

学習数理情報学研究室の紹介

東京大学 大学院情報理工学系研究科
数理情報学専攻
第6研究室

<http://www.ibis.t.u-tokyo.ac.jp/>

2020/5/23

学習数理情報学研究室

- 機械知能の本質を数理で捉える -



教授 山西健司

専門: 情報論的学習理論、データマイニング、異常検知

講義: 機械学習の数理(夏:学部)

数理情報学特別講義I データマイニングによる
異常検知(冬:大学院)



准教授 鈴木大慈

専門: 機械学習, 数理統計学, 確率的最適化

講義: 確率数理工学(夏:学部),



講師 久野遼平 (MIセンター本務、数理兼務)

専門: 経済学、ビッグデータ解析



助教

二反田篤史

専門: 確率的最適化
深層学習



特任助教 木脇太一

専門: ニューラルネット
講義: データマイニング
実践演習

研究室の特徴

- しっかりとした基礎理論に立脚しつつ、現場に生きる学習数理情報学をめざします
- 東大データサイエンス養成講座(DSS)を運営・サポートする研究室です
- 幅広い産業界との連携のもとで、研究を進めます
以下の企業とはすでに協力を開始しています。

Microsoft



hd 東大病院
The University of Tokyo Hospital

三菱商事

NTT

HAKUHODO

LAC
Little eArth Corporation

sas | THE POWER TO KNOW.

YAHOO!
JAPAN

ネットリサーチなら、マクロミル
macromill

VALUES
Consulting & Creation Group

TORAY
Innovation by Chemistry

MIZUHO

みずほ銀行

HONDA
The Power of Dreams

MODEC

山西研究室の概要

コア技術: 情報論的学習理論

目標: ディープナレッジの発見

基礎理論

潜在的ダイナミクス

潜在構造最適化

予兆情報学

Symptomatics

主要応用

医学応用

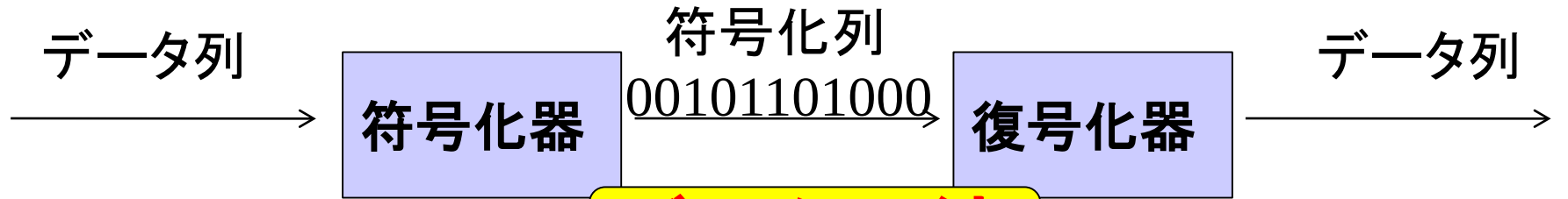
経済応用

マーケット
解析

量子
機械学習

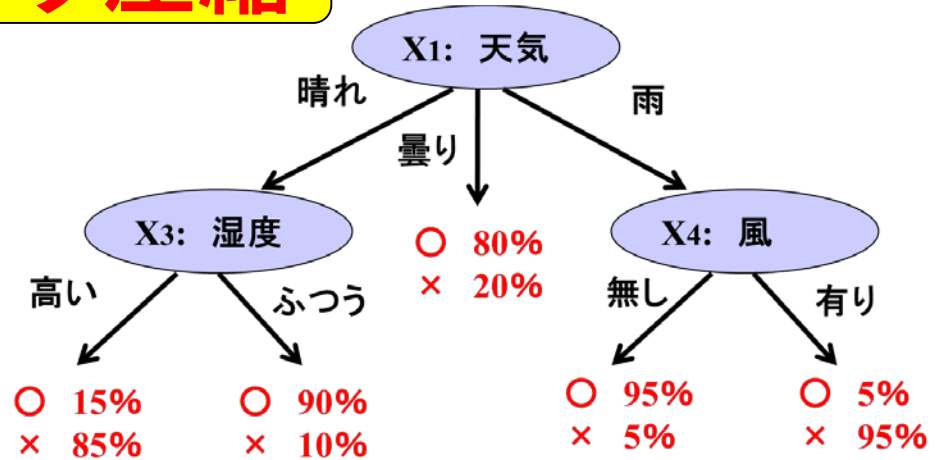
情報論的学習理論とは

学習とは最もデータ圧縮できるモデルを見つけることである



データ圧縮

X1	天気	X2	温度	X3	湿度	X4	風	Y	プレイ
	晴れ		暑い		高い		無し		×
	晴れ		暑い		高い		有り		×
	曇り		暑い		高い		無し		○
	雨		暖かい		高い		無し		○
	雨		涼しい		ふつう		無し		○
	雨		涼しい		ふつう		有り		×
	曇り		涼しい		ふつう		有り		○
	晴れ		暖かい		高い		無し		×
	晴れ		涼しい		ふつう		無し		○
	雨		暖かい		ふつう		無し		○
	晴れ		暖かい		ふつう		有り		○
	曇り		暖かい		高い		有り		○
	曇り		暑い		ふつう		無し		○
	雨		暖かい		高い		有り		×
									



$$\mathcal{L}(D^n | M) + \mathcal{L}(M) \Rightarrow \text{最小化}$$

[データ符号長] [モデル符号長] w.r.t. M

確率的コンプレキシティ

MDL (Minimum Description Length) 原理

潜在的ダイナミクス1

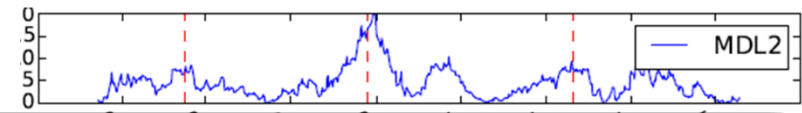
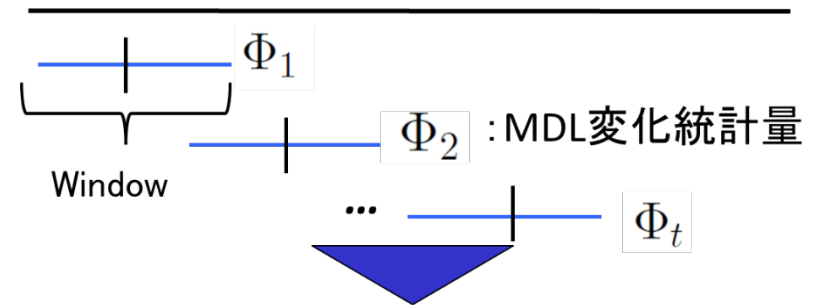
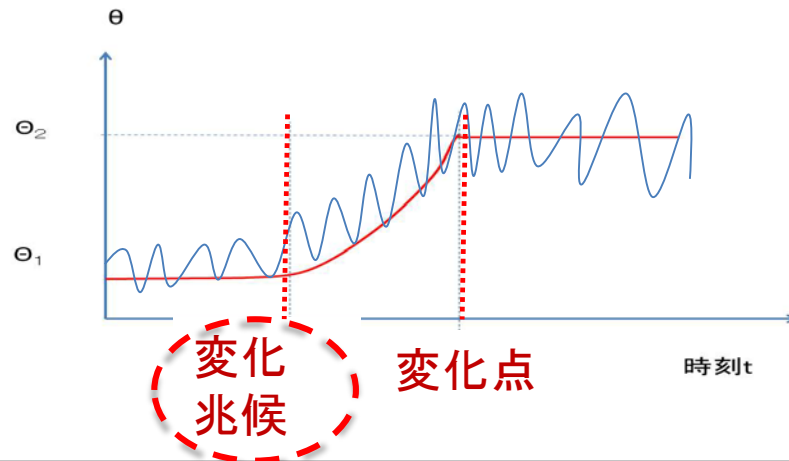
MDL変化統計量

データ圧縮度合いで漸進的変化の開始点(予兆)を検知

[Yamanishi, Miyaguchi BigData2016] [Kaneko, Miyaguchi, Yamanishi BigData2017]

[Yamanishi, Fukushima IEEE Trans Inform Theory 2018]

$$x^n = x_1 \dots x_n,$$



技術コア: **MDL変化統計量**によるオンライン変化検知を提案

MDL: Minimum Description Length

$$\Phi_t = \min_M L_{\text{NML}}(x_1^n; M) - \min_{M_1, M_2} \{L_{\text{NML}}(x_1^t; M_1) + L_{\text{NML}}(x_{t+1}^n; M_2)\}$$

変化した場合と
しない場合の
記述長の差

⇒

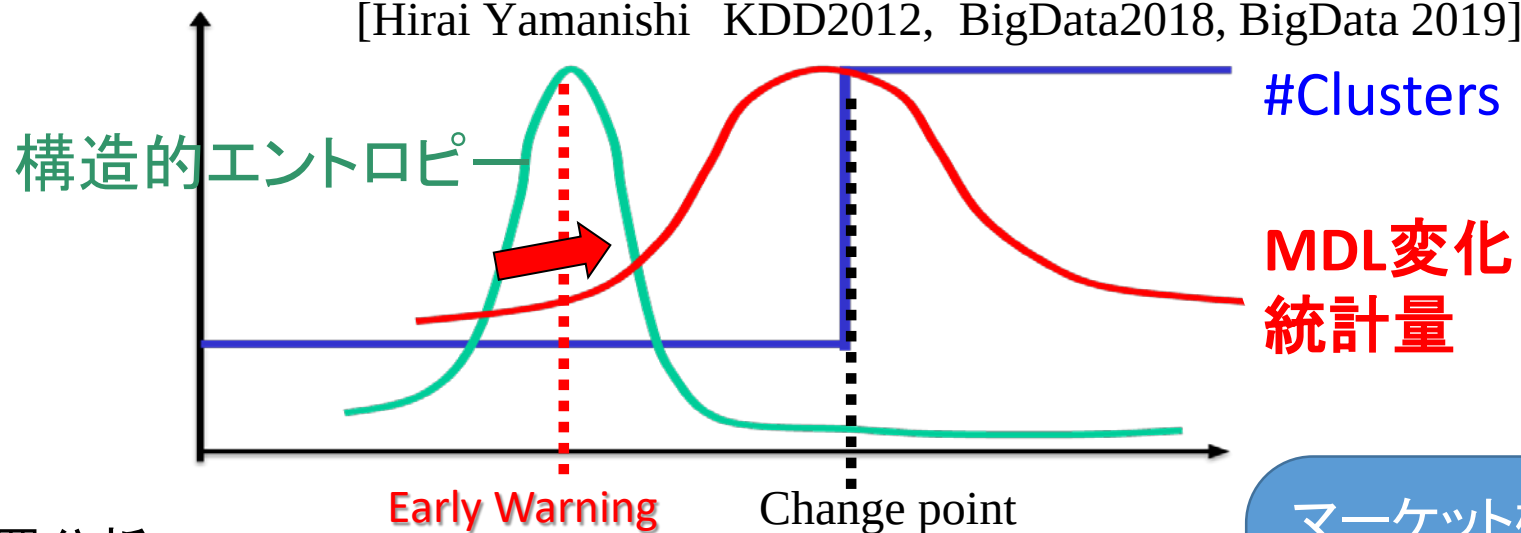
- ・固定窓で移動計算、漸進的変化検知 [BigData 2016]
- ・可変窓で移動計算、検知精度を大幅改善 [BigData 2017]
- ・工場のボイラー事故の予兆を検知(東レとの実証実験)

潜在的ダイナミクス2

潜在構造変化予兆検知

連続量(構造エントロピー、MDL変化統計量)で変化予兆を検知

[Hirai Yamanishi KDD2012, BigData2018, BigData 2019]

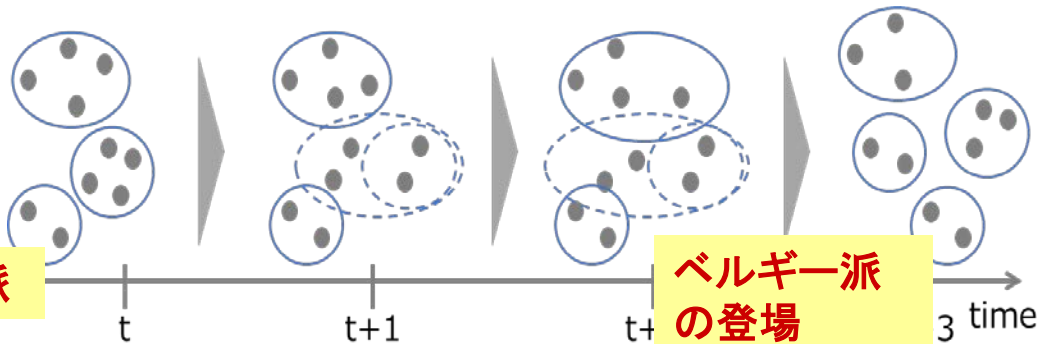


ビール購買分析

発泡酒派

地ビール派

プレミアム派



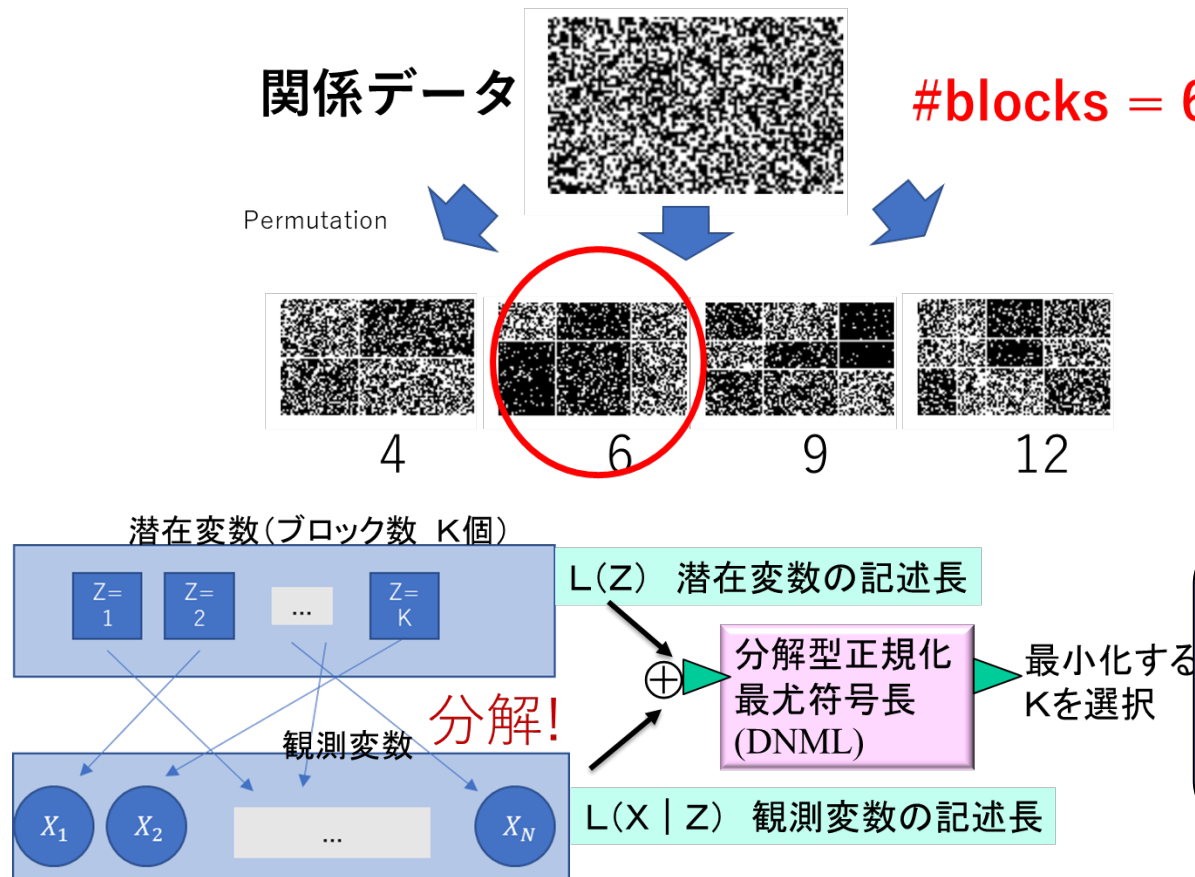
マーケット構造
変化検知
行動パターン
変化検知
に展開可能

$$H_t = - \sum_k p(k|\mathbf{x}) \log p(k|\mathbf{x}_t) \quad p(k|\mathbf{x}_t) = \frac{\exp(-\beta L_{\text{NML}}(\mathbf{x}_t; k))}{\sum_{k'} \exp(-\beta L_{\text{NML}}(\mathbf{x}_t; k'))}$$

潜在構造最適化 1

DNMLによる潜在変数モデル選択

分解型正規化最尤符号長規準(DNML)で潜在変数の数を最適化



[Yamanishi, Wu, Sugawara, Okada DAMI 2019]

[Wu, Sugawara, Yamanishi KDD2017]

[Okada, Yamanishi, Masuda RSOS2019]

- 1) 効率計算可能
- 2) 広い適用可能性
- 3) 高精度かつロバストなモデル選択

$$P(\underset{\text{観測変数}}{X}; K) = \sum_Z P(\underset{\text{潜在変数}}{X, Z}; K)$$

$$\min_K \{L_{\text{NML}}(X|Z; K) + L_{\text{NML}}(Z; K)\}$$

潜在構造最適化 2

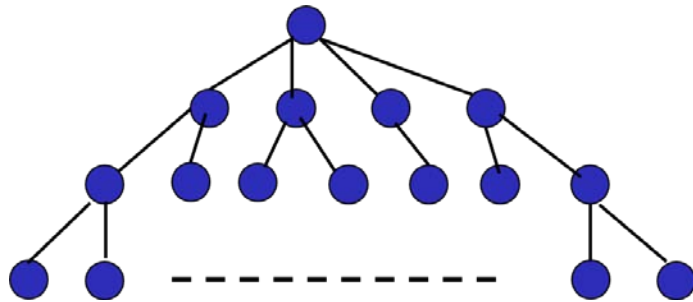
双極順序付き埋め込み

順序関係を保ちつつ双極空間に埋め込むには？

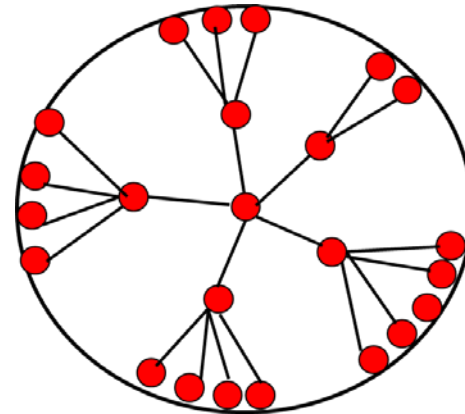
[Suzuki, Wang, Tian, Nitanda, Yamanishi ACMLI19]

[Wang, Suzuki, Xu, Fei, Yamanishi IJCAI2019]

- **Embedding:** map of objects to a metric space



高次元ユークリッド空間



低次元非ユークリッド空間 (双曲空間)

- Hyperbolic Ordinary Embedding

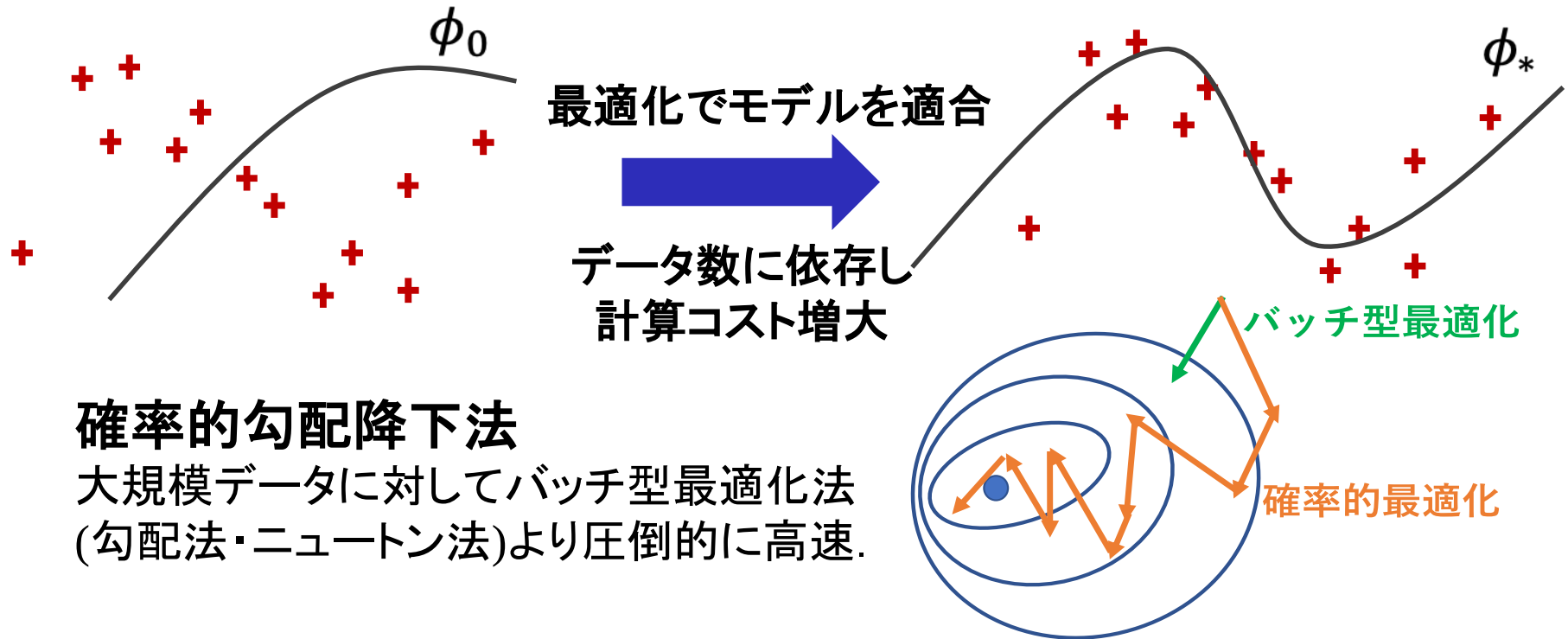
$$i \leftrightarrow j \text{ is more similar than } i \leftrightarrow k \Rightarrow d_{\mathbb{H}^D}(x_i, x_j) \leq d_{\mathbb{H}^D}(x_i, x_k)$$

- Distance function $d_{\mathbb{H}^D}(\mathbf{x}, \mathbf{y}) = \text{arcosh}\left(1 + \frac{2(\mathbf{x}-\mathbf{y})^\top(\mathbf{x}-\mathbf{y})}{(\mathbf{1}-\mathbf{x}^\top\mathbf{x})(\mathbf{1}-\mathbf{y}^\top\mathbf{y})}\right)$

効果的な可視化と
ユークリッド空間
よりも高精度な
リンク予測を実現

確率的最適化

大規模機械学習用の高速最適化手法



問題固有の構造を利用した更なる高速化の研究

- 1: **加速法＋分散縮小法** [Nitanda NIPS2014, AISTATS2016, Nitanda, Murata, Suzuki ICDM2019]
最適反復計算量を最小ミニバッチサイズで達成した世界初の手法.
- 2: **確率的DC最適化法** [Nitanda, Suzuki AISTATS2017]
DC構造を利用しボルツマンマシンのEMアルゴリズムを改良・高速化.

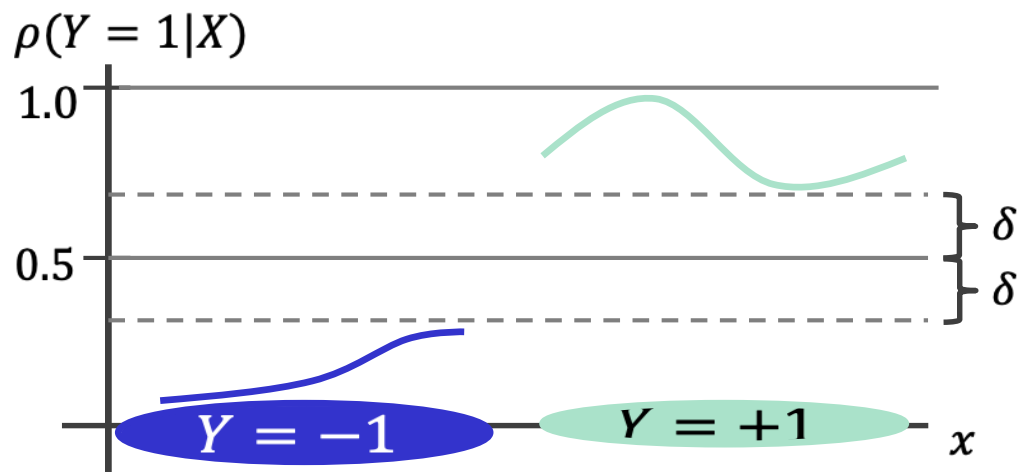
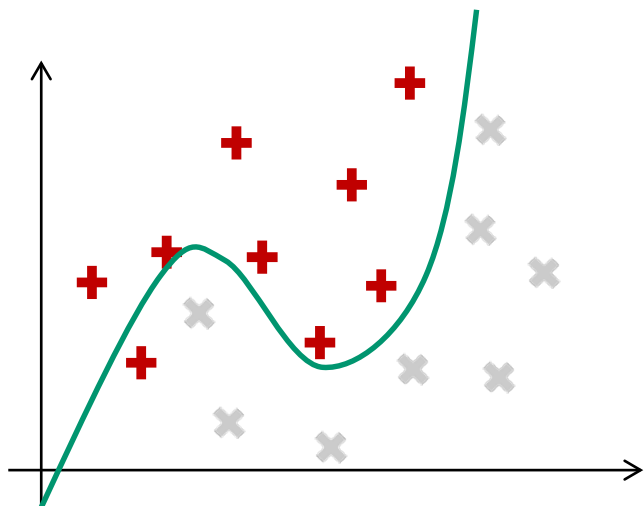
識別問題に対する確率的勾配降下法

ラベルの低ノイズ仮定の下, 識別誤差が指数収束

識別問題は確率的勾配降下法が
特に成功しているタスクの一つ
→問題特性を活かした解析をしたい

ラベルの低ノイズ条件
(ラベルの曖昧さが低い, 例: 画像)

$$|\rho(Y = 1|X) - 0.5| > \delta$$



確率的勾配降下法による識別誤差の収束性 [Nitanda, Suzuki AISTATS2019]

g_T : RKHSでのSGDの反復, g_* : ベイズ則(真の識別関数)

$$\mathbb{E} \left[\mathbb{P}[\text{sgn}(g_T(X)) \neq Y] - \mathbb{P}[\text{sgn}(g_*) \neq Y] \right] \leq O(\exp(-cT)).$$

関数勾配法による深層ニューラルネット学習

関数勾配を用いたResNetの理論解析と手法開発

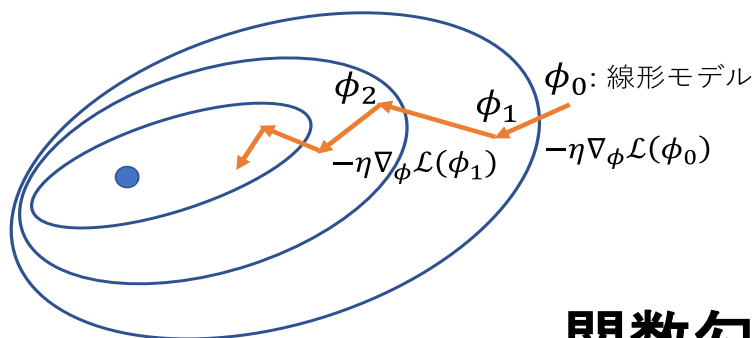
ResNet: 残差ブロックを待つ深層ニューラルネット

関数勾配法的解釈

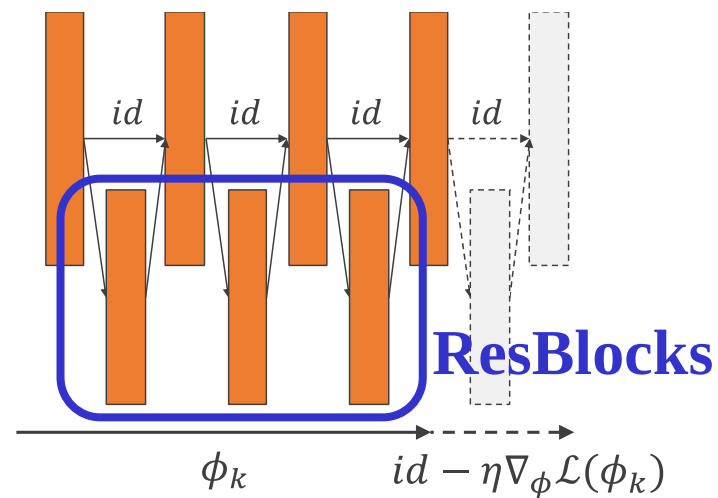
[Nitanda, Suzuki AISTATS2018, ICML2018]

$$\phi_{s+1} = (id + F_s) \circ \dots \circ (id + F_0) \circ id$$

$$-\eta \nabla_{\phi} \mathcal{L}(\phi_t) \sim F_t \circ \phi_t = \phi_{t+1} - \phi_t$$



層が積み重なる過程を
関数最適化法と見なす.



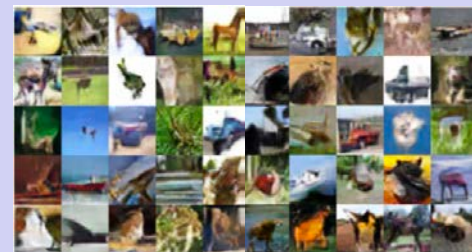
関数勾配法の応用例

1: **生成モデル学習** [Nitanda, Suzuki AISTATS2018]

輸送理論を用いた収束解析.

2: **識別学習** [Nitanda, Suzuki ICML2018]

ブースティング理論を用いた汎化解析.



交通リスクマイニング 1

教師なしリスクパターン抽出

ヘテロ情報から危険道路のパターンを抽出

[Moriya, Matsushima, Yamanishi DSAA2015, IEEE Trans ITS 2018]

GIS リンク
データ

クチコミデータ

交通流データ

事故データ
(90000件)

ブレーキ
データ

クラスター1
クラスター2
クラスター3

特徴ベースNMF
部分順序付きNMF

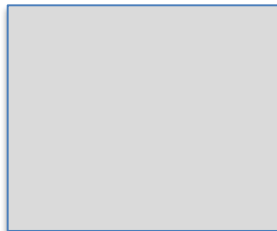


クラスター1
クラスター2
クラスター3

クラスター1
クラスター2
クラスター3

ブレ
ーキ
事故
頻度
交通
流

地点A
地点B
地点C
.....



=

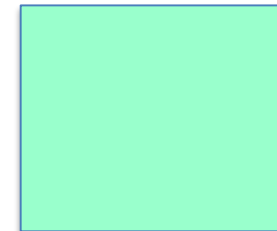
地点A
地点B
地点C
.....

クラス
ター1
クラス
ター2
クラス
ター3



×

ブレ
ーキ
事故
頻度
交通
流



クラスター1
クラスター2
クラスター3

+a + b + c

交通リスクマイニング 2

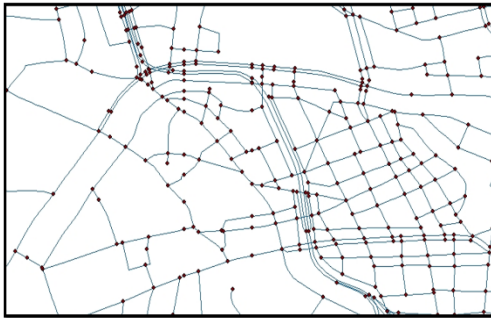
教師付きパターン抽出

地形データをネットワーク中心性で数値化、危険パターンを特徴づけ

[Kobayashi, Lee, Matsushima, Yamanishi BigData 2017]

[Lee, Matsushima, Yamanishi DAMI 2019]

地形
データ



高速パターンマイニングに基づく分類

$$\min_w C \sum_{i=1}^m L \left(\sum_{\phi \in \Phi(k)} w_{\phi} \phi(\mathbf{x}_i), y_i \right) + \|w\|_1$$

ただし, $L(f(\mathbf{x}), y) = \log(1 + \exp(-yf(\mathbf{x})))$



Network中心性
で数値表現

(次数、近傍ノード数、
ページランク、近傍中心
性、平均リンク数 etc)



低コストデータ



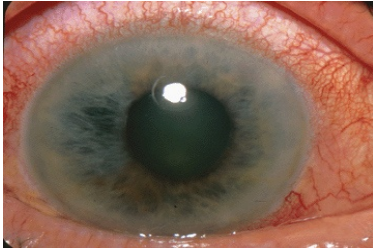
- ・危険パターンの抽出
- ・潜在的危険個所の
特定

ビッグデータから
コンパクトなパターン
を高速に抽出

緑内障進行予測 1

集団データからの時間シフト線形回帰

■ 緑内障 (Glaucoma)



緑内障眼



正常視野

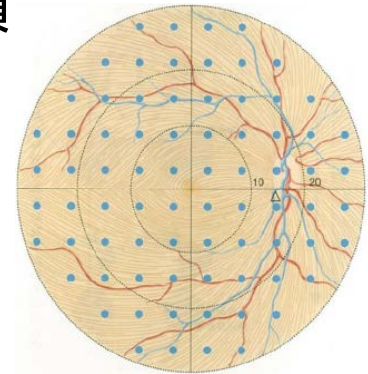


視野欠損

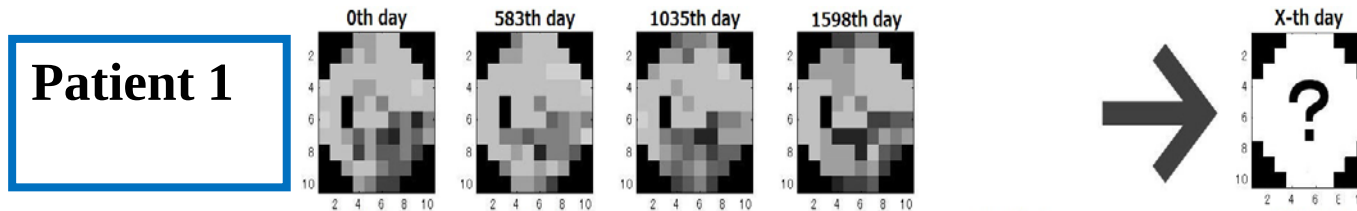
<http://en.wikipedia.org/wiki/Glaucoma>

■ データ.....68点の視野感度(Humphrey Feld Analyzer)

■ 緑内障進行予測



根本等 編
視野. 眼科プラクティス. 2007.



- 視野検査は時間がかかりノイズが大きい
- 個々の患者のデータは少ない ⇒ 似た空間パターンをもつ患者のデータを用いて時間シフト線形回帰

[Liang et al. ICDM2013][Maya et al. BigData2014] [Maya et al. KDD2015]

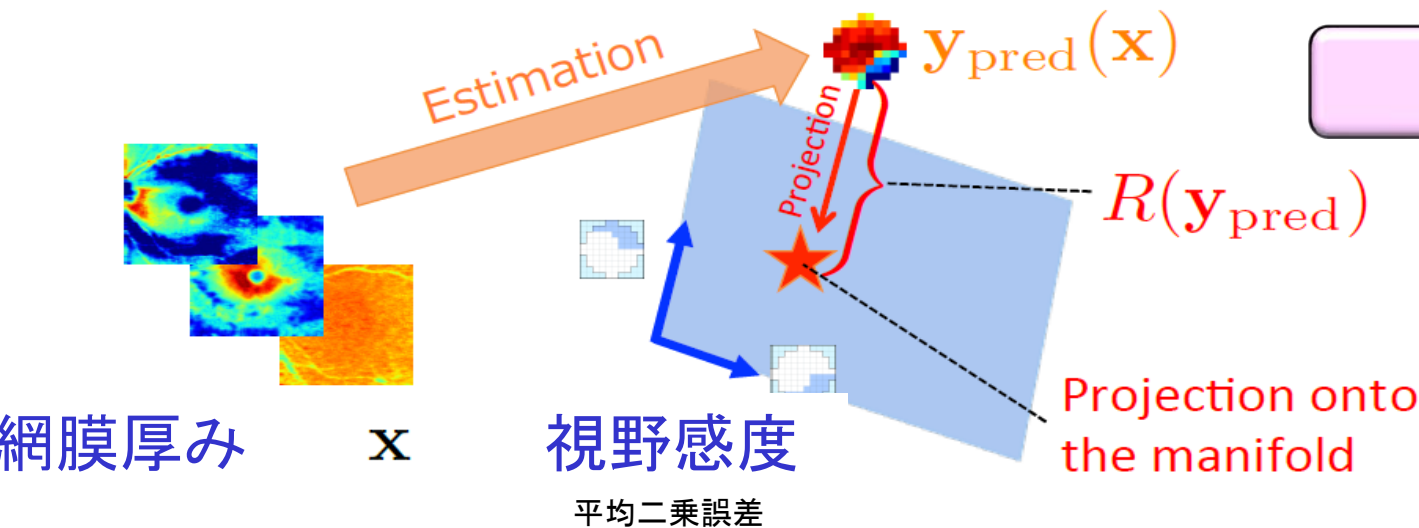
緑内障進行予測 2

パターン正則化

パターン正則化により網膜厚みから視野感度を高精度に推定

[Sugiura, Kiwaki, Siamak. Murata, Asaoka, Yamanishi KDD2018]

[Xu, et al. American Jr. Opht 2020][Hashimoto et al. British Jr. Opht. 2020]



パターン正則化

教師なしで学習した視野パターンからずれないように正則化

	Median	Mean	Worst
CNN [30]	6.68	6.76	21.7
CNN + DA	6.70	6.52	21.7
CNN + DA + PBR1(NMF)	6.41	6.30	23.5
CNN + DA + PBR1(PCA)	6.31	6.26	24.1
CNN + DA + PBR1(AE)	6.39	6.26	24.2
CNN + DA + PBR1(PCA) + PBR2	6.00	6.16	21.4
CNN + DA + PBR1(AE) + PBR2	6.02	6.16	21.5

提案手法の最新結果

$$\frac{1}{N} \sum \|\mathbf{y}_{\text{pred}}(\mathbf{x}_n) - \mathbf{y}_n\|^2 + \frac{\beta}{2} R(\mathbf{y}_{\text{pred}}(\mathbf{x}_n)),$$

$$\mathbf{y}_{\text{proj}} \stackrel{\text{def}}{=} \Theta^\top (\Theta \Theta^\top)^{-1} \Theta (\mathbf{y}_{\text{pred}} - \mathbf{b}) + \mathbf{b}.$$

東大附属病院眼科教室との共同研究

緑内障進行予測 3

DLLR(deeply regularized latent space linear regression)

網膜厚と視野感度の時系列の双方を潜在空間で統合、予測

[Tei, Xu, Kiwaki, Wang, Murata, Asaoka, Yamanishi KDD2019]



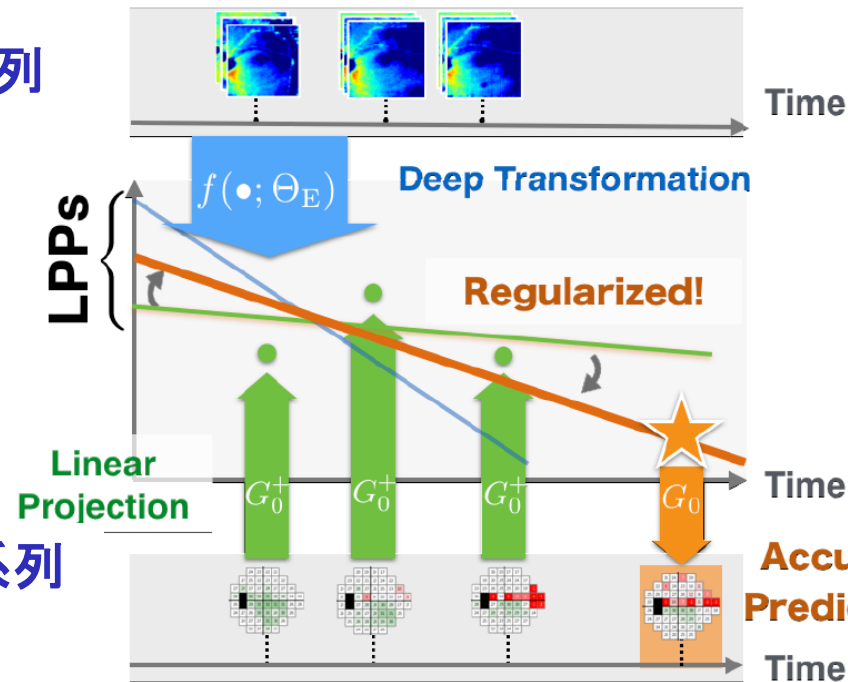
網膜厚時系列



潜在空間



視野感度時系列



視野感度の
最高予測
精度達成

潜在進行
パターン

$$\mathcal{L}_{\text{DLLR}} = \sum_{i=0}^N \|M_i \odot (F_i - G_i W_i P_i^\top)\|_F^2 + \lambda_0 \sum_{j=1}^N z_{0,j} \|W_0 - W_j\|_F^2 + \lambda_1 \sum_{i=0}^N \|f(C_i; \Theta_E) - W_i\|_F^2,$$

Latent space leaning

Collective learning

RT to VF transformation
With CNN

2015－2016の修論テーマ

- ・ネットワーク中心性を用いたテンポラルネットワークの変化検知
DSAA16採択
- ・ヘテロデータに基づく緑内障進行予測のための個々の患者に特化した病状のモデル化
Healthinf16採択 2015研究科長賞(創造)
- ・記述長最小化原理に基づく関係データ学習の研究
SDM16採択
- ・Support Vector Machine 及びトピックモデルを用いた位置情報による個人識別の研究
- ・完全変数化NML 符号長に基づく正準相関分析のランク推定の研究
DSAA17採択
- ・Study on Learning from Nonstationary Time Series
DSAA15採択
- ・サポートベクトルマシンの最適化アルゴリズムの研究
- ・野球データ解析に対する機械学習アプローチ
DSAA16採択
- ・交通データからのリスク因子発見に関する研究
2016研究科長賞(創造)
- ・緑内障眼における網膜層厚視野感度間マルチビュー学習
KDD17採択
2016研究科長賞(数理)
- ・A Study on Model Selection for Hierarchical Mixture Models
KDD17採択
- ・畳み込み行列分解のパラメータ推定と構造選択
ICDM16採択

2017-2019の修論テーマ

- Model Selection for Non-negative Tensor Factorization with MDL Entropy採択
- パターン正則化学習を用いた網膜層厚からの視野感度推定と知識発見 KDD18採択
- 分解型正規化最尤符号長を用いた連続値潜在変数モデルの選択 DAMI採択 RSOS採択
- 統計的モデル選択を用いた勾配ブースティング木の早期停止ルールの研究
- A machine learning approach to glaucoma progression prediction KDD19採択
- 双曲空間における測地線更新アルゴリズムの研究
- グラフ埋め込みにおける正則化の研究 2016研究科長賞(創造) AIPラボ長賞
- 敵対的学習における目的関数の設計及び解析に関する研究

2017-2019の博論テーマ

- Learning High-dimensional Models with the Minimum Description Length Principle SDM18採択 MLJ採択 AISTAT19採択
- Machine Learning over Space Forms ISIT118採択 ACML18採択
- Detecting Model Changes and their Early Warning Signals with the Minimum Description Length Principle BigData18採択 BigData19採択

2016-2019 の卒論テーマ

- ・可変窓上の記述長最小化に基づく時系列データの変化検知 **BigData17採択**
- ・ネットワーク情報を用いた道路の潜在的危険予測 **BigData17採択**
- ・オンライン非定常ロバスト学習とその異常検知への応用 **2017工学部長賞(計数)**
- ・複数の関係パターンを持つマルコフ確率場のMDL 原理に基づく構造学習
- ・緑内障診断のための網膜厚に対する視野感度の微分値の異常検知
- ・混合分布を用いる パターン正則化学習と その眼科学への応用
- ・潜在空間埋め込みを用いる ベイズ実験計画の研究

大学院修了生の就職先 2011-2019

■大学・研究機関

産業総合研究所、Greenwich大

■サービス、IT

Google, NTTデータ、野村総研、楽天、M3、Yahoo、
Indeed、キーエンス、ソフトバンク、Amazon

■製造業

日立製作所、キャノン、ファナック、東芝、SONY、日本IBM

■金融、保険

野村証券、明治安田生命

■広告、放送

博報堂、朝日放送