情報意味論（1）

慶應義塾大学理工学部
櫻井 彰人

この講義では

- 機械学習のいくつかの代表的な手法を知る
  - 基本原理
  - 基本アルゴリズム
  - 実際に使ってみよう
  - 少しアルゴリズムに触ってみる

講義形態

- 普通の講義形態
- できるだけ、動作例を見ててもらう
- シラバスから順序等多少変更あるかも
- 確率・統計の基礎はできるだけ省略
- Weka と R は道具として使うが概説のみ

2012年度予定

1	9月24日	月	情報と意味と機械学習
2	10月1日	月	RとWeka
3	10月15日	月	決定木と過学習
4	10月22日	月	コネクショニズム
5	10月29日	月	多層神経回路網
6	11月5日	月	ベイズ学習
7	11月12日	月	モデル選択
8	11月26日	月	EMアルゴリズム
9	12月3日	月	ベイジアンネットワーク
10	12月10日	月	事例ベース学習
11	12月17日	月	SVM
12	1月7日	月	Boosting
13	1月15日	火	相関規則
14	1月21日	月	強化学習

機械学習

- データから意味を抽出する作業を、従来から、機械学習とよんできた
- 機械学習（machine learning）：
  - データ間の規則性（意味）の抽出（学習）を計算機に行なわせる
    - これは「学習」か？ yes!
  - 知識獲得ともいう
    - 規則性が知識だって？
  - 適応（adaptation）でもある。
- データを集めて情報となすことにはかわりない

学習

- 少しずつ異なった意味で用いられるが
  - 外界と自分があるときに、自分を少しずつ変化させて、外界に適応する（よりよいメリットを得る）こと
  - すなわち、対象とする系の表現・表出に基づき、最適行動を計画・実行する
    - そのために、ある系の振舞い（データ）をもとに、その系を表現する（本質をとらえた一般記述）ことが必要

学習

- もっと一般化して考えると、学習とは
- 具体例を知り、具体例を一般化すること
  - 丸暗記という学習もある。
- 具体例（ instance ）を一般化する。
  - りんご1が落ちた、栗2が落ちた、、、
    ⇒ 物体は支えがなくなれば落ちる
  - 叩いたら痛かった：一週間前、昨日、今日、、、
    ⇒ 叩くと(いつでも)痛い
  - 隣のAさんはケイタイを持っている、会社のBさんも、、、、
    ⇒ みんなケイタイを持っている
- 特徴：間違っているかもしれない
  - わずか(有限個の)具体例に基づくので当然。

機械学習

- 「機械学習」はこの「一般化」を理論化するにあたり、結果の正しさ(という評価基準は常に必要)を、
  - 具体例が無限個になれば、正解が得られる、すなわち、
  - 具体例が無限個になれば、モデルが同定できるような

  一般化を求めることにした。
  - 後に、この「モデル同定」でない、機械学習の特徴づけ(PAC)がなされ、機械学習のさらなる発展が起こることになる
- データ(対象とする系の動作の具体値（ instance ）をもとに、その系の記述を得る、その系を同定する。
  - 2, 4, 6, 8, 10, 12,,,, ⇒偶数
  - 1, 2, 4, 8, 16, 32,,,, ⇒ 2 の冪乗

機械学習

- ますます「学習」から離れるかもしれないが、「学習」の本質は捉えている

  学習：経験(具体例)をもとにパフォーマンスを上げる

  (パフォーマンスを上げるには、未経験の事例に対しても、うまく動作する必要があるので)
  学習：経験(具体例)をもとに未知の(類似の)事態に対応すること

  そのためには、相手(外界)を知ることが必要。知るとは記述できること。
  本質：経験から(相手の)記述を帰納すること。未知事例に対して適用する。

学習の実例1　実世界

ロボットにペナルティキックをさせたい。もし関与するすべての物体の力学的性質が分かり、数値が測定可能かつ天候・芝の状態、キーパーの癖等がわかれば、最適なキック方法が選択できる。しかしそのようなことはない。どうするか？

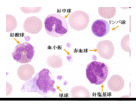
自動清掃ロボットを作りたい。顧客ごとに部屋の配置を入力させるのは(入力するのは)大変だ。ロボット自身に「学習」させたい。どうしたらよいか？

ルンバは学習しない。Brooksの基本的考え

学習の実例2 パターン認識

郵便番号(宛先)自動読み取り装置：
郵便番号・住所として書かれた文字のデータが10000組ある。これをもとに、宛先を読みとり分配するシステムを作るにはどうしたらよいか？

血液像自動分類装置：(診断に必要な)白血球・赤血球の画像とその分類例が10000枚ある。これをもとに、血球を自動分類しその個数を数える装置を作るにはどうしたらよいか。



学習の実例3　膨大なデータ

世界中にあるWWWページを自動的に収集・分類し、ユーザが指定した観点から自動的に類似性を判定し、関連性・塊を表示するシステムを作りたい。どうしたらよいか？

1TBあるアクセスログから、注文につながる、また離脱するユーザの行動に基づいて、うまくリコメンドし、注文につなげたい。

学習の実例4 IBM's Watson

- 米国のクイズ番組Jeopardy!(ジョパディ!)に挑戦し、2ゲームを通じて、最高金額を獲得した(2011年2月16日(米国時間))。
- 知識は学習手法を用いて蓄積
  - 100万冊の本を読むのに相当する自然言語で書かれた情報
- ラック10本分、総メモリー容量15TB、総プロセッサー・コア数は2,880個

学習の実例5 コンピュータ将棋

- よく知られるようになったのは、渡辺竜王 vs. ボナンザ戦(大和証券杯特別対局)。

学習の実例6 自動走行
ニューラルネットワーク

Autonomous Learning Vehicle In a Neural Net (ALVINN): Pomerleau *et al*
Navlab-5 に到り終了 (1995). 高速道路を 70mph で. "No Hands Across America"
http://www-2.cs.cmu.edu/afs/cs/user/tjochem/www/nhaa/nhaa_home_page.html

最近:

Google:

http://www.nytimes.com/2010/10/10/science/10google.html?pagewanted=1&_r=2&partner=rss&emc=rss

DARPA
2004 Grand Challenge
2005 Grand Challenge
2007 Urban Challenge

蛇足 2012 Robotics Challenge

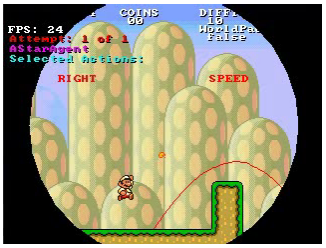
注:人工知能

- 二つの立場
  - 人間の知能そのものをもつ機械を作ろう
  - 人間が知能を使ってすることを機械にさせよう
- 後者が普通。
- 機械学習の技術も使うが、使わなくてもよい
- ロボット(知能機械)の動作に、人工知能技術は必ずしも必要ない。機械学習技術も同様
- 一方、ロボット(知能機械)でなくても、機械学習技術が必要なところはある。人工知能技術も同様

これは人工知能?

- 多分、人工知能でも、機械学習でもない
  - 勿論、使うことは可能であろうが、、、

これは？



- 人工知能です。探索技術を使っている
  - 機械学習はしていない

http://www.doc.ic.ac.uk/~rb1006/projects:marioai

ではこれは？

Michael Schmidt and Hod Lipson, "Distilling Free-Form Natural Laws from Experimental Data, Science, Vol. 324, April 3, 2009.

機械学習

- 機械学習（machine learning）:
  - データ間の規則性（意味）の抽出（学習）を計算機に行なわせる
    - これは「学習」か？ yes!
  - 知識獲得ともいう
    - 規則性が知識だって？
  - 適応（adaptation）でもある。
    - 外界（自分以外の世界）の変化に自分を合わせる

ところで、何故情報意味論？

- もともとは、データと情報と意味を議論する講義であった（にしたかった）
  - データから意味・情報をとりだす
    - 考え方と方法
  - 取り出し方
    - 学習理論とアルゴリズム
    - 2つの方法：記号的な方法、統計的な方法
  - 応用
    - 様々な adaptation
    - データマイニング
- その中でも「学習」に重点を置くことにした

情報とは何か？

- 英語では information
  - Inform がもとの動詞。どう使う？
- 日本語：いつごろ訳したか？
  - 情とは
  - 報とは

（小野厚夫, "情報という言葉を尋ねて" (1)～(3), 情報処理(2005) を参照）

（インフヲルメーションで調べてみよ）

意味とは何か

- ①記号・表現によって表される内容またはメッセージ。② 物事が他との連関において持つ価値や重要さ。（広辞苑）
- 動作で考えてみよう。例えば、「意味がある」行動とは？
- 次に、表現と意味との関係を考えてみよう。
  - 現実世界における「表現」は常に、冗長である。では、徹底して冗長性を排除したらどうなるか？
  - なぜ、冗長なのかも考えてみよう

情報理論における情報

- データを生み出す「データ源」の記述
  - 例1：0は確率1/4で，1は確率3/4でランダムに生成する
  - 例2: n番目には、n番目の素数の10進第一位を生成する

データ源の記述ができると

- データ源の記述ができると何がよいか？
  - 予測ができる
  - もしそれがノイズ源であれば、ノイズを効果的に低減することができる
  - （もっと一般的には）制御することができる

例えば、

- 一つの音源の音を正確に採取するために、複数のマイクを使う。
  - 2つのデータ中の相関の大きな成分が当該音源の音である（ノイズには相関がない）
- 経済予測：株価予測、売上予測
  - 潜在需要の発見とその利用（刺激して新市場創造）
- 物理現象・化学現象・社会現象の記述と予測

つまり、機械学習

- 要は、
  - 目的、方法、評価方法は様々であれ、
  - データから意味（これって、目的によって変わります）をとりだすこと

が機械学習

データマイニングとは？

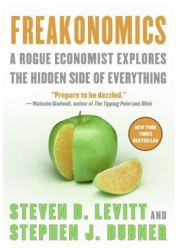
- データマイニング(データベースからの知識発見):
  - 興味深い(当たり前でない、潜在的、これまで知られていなかった、しかも、役に立つと思われる)情報あるいはパターンを大規模データベースから抽出すること
- データマイニングの別名
  - データマイニング：命名を間違えた？
  - データベースからの知識発見(Knowledge discovery in databases, KDD)、知識抽出、データ/パターン解析、データ考古学、情報収穫、ビジネスインテリジェンス、など
- データマイニングでないのは何か？
  - (演繹)質問応答処理
  - エキスパートシステムあるいは小規模な機械学習システム/統計パッケージ

データマイニングの応用例

- データベース解析と意思決定支援システム
  - マーケット分析とマネジメント
    - ターゲット・マーケッティング、CRM(customer relation management)、購入品目分析（market basket analysis）、マーケット区分（market segmentation）
  - 危機分析とマネジメント
    - 予測、顧客維持、保険の査定の改善、品質管理、競争力分析
  - 不正検知と管理：アクセスログ解析
- 他の応用
  - テキストマイニング(電子メール、webドキュメント、ブログ)
  - Web アクセスログ解析
  - 遺伝子解析（文献解析含む）

これはデータマイニング？

- 経済学？ yes.
  - 経済的インセンティブを取り扱っている
- データマイニング？ yes
  - 多量データの分析結果に基づく



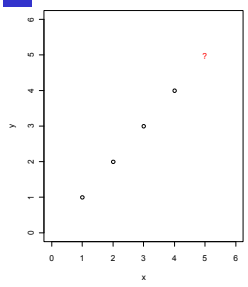
例えば

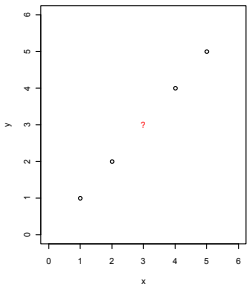
- 相撲の星取り表から
- 全米統一テストの結果から
- 中絶率と犯罪の発生件数から
- ベーグルの料金回収率から

蛇足: なぜ機械学習か？

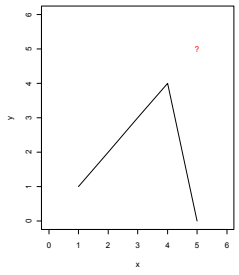
- 様々な意味で「計算能力が向上」
  - データベースマイニング: データを知識に
  - 自動カスタマイズプログラム: ニュースのフィルタ、適応的な監視カメラ
  - 行動の学習: ロボットの計画, 制御の最適化, 決定支援
  - プログラム困難なアプリケーション: 自動運転, 音声認識
- 人間の学習や教育のよりよい理解を求めて
  - 認知科学: 知識獲得の理論 (e.g., 実践を通じて)
  - パフォーマンス向上: 推論・推測, 推薦システム
- 時は今、、、
  - 学習アルゴリズムや理論の最近の進歩は目覚しい
  - 様々なソースから大量のオンラインデータが提供される
  - 計算機は安価・高速
  - 機械学習を用いた事業が発生・成長 (e.g., データマイニング/KDD)

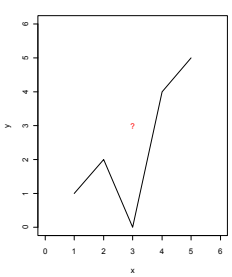
予測と推測・推定

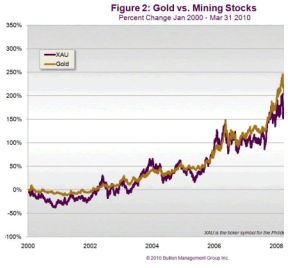




予測と推定・推測







http://www.safehaven.com/article/17497/why-bullion-is-outperforming-mining-stocks

ランダムウォーク S が $2n$ 歩後に $2l$ ($-n \le l \le n$ とする) の地点にいる確率は

$$P(S_{2n}=2l)=\binom{2n}{n+l}\frac{1}{2^{2n}}=\frac{(2n)!}{(n+l)!(n-l)!}\frac{1}{2^{2n}}$$

ランダムウォーク S が $2n$ 歩後に $a\sqrt{2n}$ 以上 $b\sqrt{2n}$ 以下である確率は

$$P\left(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}\right)=\frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{1}{2}y^2}\,dy$$

ただし、$-\sqrt{2n} \le a \le b \le \sqrt{2n}$

逆正弦定理

定理(逆正弦法則) ランダムウォーク S が $2n$ までの間に正の側で $2k$, 負の側で $2n-2k$ 過ごす確率 $P(n, k)$ は

$$P(n,k)=u_k u_{n-k}$$

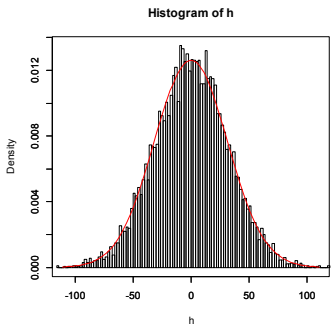
である

定義 $$u_0=1, u_n=\binom{2n}{n}\frac{1}{2^{2n}}=\frac{(2n)!}{n!n!2^{2n}}$$

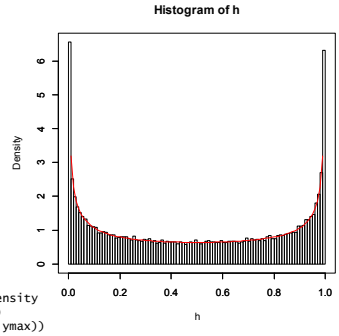
P(ランダムウォーク S が $2n$ までの間に正の側にいる割合 $\le \alpha$)

$$=\sum_{k\le \alpha n} P(n,k) \approx \sum_{0\le \frac{k}{n}\le \alpha} \frac{1}{\pi\sqrt{k(n-k)}}=\sum_{0\le \frac{k}{n}\le \alpha} \frac{1}{\pi\sqrt{k(n-k)}}=\frac{1}{\pi}\sum_{0\le \frac{k}{n}\le \alpha} \frac{\frac{1}{n}}{\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)}} \approx \frac{1}{\pi}\int_0^{\alpha}\frac{dx}{\sqrt{x(1-x)}}=\frac{2}{\pi}\arcsin\alpha^{\frac{1}{2}}$$

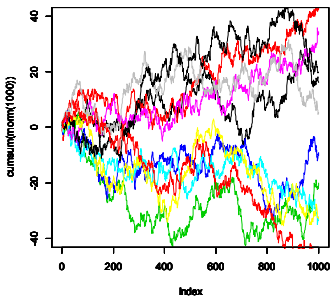
http://elis.sigmath.es.osaka-u.ac.jp/~nagahata/20070816/arcsin.pdf




```
set.seed(123)
rep <- 10000
N <- 1000
br <- 100
h <- numeric(rep)
h<-numeric(rep); for( i in 1:rep) h[i]<-sum(rnorm(N))
hc <- hist(h,freq=F,breaks=br)$density
ymax <- max(hc)
hist(h,freq=F,breaks=br,ylim=c(0,ymax),xlim=c(-3.5*sqrt(N),3.5*sqrt(N)))
par(new=T)
plot(function(x) dnorm(x,0,sqrt(N)), col=2, ylim=c(0,ymax),
      xlim=c(-3.5*sqrt(N), 3.5*sqrt(N) ) , xlab="", ylab="")
```

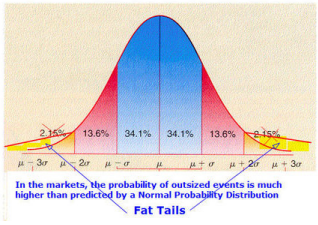



```
set.seed(123)
rep <- 50000
N <- 1000
br <- 100
h <- numeric(rep)
for ( i in 1:rep) {
  t <- cumsum(rnorm(N))
  h[i] <- length(t[t >= 0])/N
}
hc <- hist(h,freq=F,breaks=br)$density
ymax <- max(hc[1],hc[length(hc)])
hist(h,freq=F,breaks=br,ylim=c(0,ymax))
par(new=T)
plot(function(x){(1/pi)/sqrt(x*(1-x))},col=2,
    xlim=c(0,1),ylim=c(0,ymax),xlab="",ylab="")
```

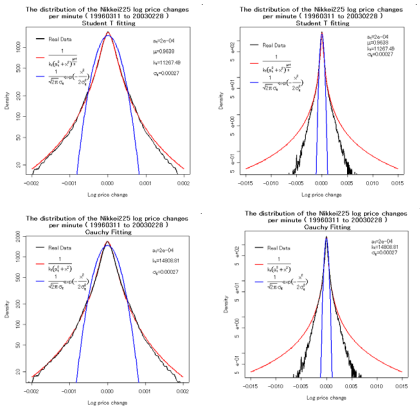



```
set.seed(123)
for ( i in 1:10)
  { plot(cumsum(rnorm(1000)),col=i,type="l",ylim=c(-40,40));par(new=T) }
```

正規分布でない世界なら



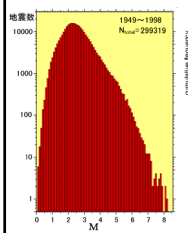
http://stephenvita.typepad.com/alchemy/2010/08/adjustments-8312010.html

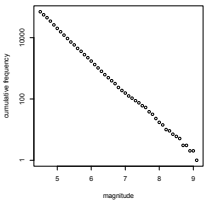


安定分布(stable distribution) がよくフィットするとされている。しかし、実データは安定分布より早く裾が落ちる。そこで、切断安定分布(truncated stable distribution) を考えることがある。しかし、今度はきりすぎになりやすい。

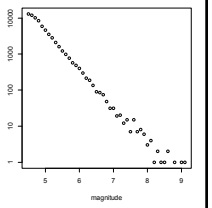
http://www18.tok2.com/home/kabutaro/econophysics/N225perMinDistribution.htm

例えば、地震





全世界で2000年1月1日～2012年8月30日間に検知された地震のマグニチュードに対する積算度数

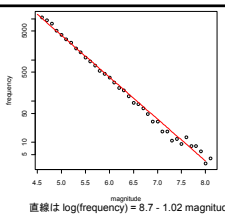


全世界で2000年1月1日～2012年8月31日間に検知された地震のマグニチュードに対する度数

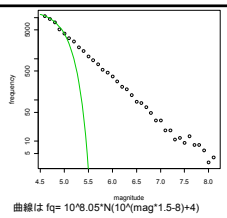
http://earthquake.usgs.gov/earthquakes/eqarchives/epic/epic_global.php

1949～1998年の50年間に日本周辺で検知された地震のM別頻度分布（気象庁データによる）

http://www.hinet.bosai.go.jp/about_earthquake/sec1.2.html

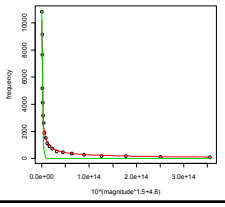


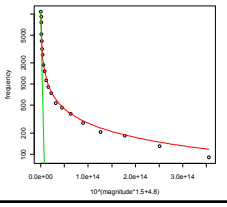
直線は log(frequency) = 8.7 - 1.02 magnitude.



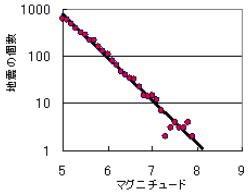
曲線は fq= 10^8.05*N(10^(mag*1.5-8)+4)

直線は log(frequency) = 8.7 - 1.02 magnitude. なお、USBSのデータを用いているため、マグニチュードはモーメントマグニチュードである。いわゆるエネルギーの対数の線形関数である

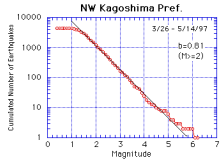




地震の規模と発生頻度との関係

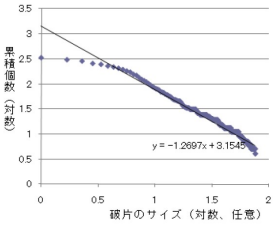


データは理科年表（国立天文台編）2005の693ページのデータを使用。（北緯25～48°、東経125～150°の範囲[日本列島を囲む範囲]で、1961年から1999年の間に気象庁が決めたM5以上の地震）
ただし、M8.0の地震の個数が0であるためにM8.0以上の2個のデータを除外している。採用したデータが示す直線はグーテンベルグ-リヒターの関係を示しており、式は log n(M)=7.490-0.919M で表される。



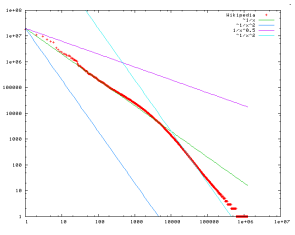
マグニチュードの度数分布. b値は0.81程度と推定された. 1997年5月13日の鹿児島県北西部地震前の地震

べき分布



はやぶさの分裂破片のサイズ分布

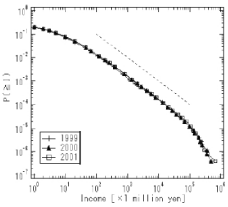
http://pholus.mtk.nao.ac.jp/~satomk/hayabusa/result/



英単語の出現頻度の Zipf 則。横軸が（wikipedia に書かれている）単語の出現数の順位、縦軸が出現数。

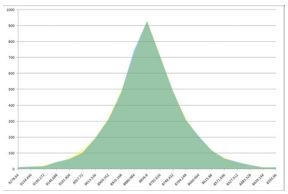
http://en.wikipedia.org/wiki/Zipf%27s_law

べき分布（続）



日本の企業の年間の総収益の分布。べきが1のべき分布がきれいに見えている。（東工大・高安研究室提供。"Misako Takayasu and Hideki Takayasu" To appear in Encyclopedia of Complexity and Systems Science,S.Havlin(Eds.), Springer Fractals and Economics.

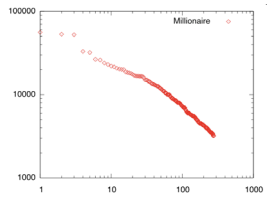
http://www.gakushuin.ac.jp/~881791/modphys/08/



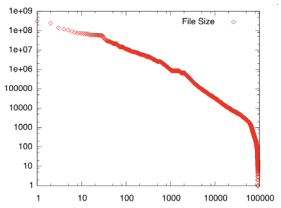
過去20年間の日経平均の騰落率（終値から終値まで1日間）のグラフに指数ベキ分布を重ね合わせたグラフです。（度数と分散が同じになるように調整した後、pを調整して相関係数が高くなるように調整してあります。p=1.278。

http://reiko0309.blog41.fc2.com/blog-entry-486.html

べき分布（続）



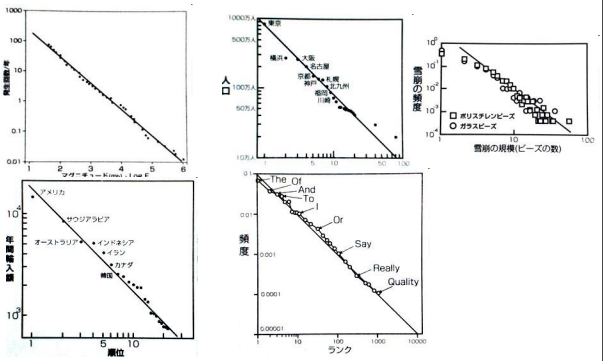
世界の長者番付データ（Forbes 2007. Wikipedia 2007年度世界長者番付）をもとにした、富豪の資産の順位



増井俊之氏のホームディレクトリに入っているファイルを大きさ対順位

http://www.pitecan.com/articles/WiredVision/wv06/index.html

べき分布（続）



http://hiroshi-kobayashi.at.webry.info/200805/article_27.html

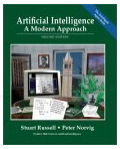
関連領域

- 認知科学: 言語獲得、推論の学習
- 統計学: バイアス vs. 分散, 信頼区間, 仮説検定
- ベイズの方法: ベイズの定理、欠測値の推定
- 人工知能: 記号表現、計画、知識を用いた学習
- 計算の複雑さの理論: PAC 学習、VC次元、誤差限界
- 制御理論: 最適化、動的計画、予測の学習
- 情報理論: エントロピー, MDL, 情報源符号化
- 神経科学: 人工神経回路網、脳（大脳、小脳、視床下部）
- 哲学: オッカムの剃刀, 帰納的一般化
- 心理学: 練習の冪法則(Power Law of Practice) 発見的学習

機械学習環境

- Weka: Waikato大学開発
  - http://www.cs.waikato.ac.nz/ml/weka/
- RapidMiner:
  - http://rapid-i.com/content/blogcategory/10/69/
  - 旧名: Yale: yet another learning environment
    - http://www-ai.cs.uni-dortmund.de/SOFTWARE/YALE/index.html
- R: 統計計算用言語・パッケージ
  - http://www.r-project.org/
- 掲示板
  - http://www.kdkeys.net/forums/

参考書等



- **パターン認識と機械学習**
- **Thomas Mitchell, Machine Learning, McGraw-Hill.**
- **Stuart Russell, Peter Norvig, エージェントアプローチ人工知能, 共立出版**
  - Artificial Intelligence: A Modern Approach (2nd edition), Prentice Hall
- http://www.sakurai.comp.ae.keio.ac.jp/