    - 反論: 答えになっていない。実際には「短い仮説」に関する議論が重要
- 「短い仮説」であって、どうして他の「小さい仮説空間」ではないのか?
  - 小さい仮説集合を定義する方法はいろいろとある.
  - 選好バイアスで用いる size が何であっても, 適当に基準 S を選べば *size*(h) をその限界内に制限することができる (i.e., "S に合致する木のみ受理する")
    - e.g., 節の個数が素数であって, 文字 "Z" で始まる属性を用いている木
    - なぜ「小さな木であって, (例えば) A_1, A_2, …, A_{11} を順番にテストするもの」ではないのか?
    - *size*(h) に基づいて小さな仮説集合を定義することに、特別の意味があるのか?
  - 参考: Chapter 6, Mitchell's Machine Learning

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
- Bayesian アプローチと比較してみよう