

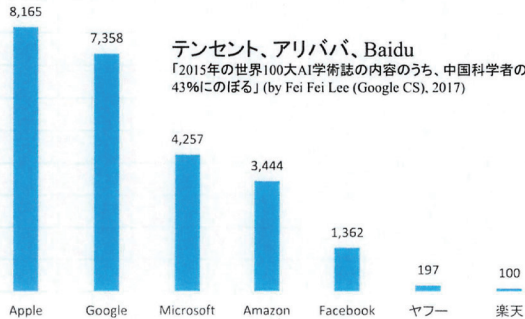
ビッグデータ、AI時代に必要とされる 統計的推論法の習得に必要な数学教育

樋口知之（情報・システム研究機構 統計数理研究所）

1/40

投資額の格差(USのAI×データ企業との比較)

設備投資額（億円/年）



テンセント、アリババ、Baidu
「2015年の世界100大AI学術誌の内容のうち、中国科学者の貢献が43%にのぼる」(by Fei Fei Lee (Google CS), 2017)

各社のFY2013 IR資料：有形固定資産の取得、100円/\$で換算

3/40

個人的意見

以下の講演内容は、あくまでも個人的意見であり、統計数理研究所長としての見解でないことをご承知ください。

また、以下の意見に大賛成です。

- データリテラシー教育は初等教育から行う
- 初歩的な統計分析はすべての学部学生に習得させる
- 成人の学び直しとして統計学は有用

- ✓ 昨今、国内外で話題となっている『 p 値』についての議論は本日はしません。
- ✓ 習得内容が学部後期あるいは大学院レベルになっています。

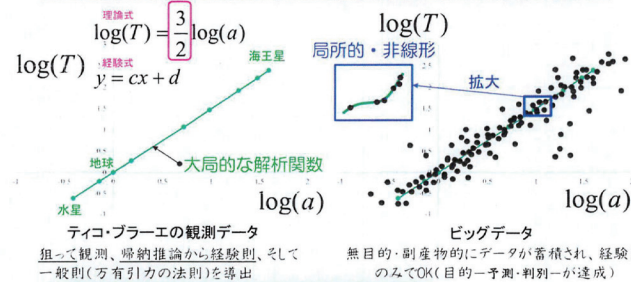
2/40

概括：帰納法とデータ今昔

ケプラーの第三法則

師匠であったティコ・ブラーエの観測記録から推定し定式化

惑星の公転周期 T の2乗は、楕円軌道の半長軸 a の3乗に比例する



ピンク圏で示す良質な空間を見つけること

- 1) 説明変数の選択
- 2) 線形性
- 3) 少ないパラメータ

4/40

急がば回れ： まずは画像へ変換

2017年

DeepVariant: ゲノムのmutation caller (変異検出器)

Creating a universal SNP and small indel variant caller with deep neural networks
<https://www.biorxiv.org/content/early/2016/12/21/092890>

リファレンス配列にアライメントされたリード配列の積み重ね (pileup) を画像データとして入力し学習させている

結果、非常に高精度の変異 (多様性) 検出ができるようになっており、生殖細胞系列における変異 (germline variant) の検出のコンペティションにおいて賞を受けている。
 pileupの画像は、通常の研究現場では、何らかの方法で見つけられた変異の候補を人間の専用のviewerで、本物かどうか確認・検討する際に目にするもの
 その画像そのものを変異検出に用いた点がすごいところ

プラント内で既に蓄積されている様々な数値 (温度、圧力、流量等) を解析し、現在のデータと比較することによって、プラントの異常の検知や未来の変動の予測を行うシステムを開発
 エッジコンピューティング、IoTでもこのような流れか?

9/40

この20年間の内挿手法の劇的な性能向上

線形・非ガウス→スパースモデリング

変数 (要素) 間の関係 最適化関数のクラス Tibshirani, R. (1996)

Candès, E.J. and Tao, T. (2005)

非線形・ガウス→深層学習

Hinton, G.E. (2006)

線形・非ガウスモデルと非線形・ガウスモデルがマシンで得られる時代になった!

モデリングの技

10/40

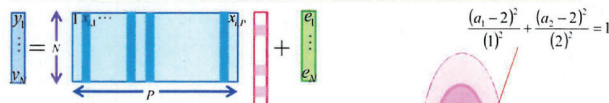
スパース (疎性を利用した) 最適化 1:

LASSO: Least Absolute Shrinkage and Selection Operator

例: 多変量回帰

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \dots + a_p x_{i,p} + e_i \quad (i=1, \dots, N)$$

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}$$



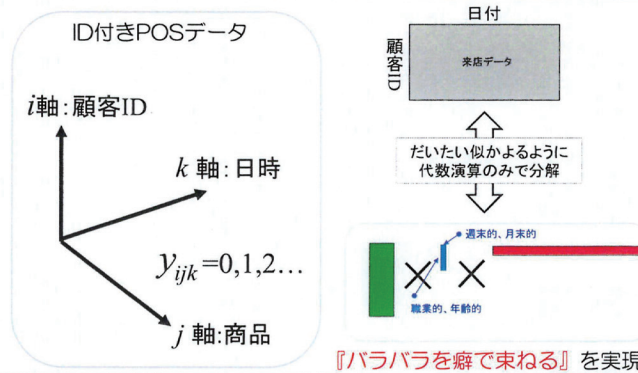
新NP問題 ($N \ll p$)

$$\mathbf{a}^* = \min_{\mathbf{a}} \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2 + \lambda \|\mathbf{a}\| \right\}$$

$$\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{X}\mathbf{a}\|^2 \text{ subj. to } \|\mathbf{a}\| \leq \alpha$$

11/40

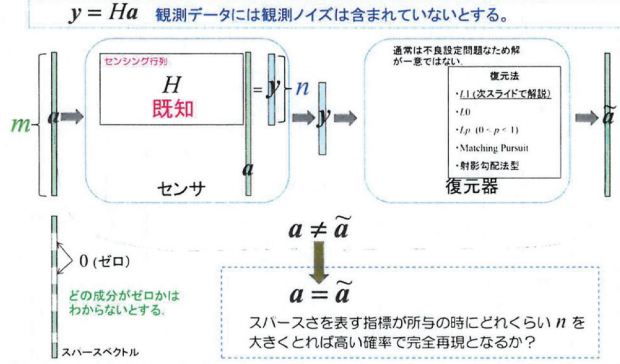
スパース最適化 2: 行列、テンソル分解



12/40

スパース最適化 3 : 圧縮センシング: Compressed Sensing

E.J. Candes and T. Tao (IEEE IT, 2006), D.L. Donoho (IEEE IT, 2006) ※ 2003 年頃から研究が始まる



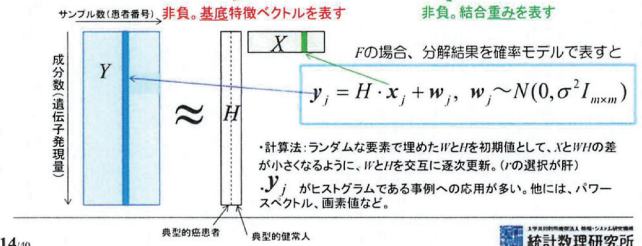
13.40

スパース最適化 4 : NMF: Non-Negative Matrix Factorization

D. D. Lee and H. S. Seung (Nature, 1999) (NIPS, 2000)

$Y_{m \times n}$: 非負値データ行列。 m 次元特徴ベクトルを n サンプル分横に並べた行列

$$\min. \|Y - H_{m \times r} \cdot X_{r \times n}\|_{F \text{ or } ID}^2$$

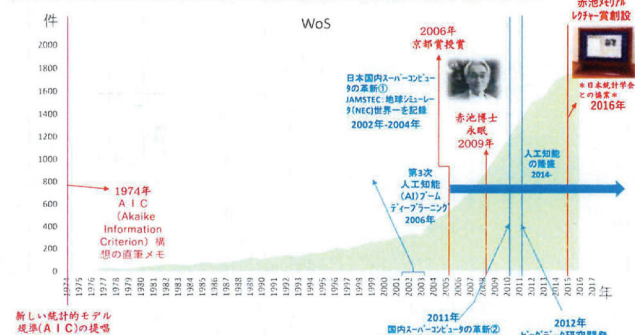


14.40

統計数理の特質

評価されるまでに時間がかかるが、一度評価されると長期間に渡り利用される

Year by Year Citation counts of "NEW LOOK AT STATISTICAL-MODEL IDENTIFICATION" (IEEEac, 1974) の被引用件数 (Data from Web of Science on 21 Apr. 2017) グラフに見る故赤池博士の業績と統計研究を取り巻く世情の流れ



15.40

深層学習の強み: 人工知能の「度量衡」

長さを計る基準を「度」、体積は「量」、重さは「衡」と定め、...

2018年 人工知能学会誌 特集「AIとデータ」に招待論文を寄稿

- ① パラメータ学習がバックプロパゲーションと確率的勾配降下法で統一
- ② 計算プラットフォームが汎化
 - a. 専用計算機の整備(比較的廉価)
 - ・ GPGPU クラスタ
 - b. オープン・無償な開発プラットフォームの整備
 - ・ オープンソース: TensorFlow, Caffe, Keras, Chainer
- ③ オープンかつリアルタイムな成果公開
 - ✓ 査読プロセスをとらず、プレプリサーバ(arXiv)にアップ

理論的には第二次AIブームから大きな発展はこれから

16.40

Back Propagation

コスト関数

$$E = \sum_n E_n = \sum_n \sum_j (y_j^{[n]} - x_j^{(last)})^2$$
11番目のデータの最後の層における / 番目の素子の出力

線形結合

$$a_j = \sum_i w_{i,j} x_i^{(-)} + b_j$$

活性化関数

$$x_j = f(a_j) \quad \frac{\partial a_j}{\partial w_{i,j}} = x_i^{(-)}$$

微分の連鎖率

$$\frac{\partial E_n}{\partial w_{i,j}} = \frac{\partial E_n}{\partial a_j} \frac{\partial a_j}{\partial w_{i,j}} = x_i^{(-)} \cdot \delta_j \quad \delta_j = \frac{\partial E_n}{\partial a_j}$$

$$a_k^{(+)} = \sum_m w_{m,k}^{(+)} f(a_m) + b_k^{(+)}$$

$$x_k^{(+)} = f(a_k^{(+)})$$

$$\delta_j = \sum_k \frac{\partial E_n}{\partial a_k^{(+)}} \frac{\partial a_k^{(+)}}{\partial a_j} = \sum_k \delta_k^{(+)} \cdot w_{j,k}^{(+)} f'(a_j) = f'(a_j) \sum_k \delta_k^{(+)} \cdot w_{j,k}^{(+)}$$
j→kしか残らない。灰色矢印の部分はゼロ。

17/40

4 dimensional Variational method

Optimize only for initial state vector

$$p(x_0^{[k]} | y_{1:T}) \propto p(y_{1:T} | x_0^{[k]}) p(x_0^{[k]} | y_{\cdot,0})$$
Dimension is huge, so we need an elegant fast calculation scheme

$\frac{\partial \log p(y_{1:T} | x_0)}{\partial x_0}$ Calculation method for efficiently deriving differential vector
 $x_0 = x_0^{[k]}$ x_0' is a transposition of x_0

- Chain law of differential

$$\frac{\partial \log p(y_{1:T} | x_0)}{\partial x_0} = \frac{\partial \log p(y_{1:T} | x_0)}{\partial x_T} \frac{\partial x_T}{\partial x_{T-1}} \frac{\partial x_{T-1}}{\partial x_{T-2}} \dots \frac{\partial x_1}{\partial x_0}$$
System model $x_T = f(x_{T-1})$
- Decomposition formula of likelihood function

$$\log p(y_{1:T} | x_0) = \sum_{t=1}^T \log p(y_t | y_{1:t-1}, x_0) = \sum_{t=1}^T J_t = J_{1:T} \quad J_t = \log p(y_t | y_{1:t-1}, x_0)$$

$$J_1 = \log p(y_1 | x_0)$$

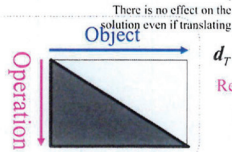
18/40

Top-down automatic differentiation → Ajoint method

Back propagation in NN is a special type of top-down automatic differentiation.

$$\frac{\partial J_{1:T}}{\partial x_0'} = \frac{\partial J_0}{\partial x_0'} + \left(\frac{\partial J_1}{\partial x_1'} + \left(\dots \left(\frac{\partial J_{T-2}}{\partial x_{T-2}'} + \left(\frac{\partial J_{T-1}}{\partial x_{T-1}'} + \frac{\partial J_T}{\partial x_T'} \right) \frac{\partial x_{T-1}}{\partial x_{T-2}'} \right) \frac{\partial x_2}{\partial x_{T-1}'} \right) \frac{\partial x_1}{\partial x_{T-2}'} \right) \frac{\partial x_0}{\partial x_{T-2}'}$$

$$d_T = \frac{\partial J_T}{\partial x_T'}, \quad d_{T-1} = \frac{\partial J_{T-1}}{\partial x_{T-1}'} + d_T \frac{\partial x_T}{\partial x_{T-1}'}, \quad d_{T-2} = \frac{\partial J_{T-2}}{\partial x_{T-2}'} + d_{T-1} \frac{\partial x_{T-1}}{\partial x_{T-2}'}$$



There is no effect on the solution even if translating

Recurrence formula going back in time (Recurrence backward formula)

$$d_0 = \frac{\partial J_0}{\partial x_0'} + d_1 \frac{\partial x_1}{\partial x_0'} = \frac{\partial J_{1:T}}{\partial x_0'} \quad J_0 = \log p(y_0 | x_0) = 0$$

Updated by descent method $x_0^{[k+1]} = x_0^{[k]} + s \cdot d_0^{[k+1]} \quad (k=0,1,\dots)$

19/40

確率的勾配降下法 (SGD: Stochastic Gradient Decent)

目的(最適化)関数 $N \leftarrow$ 超ビッグ!

$$L(w) = \sum_{i=1}^N \ell(z_i, w) + \psi(w)$$
データに依存しないペナルティ項 あとで考えれば良い

$$z_i \equiv (y_i, x_i)$$
ラベル 説明変数ベクトル

パラメータの次元は、億以上
 尤度関数部分に相当
 事前分布部分に相当

最急降下法

$$\ell(z_i, w) := Q_i(w) \quad w^{(k)} = w^{(k-1)} - \eta \nabla \left(\sum_{i=1}^N Q_i(w^{(k)}) \right)$$

SGD $w^{(k)} = w^{(k-1)} - \eta_k \nabla Q_{i^*}(w^{(k-1)})$
 モーメンタム法 i^* は、1~Nの中からランダムに選択(乱択) ミニバッチ: データ分割法

Nestrov加速 $\tilde{w}^{(k)} = w^{(k-1)} - \eta_k \nabla Q_{i^*}(w^{(k-1)})$

Yurii Nesterov (1983) $w^{(k)} = \tilde{w}^{(k)} + \frac{k-2}{k+1} (\tilde{w}^{(k)} - \tilde{w}^{(k-1)})$

20/40

確率的勾配降下法の改良

機会学習ではオンライン学習と呼ばれる研究分野

確率的分散縮小勾配法 (SVRG)

Outer loop : $w^{[m]}$ を更新
Inner loop : $w^{(k)}$ を更新

i^* は、1~Nの中からランダムに選択 (乱択)

$$w^{(k)} = w^{(k-1)} - \eta_{m,k} \left[\nabla Q_{r^*}(w^{(k-1)}) + \left\{ \nabla Q_{r^*}(w^{[m]}) \right\} + \nabla \left(\sum_{i=1}^N Q_i(w^{[m]}) \right) \right]$$

アイデア自体はそれほど新規味は感じないが、収束速度の証明は大変

二重加速確率的分散縮小勾配法 理論的には世界最速! (Suzuki)

$$z^{(k)} = \tilde{w}^{(k-1)} + \frac{k-2}{k+1} (\tilde{w}^{(k-1)} - \tilde{w}^{(k-2)}) + \frac{k-1}{k+2} (w^{(k-1)} - \tilde{w}^{(k-2)})$$

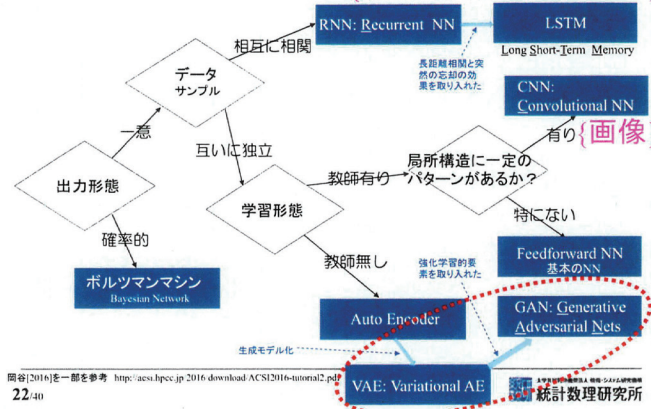
Nestrov加速

$$(\tilde{w}^{(k)}, w^{(k)}) = \text{inner loop}(z^{(k)}, \tilde{w}^{(k-1)})$$

勾配をSVRGと加速法で計算

DNNの分類

{系列データ: 音声、テキスト}



生成モデルと識別モデル (関数) Generative Model vs. Discriminative Model

生成モデル (広義)

$$p(y | \theta, z)$$

データ パラメータ 説明変数(アスペクト)

生成モデルからベイズの定理を経由して条件付き確率を計算

クラス分類問題

$$p(C = C_i | y, \theta) = \frac{p(y, C = C_i | \theta)}{\sum_j p(y, C = C_j | \theta)}$$

例: $C=1$ or 0

生成モデル (狭義)

$$p(y, C | \theta) \text{ 同時分布}$$

データ クラス $p(y | C, \theta_2) \cdot p(C | \theta_1)$

識別モデル

$$p(C | y, \theta) \text{ 条件付き分布}$$

識別関数: y を C に写像する関数

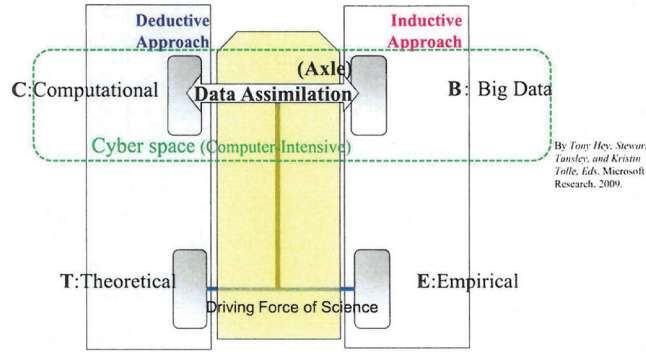
統計学の強みは生成モデルの取り扱いに慣れていること

シミュレーションと機械学習の融合

Smart Simulation

- ① Emulation
 - ✓ Sparse modeling
- ② Neo experimental design
 - ✓ Bayesian optimization
- ③ Automation of making a generative model
 - ✓ Reversal formula by Bayese Theorem
 - ✓ GDAE, VAE, GAN

Science Paradigms: Four Kind of Methodology of Science



25/40

① Emulation 気候変動の確率地図 (上野玄太)

文科省創生プログラム「気候変動予測データの統計学的解析手法の開発」

複数の全球気候モデルのシミュレーション結果を用い、気候変動を確率的に評価する手法を開発し、21世紀後半の気温上昇の確率地図東アジア及び日本における気温の将来変化の確率地図を初めて作成した。

観測値 シミュレーション

スパース回帰モデル

$$J(\beta_0, \beta_1, \dots, \beta_N) = \sum_{i=1}^T \left(y_i - \beta_0 - \sum_{n=1}^N \beta_n x_i^{(n)} \right)^2 + \gamma \left[(1-\alpha) \sum_{n=0}^N \beta_n^2 + \alpha \sum_{n=0}^N |\beta_n| \right]$$

(例) 東京の気温

- 白: 観測値
- 灰: シミュレーション
- 緑: アンサンブル平均
- 赤: 推定した確率分布

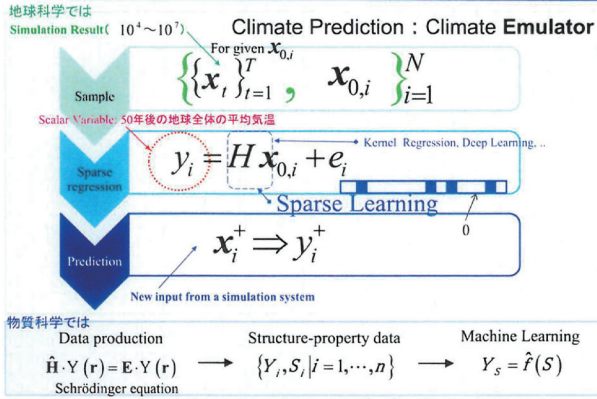
2°C以上昇温する確率

東アジア及び日本の多くの地域で2度以上月平均気温が上昇する確率が70~80%以上。

26/40

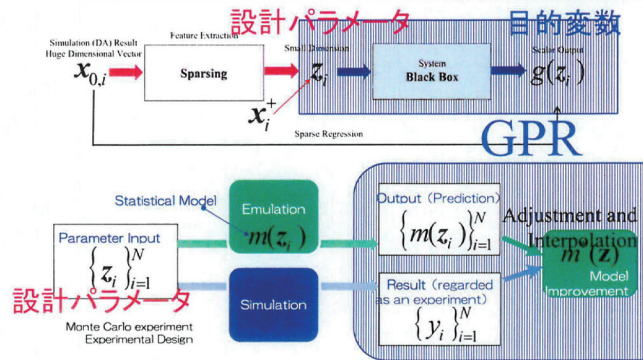
26/40

① Emulation エミュレーション: スパースモデリング



27/40

Emulator and Emulation ② Bayesian optimization Emulate an output of simulation given parameters



28/40

ベイズの反転公式

③ a. Reversal formula by Bayese Theorem

x : 興味のある対象

順解析: シミュレーション等

y : データ

ベイズの反転公式

$$p(x|y) = \frac{p(y|x)p(x)}{\sum p(y|x)p(x)}$$

逆解析

ベイズの定理。等式の右側と左側で、赤と青で示した変数部分の確率との相対関係が反転していることがわかる。この事実により、ベイズの反転公式と呼ばれる。

29/40

統計数理研究所

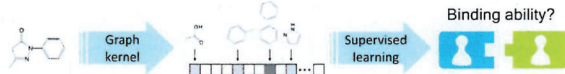
Data Science Center for Creative Design and Manufacturing
established in July, 2017

Prof. Yoshida 

Virtual screening, QSAR modeling: P(Y|G) Forward Problem

Quantitative Structure-Activity (Affinity) Relationship

Develop statistical models to predict biochemical or physicochemical activities Y of an input chemical structure G



Chemical design, Inverse-QSAR: P(G|Y=y) Inverse Problem

Generate novel chemical structures G achieving desired activities $Y=y$

Preimage reconstruction of the graph kernel using a MCMC algorithm



31/40

(Thanks to Prof. Yoshida; 吉田@統計数理研から借用、改変)

統計数理研究所

非効率（現実には機能しない）探索法

構造 x

$f(x)$

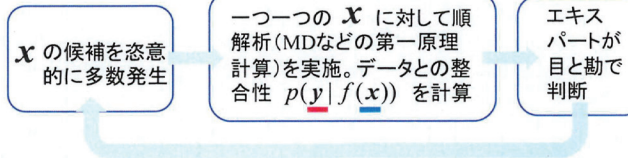


物理あるいは化学的指数や係数、特性の計算値

機能発現 y

その(実験値あるいは経験値)データ

膨大な数のモンテカルロ計算



30/40

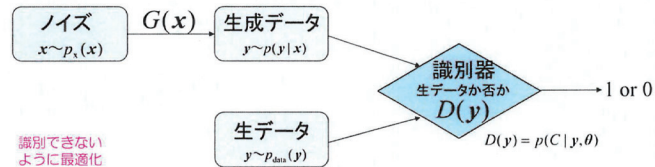
統計数理研究所

GAN (敵対的生成ネットワーク) の基本アルゴリズム

Generative Adversarial Nets; Goodfellow, Ian et al. Advances in Neural Information Processing Systems 27. 2014年

$D(y)$: y が学習データである確率を与える写像(関数)。sigmoid を利用

エミュレータ $G(x)$: ノイズあるいは潜在変数 x からデータ空間 y への写像(関数)



識別できないように最適化

識別できるように最適化

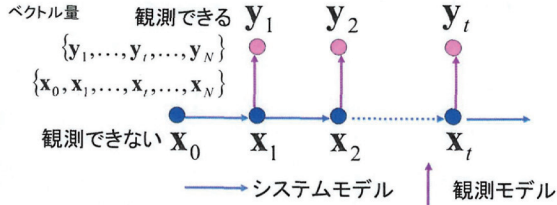
$$\min_G \max_D E_{y \sim p_{data}(y)} [\ln D(y)] + E_{x \sim p_x(x)} [\ln(1 - D(G(x)))]$$

Discriminative NetworkはGenerative Networkの学習で使うときは訓練しない

32/40

統計数理研究所

Chain Structure Graphical Model



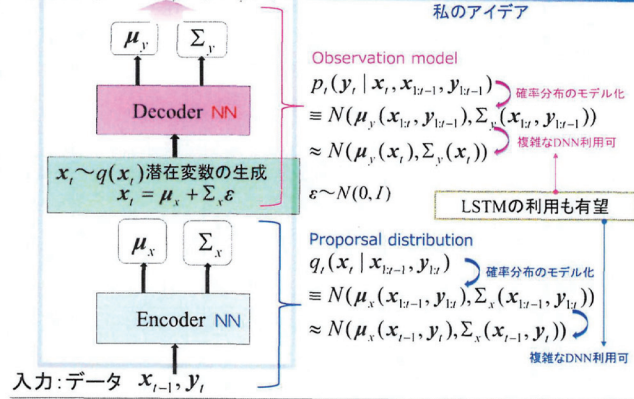
$$p(y_{1:T-1}, y_T, x_{1:T}) = \prod_{t=1}^T p(y_t | y_{1:t-1}, x_{1:t}) \cdot p(x_t | y_{1:t-1}, x_{1:t-1})$$

一般型

$$\Rightarrow \prod_{t=1}^T p(y_t | x_t) \cdot p(x_t | x_{t-1})$$

33/40

VAEと粒子フィルタの融合



34/40

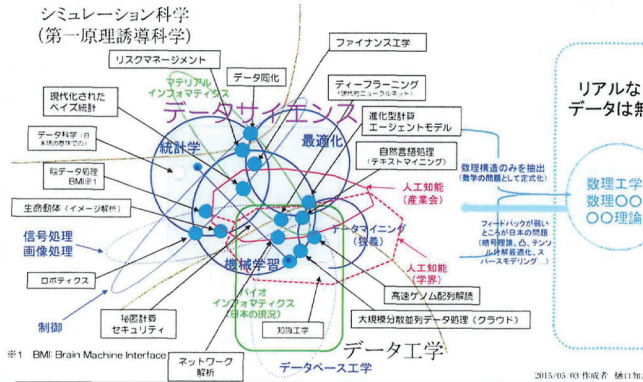
1980年代数理・情報技術の再興

- ニューラルネットワーク → 深層学習
- 情報統計力学
 - ✓ イジングモデル → 量子アニーリング
 - ✓ シミュレイテッドアニーリング
- 遺伝的アルゴリズム、遺伝的プログラミング → 超並列計算による大規模離散最適化
- 連続最適化 (降下法) → 確率的勾配降下法、オンライン学習
- 制約付き最適化 → スパースモデリング
- 次元圧縮 → 行列、テンソル分解

最適化

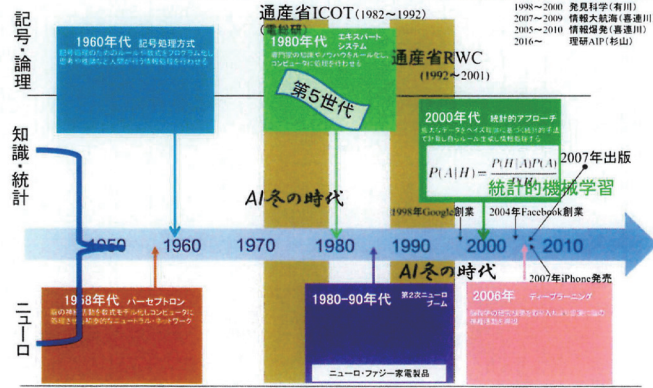
35/40

データに関連した数理技術の俯瞰図



36/40

人工知能研究および研究開発プロジェクトの歴史



* http://blogs.itmedia.co.jp/itsolutionjuku/2015/07/post_105.html を参考として作成
 * 統計研・福水教授のスライド内容を一部借用
 統計数理研究所

再掲： 統計学は対応できているのか？

- 計算量に配慮しているか？ : 計算量が $O(N)$ では役に立たない
- p 次元空間内の分布の表現をどうするのか？
- オンライン学習(推定)法の研究は？
- 統計的モデリングはエキスパートがやるのか？ : 深層学習とスパースモデリングの勃興
- 可視化技術は人が見るため？ : 深層学習(とくにCNN: Convolutional NN)の画像処理における圧倒的性能
- 計算プラットフォームの整備は？
- 目的特化型計算機への実装は？

何が肝か？

薄く幅広く習得するのがよい。引き出しをたくさん持っておくこと。

理由： Webでいくらかでも自学自習できる。実務でもしっかり体得できる。「厳密・唯一」から、「近似・アンサンブル」の世界観を身につけさせる。

- 高次元
- 超大サンプル数 (オンライン)
- 非線形
- 最適化
- シミュレーション
- プログラミング
- ビッグデータをベースに、従来、分断されていた教育内容を整理・融合
- 学部前期教育もそれに向けた数学の基礎とデータリテラシーに

人生の分水嶺



- 基礎的な数学を修得するかどうかで、人生の分水嶺は決まる
- 『仕組み』をつくるものが勝利を得る世界
- そこに資するのは数学とデザインセンス！