階層的自然観とスパースモデリング

東京大学　大学院新領域創成科学研究科
複雑理工学専攻
岡田真人

本スライドのまとめ

- 自然への理解を目指す科学には以下の二つのアプローチが存在する。
  - ミクロな基礎方程式からの演繹で、自然を理解できるという要素還元主義
  - ミクロな基礎方程式からの演繹だけでは、自然が理解できないという階層的自然観
- 本スライドでは、後者の階層的自然観の立場をとる。
- ある階層の数理的な記述に関して、その下のレベルからの演繹だけでは記述できないという立場をとる。
- その際には、観測データを解析する基盤技術として、スパースモデリングが必須である。
- 実験観測データの生成モデルが数理的に明らかでない場合、データの特徴量をスパースモデリングで選択することが本質的である。
- スパースモデリングにより、データから知りたい特性を予測することができるようになる。
- さらに、特徴量選択を用いることで、データの生成モデルの候補を絞り込むことができる可能性がある。

要素還元主義の破綻

- 要素還元主義：ミクロな基礎方程式からの演繹で、自然を理解できるという立場
- 要素還元主義はとても美しい立場であるが、これまでの科学の以下の発展の事例から考えて、要素還元主義は破綻していると言わざるを得ない。
- 事例1：量子力学と電磁気学で第一原理的な方程式が書き下せる物性物理学において、強創刊電子系などの多対効果が無視できない系においては、計算機を使った第一原理計算でも、系の性質を記述することができないという事実がある。
- このような場合は、実験計測使って、系の性質を決めるパラメータをデータ駆動的に決めざるを得ない。
- 事例2：例えば恋愛感情のような高度な認知機能のメカニズが量子力学や電磁気学から演繹できるとは誰も思わない。

階層的自然観に対する
普遍的アプローチは存在するか？

- 階層的自然観に対して、各階層個別にしか研究できないのか? 各階層に依存せずに、研究することな可能な普遍的枠組みは存在するか？
- 我々は、その普遍的アプローチの数理基盤の一つがスパースモデリングであるというパラダイムを提案する。
- 階層によらずに、以下のような状況が存在する。
  - 実験計測データを説明する数理モデルが存在しない。
  - そこで、スパースモデリングを用いて、データ駆動的にデータをもっとよく説明できるモデルを構成する。
  - 具体的には、データもしくはデータの特徴量から、データそのものを再構成したり、別のデータを予測することを考える。
  - そこで用いる学習機械は線形回帰などの単純なモデルを一般的には用いる。
  - スパースモデリングによる特徴量抽出を行い、データの再構成や予測を行う。
  - 抽出された特徴量を用いて、データを生成する数理モデルの構築の一助とする。

スパースモデリングとは

スパースモデリング基本的な考え方は

(1) 高次元データの説明変数が次元数よりも少ない(スパース(疎)である)と仮定し,

(2) 説明変数の個数がなるべく小さくなることと,データへの適合とを同時に要請することにより,

(3) 人手に頼らない自動的な説明変数の選択を可能にする枠組みである.

スパースモデリングの具体例
線形回帰における定式化

目的変数 (機能) $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N$ 説明変数

$$= \boldsymbol{\beta}\mathbf{x}$$

サンプル数p, 説明変数の数 N

回帰係数 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, 0, \ldots, \ 0, \beta_N)$

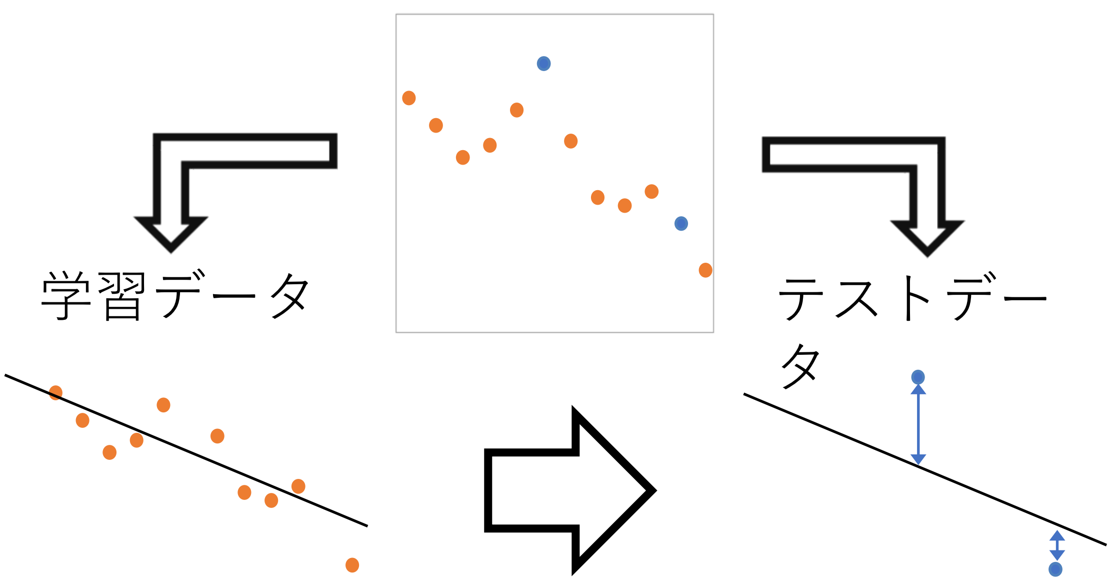
インディケータ $\mathbf{c} = (1,\ 1, \ldots, 0, \ldots, 0, 1)$

$\mathbf{c} = \{0,1\}^N \quad 2^N$ - 1 通り

全状態探索(Exhaustive Search, ES) 法

Garside 1965, Ichikawa et al., 2014, Nagata et al., 2015, Igarashi et al., 2016, Igarashi et al., 2018

1. $\mathbf{c}$ を決めると記述子の係数 $\boldsymbol{\beta}$が決まる
2. その $\mathbf{c}$ に対する回帰モデル($\boldsymbol{\beta}$)の交差検証誤差(CVE)を求める
3. 各記述子の組み合わせ($\mathbf{c}$)を評価



交差検証によって，限られたデータから汎化誤差を推定する

線形回帰における全状態探索(ES-LiR) 法

インディケータ $\boldsymbol{c}$		$\boldsymbol{\beta}$ を用いて$\mathbf{y}$ の推定		CVE
$\boldsymbol{c}=(1,0,0)$	$\Longrightarrow$	$\hat{y}=b_0+b_1x_1$	$\Longrightarrow$	$error=\boldsymbol{0.20}$
$\boldsymbol{c}=(0,1,0)$	$\Longrightarrow$	$\hat{y}=b_0+b_2x_2$	$\Longrightarrow$	$error=\boldsymbol{0.21}$
$\boldsymbol{c}=(0,0,1)$	$\Longrightarrow$	$\hat{y}=b_0+b_3x_3$	$\Longrightarrow$	$error=\boldsymbol{0.23}$
$\boldsymbol{c}=(1,0,1)$	$\Longrightarrow$	$\hat{y}=b_0+b_1x_1+b_3x_3$	$\Longrightarrow$	$error=\boldsymbol{0.16}$
$\boldsymbol{c}=(1,1,0)$	$\Longrightarrow$	$\hat{y}=b_0+b_1x_1+b_2x_2$	$\Longrightarrow$	$\boxed{error=\boldsymbol{0.14}}$
$\boldsymbol{c}=(0,1,1)$	$\Longrightarrow$	$\hat{y}=b_0+b_2x_2+b_3x_3$	$\Longrightarrow$	$error=\boldsymbol{0.18}$
$\boldsymbol{c}=(1,1,1)$	$\Longrightarrow$	$\hat{y}=b_0+b_1x_1+b_2x_2+b_3x_3$	$\Longrightarrow$	$error=\boldsymbol{0.20}$
	線形回帰		CVEの計算	

2^3-1 = 8 通り，すべてのCVEをチェックする

記述子の個数Nが十分に大きい場合，計算量O(2^N)が膨大になる

スパースモデリングは深層学習を含めほとんどの学習機械に適用可能

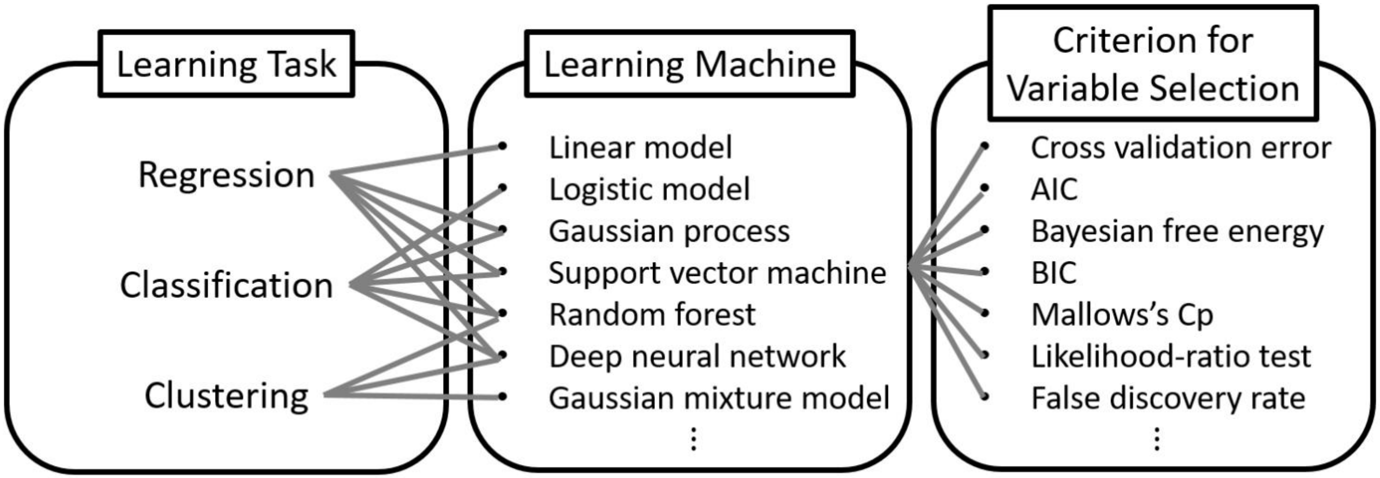


Figure 1. Constitution of exhaustive search. Exhaustive search is conducted on every triplet of learning task, learning machine and criterion for variable selection.

Yasuhiko Igarashi, Hiroko Ichikawa, Yoshinori Nakanishi-Ohno, Hikaru Takenaka, Daiki Kawabata, Satoshi Eifuku, Ryoi Tamura, Kenji Nagata, and Masato Okada
“ES-DoS: Exhaustive search and density-of-states estimation as a general framework for sparse variable selection”
International Meeting on "High-Dimensional Data-Driven Science" (HD3-2017)
Journal of Physics: Conference Series, 1036, 012001, (2018)