

# 知的情報処理

## 1. 導入

櫻井彰人  
慶應義塾大学理工学部

## 本講義の目的

- データに基づき未知を予測・推測する方法の基礎を知る
  - 「未来」の予測とは限らない
  - 「未知」の推測もある
- その基礎である機械学習を知る
  - 統計的手法も知る
- 道具として用いる R を知る

## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- この講義について

## 本日の第一部



### 予測

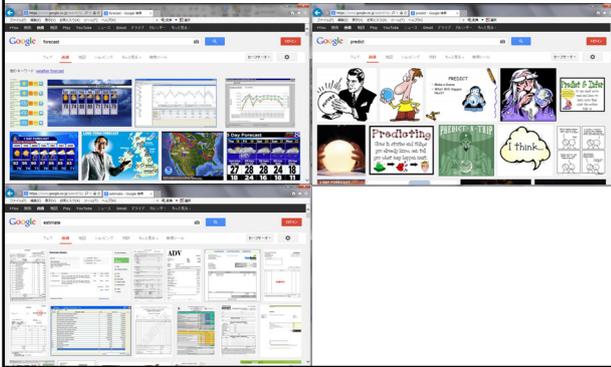
- 事の成り行きや結果を**前もって**おしはかること。また、その内容。(デジタル大辞泉)  
例文: 10年後の人口を予測する
- **将来**の出来事や状態を**前もって**おしはかること。また、その内容。**科学的根拠**が重んじられる。(大辞林)
- **将来**どうなるかを得られた情報などに基づいておしはかること。また、そのようにして得たもの。「予想」は将来を推測する意で広く使い、「予測」は**具体的なデータ**などに基づく意で使うことが多い。(明鏡国語辞典)
- Predict: prae "before" + dicere "to say"  
「予言=予め言う」にそっくり!



### 推定

- ①周囲の状況や情報に基づいて、おしはかって決めること。推測決定すること。また、そのようにして得たもの。②法律で、明瞭でない法律関係または事実関係について、否定する反証が成り立つまで、それを正当なものとして扱うこと。(明鏡国語辞典)
- (1)はっきりとはわからないことをいろいろな根拠をもとに、あれこれ考えて決めること。(2)[法] 明瞭でない法律関係・事実関係について一応の判断を下すこと。(3)[数] 統計で、ある母集団から取り出された標本をもとにその母集団の平均・分散などを算出すること。(4)文法 ... (大辞林)

## Forecast, predict, and estimate



## 推測

- ある事柄をもとにして推量すること。(デジタル大辞泉)

## と、まあ、言ったが

- 我々が(理工学的に)できることは、

穴埋め

です。

## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- この講義について

## 例1

一期○会  
三○一体  
無○乾燥  
○面楚歌  
八面○臂  
無我○中  
我○引水  
支離○裂  
単○直入

- 解けますか？
- コンピュータに解かせるには何が必要ですか？

そうです、「答え」が必要です。 なぜなら、規則性がない、から  
人間だって、覚えていなければ答えられません。  
ましてや、コンピュータも覚えなければ答えられません。  
しかし、覚えれば答えられる

## 例2

1	3	8	11	5	5	1
2	1	10	11	5	2	0
4	3	1	4	8	6	1
8	7	4	11	5	9	1
??	6	3	??	9	5	??
32	6	4	10	6	2	0
64	10	5	15	7	10	1
128	8	2	10	5	8	1

違いは、何でしょうか？

### 例3

星新一「ぼろ家の住人」より

- 以下、空欄に「は」か「が」を入れて、文法的に正しい日本語文にしてください。

  - それでまた金をむだ使いし、あとにはさらに大きなむなしさ（ ）残る。
  - 現実形に形となって残るの（ ）、ふえてゆく借金ばかり。
  - 世の中（ ）太平ムードで好景気というのに、おれだけ（ ）例外。
  - 番組にのせる、なにかいい題材（ ）ないものかと考えながら。
  - ごみごみと、古くきたない家々（ ）密集している地域だった。
  - うむ、この経過（ ）いいテーマかもしれぬ。
  - 都市（ ）再開発されてゆくの、具体的にとらえるのだ。
  - 「それ（ ）ありがたい。あわれであればあるほど、びったりです。で、それどこですか。」
  - このへんの建物（ ）どれもぼろだが、そのなかでも最もぼろで最も小さく、建物というより小屋に近い。
  - ひとりの老人（ ）いた。
  - 同情（ ）視聴者のすることであり、テレビ関係者（ ）まず番組のことを考える。
  - 「生活保護（ ）受けていますか」
  - 「そんなもの（ ）知らん。」
  - 会話をしているうちに、この老人だけで番組（ ）一つできると思った。

### 例3 補足

- 日本語を母語とする人なら、まず、正解する。
- どうしてだろうか？
- 記憶している？ NO! 同じ文を見たこと・聞いたことは、まず、ない
- 規則を知っている？ NO! 次のスライドのような説明ができますか？
- 生まれつき知っている？  母語は、生まれ育った環境に依存する
- では、教わった？  母語教育はあるが、大抵は、かなりできるようになってから行われる
- つまり、
  - 生後、自力で学習した
  - 丸暗記ではない
  - 学習結果(規則)を口頭で表現することはできない。

<http://techieminx.com/listen-no-one-can-tell-you-who-is-looking-at-your-twitter-profile.html>



### 例3の説明

文1. それでまた金をむだ使いし、あとにはさらに大きなむなしさ（ ）残る。

この文は出来事を表す「現象文」である。「残る」は自動詞である。「現象文」の中では、主語に「が」をつけるのが普通である。そして、「あとには」の中に取立ての「は」が入っているの、主語に対してもう一つの取立ての「は」が入りにくい。これにより、この文には「が」しか使えない。

文2. 現実形に形となって残るの（ ）、ふえてゆく借金ばかり。

この文は名詞述語の「判断文」である。文末に【である】が省略されているが、「判断文」という性質に変わりはない。「判断文」の主語に「は」をつけるのが普通である。

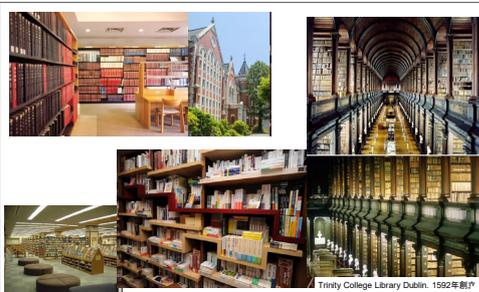
説明できなくても正解できますよね？

ある日本語研究・教育用テキストから

### 例3. 本題に戻ろう



- コンピュータでできるか？
- もし、人間が行うように、生後聞いた母語をすべて与えたらできるかもしれない。
  - なぜなら人間は皆そうして学習しているから
    - ・(脱線)「生まれる前から知識を持っている」と主張する人と「まったくの白紙から学習する」と主張する人がいる
- つまり、



Trinity College Library Dublin. 1992年制作

英語ですが

大量の「日本語文」があれば

それでまた金をむだ使いし、あとにはさらに大きなむなしさ(が)残る。

### つまり

- 人間であれば、  
データ + 学習(の結果) ⇒ 穴埋め
- コンピュータであれば、  
データ + アルゴリズム ⇒ 穴埋め

ということを期待してよいだろうか？



[http://hararie-japan-tokyo-tokyo.com/japanese\\_alphabet/japanese-study/various-expressions-of-yes/](http://hararie-japan-tokyo-tokyo.com/japanese_alphabet/japanese-study/various-expressions-of-yes/)

## つまり

人間の学習

• 人間であれば、  
データ + 学習(の結果) ⇒ 穴埋め

• コンピュータであれば、  
データ + アルゴリズム ⇒ 穴埋め

コンピュータの学習  
すなわち  
機械学習

## 整理しよう

穴埋め1	穴埋め2-1	穴埋め2-2	穴埋め2-3	穴埋め3
一期○会	1	6 9 15	8 8 1	1. それでまた金をむだ使
三〇一体	2	5 1 6	6 7 1	2. 現実に形となって残る
無○乾燥	4	5 2 7	4 7 1	3. 世の中( ) 本平ムー
○面楚歌	8	4 10 14	1 3 0	4. 番組にのせる、なにか
八面○臂	??	6 10 ??	8 6 ??	5. ごみごみと、古くきた
無我○中	32	7 6 13	10 1 1	6. うむ、この経過( )
我○引水	64	10 7 17	3 7 1	7. 都市( ) 再開される
支離○裂	128	10 9 19	6 6 1	8. 「それ( ) ありがた
単○直入				9. このへんの建物( )
同一	一次元系列	連続関数	不連続関数	10. ひとりの老人( )
				11. 同情( ) 視聴者の
				12. 「生活保護( ) 受
				13. 「そんなもの( )
				14. 会話をしているうち

## 整理しよう

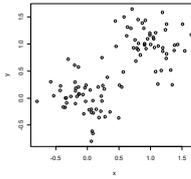
穴埋め1	穴埋め2-1	穴埋め2-2	穴埋め2-3	穴埋め3
一期○会	1	6 9 15	8 8 1	 <p>1. それでまた金をむだ使 2. 現実に形となって残る 3. 世の中( ) 本平ムー 4. 番組にのせる、なにか 5. ごみごみと、古くきた</p>
三〇一体	2	5 1 6	6 7 1	
無○乾燥	4	5 2 7	4 7 1	
○面楚歌	8	4 10 14	1 3 0	
八面○臂	??	6 10 ??	8 6 ??	
無我○中	32	7 6 13	10 1 1	
我○引水	64	10 7 17	3 7 1	
支離○裂	128	10 9 19	6 6 1	
同一	一次元系列	連続関数	不連続関数	
			文字列	

## 目次

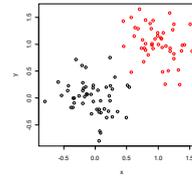
- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- この講義について

## 穴埋めではないが、類似

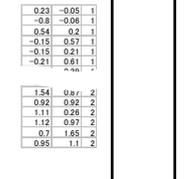
0.23	-0.05
-0.8	-0.08
0.54	0.2
-0.15	0.27
-0.15	0.21
-0.21	0.61
	0.61



1.54	0.67	2
0.92	0.92	2
1.11	0.26	2
1.52	0.97	2
0.7	1.65	2
0.95	1.1	2



0.23	-0.05	1
-0.8	-0.08	1
0.54	0.2	1
-0.15	0.27	1
-0.15	0.21	1
-0.21	0.61	1
	0.61	1

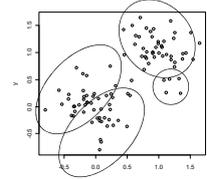


## クラスタリングという

クラスタリングとは  
「いくつか」の「かたまり」に分けること

課題  
「かたまり」?  
「いくつ」?

応用はあるの?  
はい!

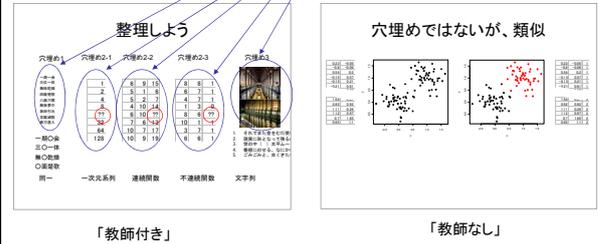


こんな風に考えては、悪いのか?

## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- この講義について

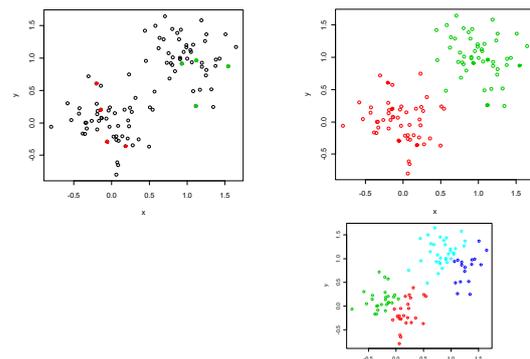
## 教師



## 教師なしデータの有用性

- 教師データ作成コストは高い
- 例えば 次のアプリを考えよう(実問題)
  - バイオ系の論文で、化合物・タンパク質間の相互作用について記述した論文を選び出したい。
  - この「相互作用」の表現をリストアップすることができない(googleの検索(つまり全文検索)では見つけられない)。
  - 専門家が論文を読んで、その表現を探し出す必要がある。
- ならば、その少数例(教師データ)と、多数の文から、仮の(間違っているかもしれない)教師データを作れないか？
  - 実は、この問題は、この方法では難しい。しかし、雰囲気は分ろう

## 半教師付き学習



## もう一度、整理

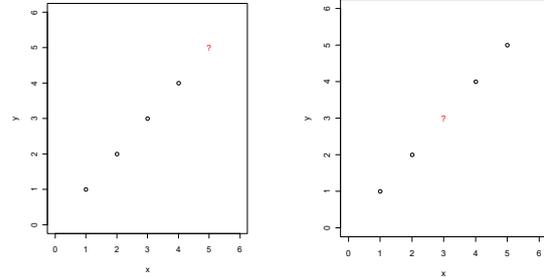


## 目次

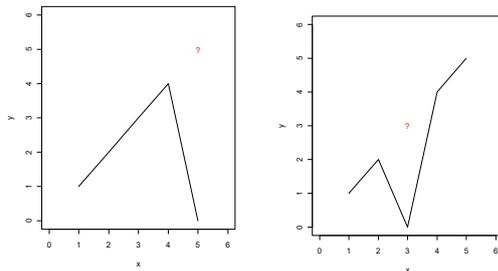
- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- この講義について

## 本日の第二部

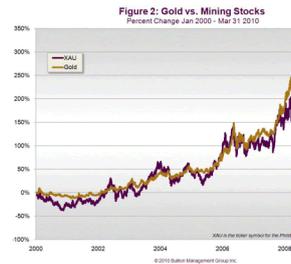
## 予測と推測・推定



## 予測と推定・推測



<http://heavenawaits.wordpress.com/god-man-and-stock-market-wave-theories/>

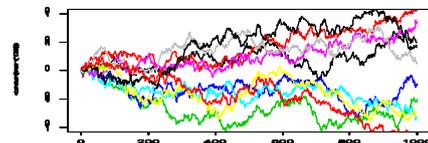


<http://www.safehaven.com/article/17497/why-bullion-is-outperforming-mining-stocks>

## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
    - Random walk - ランダムな時系列
    - ベキ分布とBlack swan
    - 機械学習の位置づけ
- この講義について

### Random walk の意外な性質



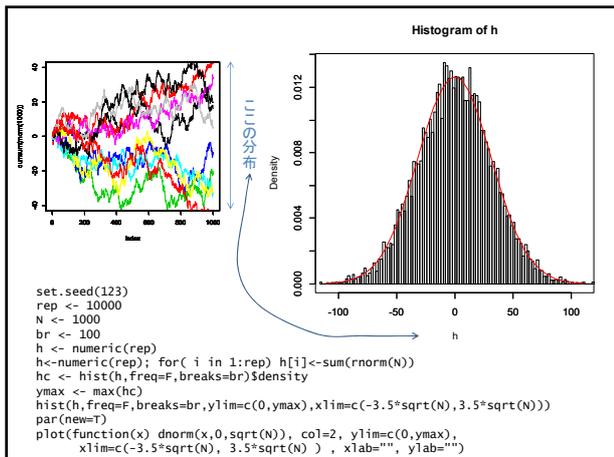
ランダムウォーク  $S$  が  $2n$  歩後に  $2l$  ( $-n \leq l \leq n$  とする) の地点にいる確率は

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

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

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

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



## 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

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

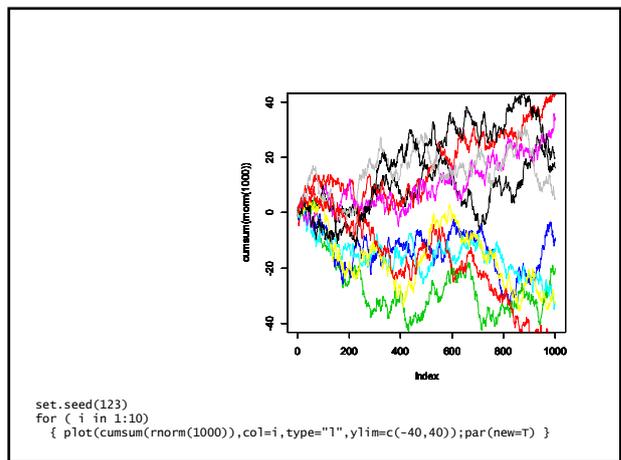
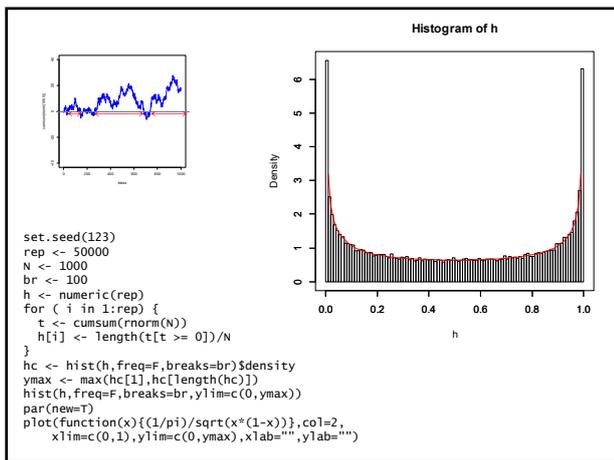
である

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  $\leq a$ )

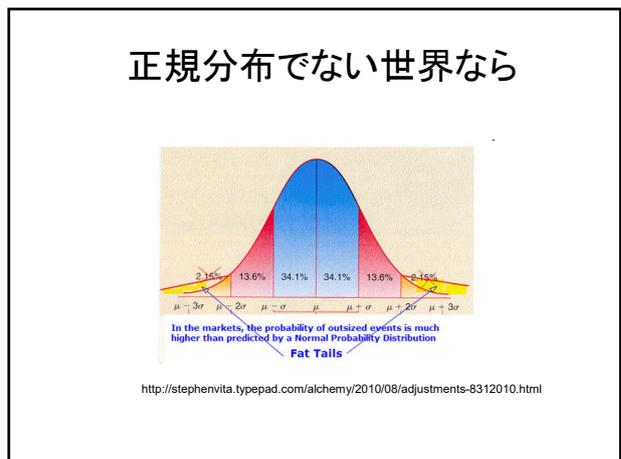
$$= \sum_{k=0}^{\lfloor a n \rfloor} P(n, k) \approx \sum_{0 \leq k \leq a n} \frac{1}{\pi \sqrt{k(n-k)}} = \frac{1}{\pi} \sum_{0 \leq k \leq a n} \frac{\frac{1}{n}}{\sqrt{\frac{k}{n} \left(1 - \frac{k}{n}\right)}} \approx \frac{1}{\pi} \int_0^a \frac{dx}{\sqrt{x(1-x)}} = \frac{2}{\pi} \arcsin a^{\frac{1}{2}}$$

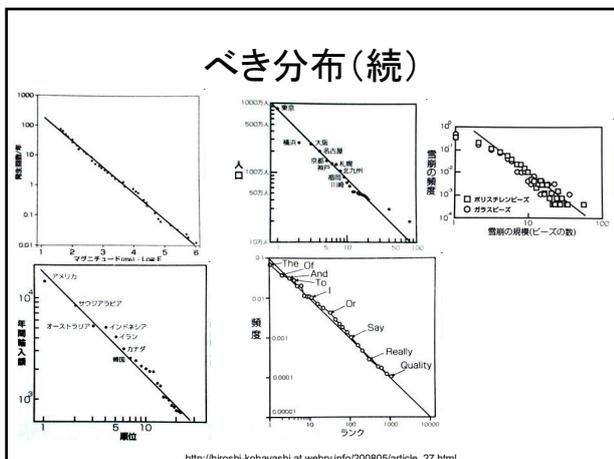
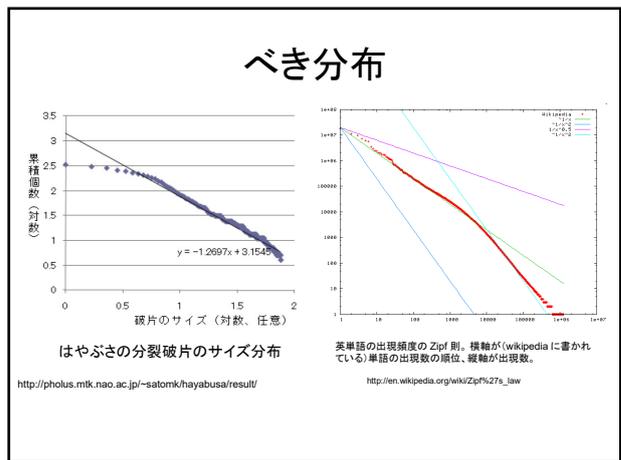
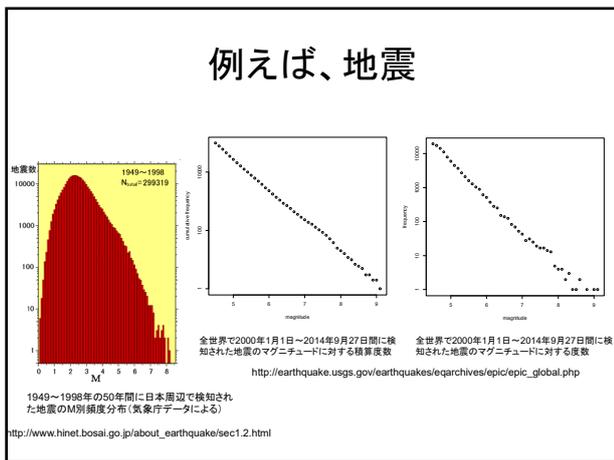
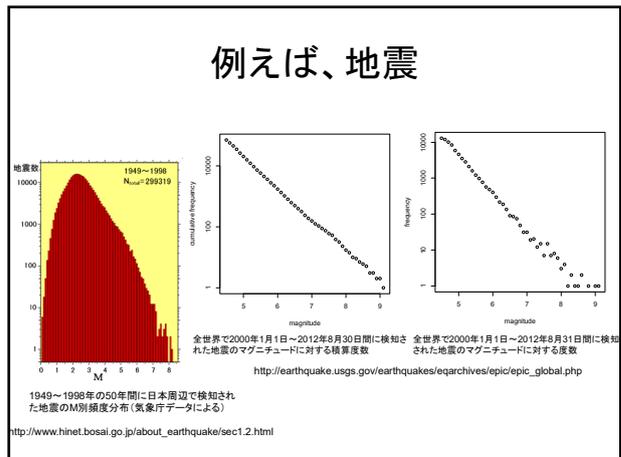
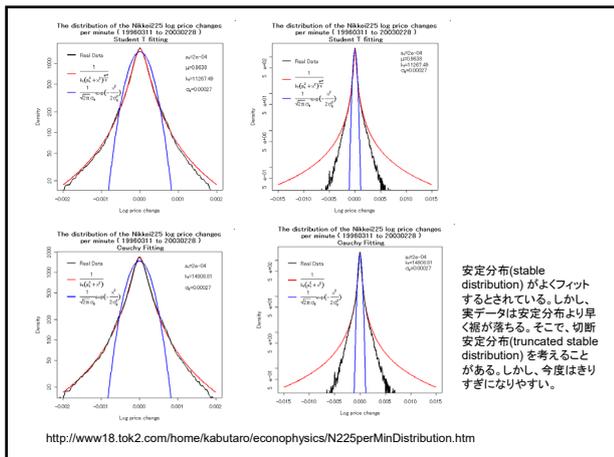
<http://elis.sigmath.es.osaka-u.ac.jp/~nagahata/20070816/arcsin.pdf>



## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - べき分布とBlack swan
  - 機械学習の位置づけ
- この講義について





### 現実のデータ

- 正規分布に従わないものがある
  - 冪分布に従うと、fat tail である。
  - その結果、予測誤りの影響が大きくなる
- しかも、現実にはデータ量が少ない
  - 絶対量が少ない場合
  - 相対量が少ない場合

世の中ビッグデータだと騒いでいるのに？

## Swan and Black Swan



## Swan and Black Swan

- "Black Swan" はTalebの極めて有名な著書
  - 最近は、"Black Swan" とgoogleで引くと、別のものが大量に出てきて困ります。
    - 一昨々年よりは、よい。昨年と比べれば、同じ(かな)。
- Swanは白い鳥だと誰もが信じていた。Black Swan が発見されるまでは。
  - 「これはバブルではない、わが国経済の実力である」と誰もが信じていた。バブルが崩壊するまでは。

## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - べき分布とBlack swan
  - 機械学習の位置づけ
- この講義について

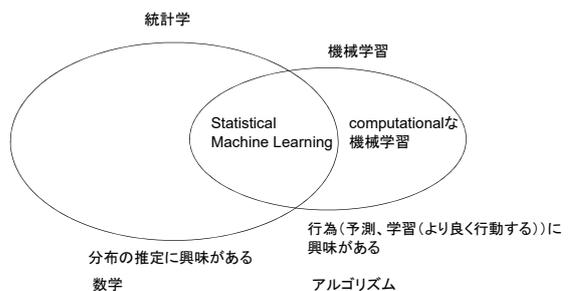
## 統計学との違い

- 統計では、分布の推定や(ほぼ同じことであるが)パラメータの推定を目指す。
- 機械学習では、「論理・ロジック」と組み合わせた知識表現(記号表現)も推定対象に入る。
- 統計で目的とする分布は、数式で書ける分布が多い。
  - モデルを単純にし、理論的に正確に。ノンパラメトリックという手法はある
  - もっとも、Bayes系のモデルはかなり複雑
- 機械学習では、数式で書けないような分布を対象とする
  - モデルは複雑(かどうかは分からないが)に、予測結果は正確に。
  - 機械学習でも「分布の推定」ということを行うし、研究ではその評価式がたくさん出てくるのですが、実用上は、その推定精度は大したことはない。結果としての予測精度が重要。
  - 分布の推定精度が測れるほどのデータ量が、実は、ないのが原因。
    - モデルパラメータが多すぎるのが原因。
- 融合が進んでいますが、相変わらず、違いがある

## PAC学習可能

- Probably Approximately Correct 学習可能
  - 大抵の場合、近似的に、正しく学習できる
- 大抵の場合:
  - (e.g.) 100回学習すると、そのうちの95回
- 近似的に正しい:
  - (e.g.) 真の解との誤差は 5% 以内で

## 統計学と機械学習



## 目次

- 第一部
  - 予測と推定
  - 機械学習とは、(アルゴリズムを用いた)穴埋めである
  - クラスタリング
  - 教師付、半教師付、教師なし
- 第二部
  - 予測
  - Random walk - ランダムな時系列
  - ベキ分布とBlack swan
  - 機械学習の位置づけ
- [この講義について](#)

## この講義の目的

- 知的な情報処理を実現する技術の一つである「機械学習・データマイニング」技術の基礎を知る
- Rという統計パッケージに慣れておこう
  - きっと卒論で役に立つよ

## 講義の進め方

- 講義中心、しかし、手を動かす。
- Rを使用します(ほどほどに)

## Rとは

- (元は)統計計算とその結果表示のための言語・環境
  - いろいろな統計手法が、パッケージ(オープンソース)とされ、簡単に組み込める
- ところが、今では、様々な機械学習手法も入っている。
  - そのため、機械学習の手法の学習には使えない。
  - しかし、使い方の学習にはもってこい!
- フリーソフト

## 評価他

- レポート(3回ほど)と試験(またはレポート(4回ほど))に基づく
  - 講義の進行状況に依存して決める。
  - レポート採点は、考察重視
  - 出席はとらない予定
  - ただし、たいていは、簡単な即レポで代替するので、ご注意ください
- 講義資料は、[櫻井研究室 website](http://www.sakurai.comp.ac.keio.ac.jp/) に掲載予定 (google で 櫻井研究室 で検索すればよい)  
<http://www.sakurai.comp.ac.keio.ac.jp/>

## 予定

1	9月26日	火	予測と推測と機械学習
2	10月3日	火	R 超入門
3	10月10日	火	最近傍法 - 近さの利用
4	10月17日	火	ナイーブなベイズ法
5	10月24日	火	ナイーブベイズと自然言語処理
6	10月31日	火	決定木 - 生成と剪定
7	11月7日	火	オッカムの剃刀と過学習
8	11月14日	火	誤差の話と演習
9	11月28日	火	ニューラルネットワーク - 夢と限界と広がり
10	12月5日	火	実用的なニューラルネットワーク
11	12月12日	火	SVM - 強力な分類法
12	12月19日	火	クラスタリング - グループ分け
13	12月26日	火	モデル選択・Deep Learning
14	1月16日	火	最終課題