2022年3月21日

第５回 統計教育シンポジウム

小中の実践発表を受けて・大学での取組

お茶の水女子大学
附属高等学校
理学部 情報科学科
吉田 裕亮

小学校での実践発表を受けて

小学校での実践事例

さつまいもの分け方

13個のサツマイモを 4 人の先生方にお渡しするため, その分け方を考える

各サツマイモの重さは,
110, 120, 120, 150, 180, 230, 230, 280, 310, 330, 350, 390, 470
である.

★ 目標は, 4人の先生にできるだけ公平になるように分けたい.

離散最適化（組合せ最適化）問題

離散最適化（組合せ最適化）　問題

離散的な組み合せの中（場合の数）から，ある基準に最も適する場合を見つけ出す問題

ある基準は評価関数（リスク関数）とも呼ばれ，この問題は，関数の最大（最小）問題と考えられる．

一般には, 比較的個数が少ない場合でも, 組合せ(場合の数)が膨大になり, 全ての場合を尽くして調べるのが現実的でないこともある.

＝＞　確率的なアルゴリズムで最適解を探索する．代表的な手法

- 焼きなまし法（Simulated Annealing）
- 遺伝的アルゴリズム（Genetic Algorithm）

離散最適化（組合せ最適化）問題

離散最適化問題への戦略

代表的な手法の

- 焼きなまし法（Simulated Annealing）
- 遺伝的アルゴリズム（Genetic Algorithm）

においては，どちらも適当な初期値から始めて，局所的に入れ替えを行って，入れ替えをすべきか否かを評価関数の値を比較することで最適解に進んで行くような方法を取る．

子ども達も基本的には，このような戦略を取っている．

評価関数1

評価関数 1

基準1.　各クラスの重さの和が出来るだけ均等になるように分けたい

サツマイモの重さの合計を 4 で割った値が 817.5 なので
この基準のための評価関数は クラス j の重さの合計を w_j として

$$R_1 = \sum_{j=1}^{4} \left| w_j - 817.5 \right|$$

の関数を考えている

評価関数1

$$R_1 = \sum_{j=1}^{4} \left| w_j - 817.5 \right|$$

R_1 による最適化の結果は, 各クラスの重みの組合せは
{ 810, 820, 820, 820 } が最適であることが分かる.

一般に、平均に関連した最適化には、絶対誤差和よりも２乗誤差和

$$\widetilde{R_1} = \sum_{j=1}^{4} \left(w_j - 817.5 \right)^2$$

を評価関数に用いることが多い.

このサツマイモの場合は, ２乗誤差和でも同じ重みの組合せが最適解

評価関数2

評価関数２

基準2.　各クラスの個数が出来るだけ均等になるように分けたい

サツマイモの個数を 4 で割った値が 3.25 なので
この基準のための評価関数は クラス i の個数を c_j として

$$R_2 = \sum_{j=1}^{4} \left(c_j - 3.25\right)^2$$

となる.

このときの最小化は $\{3, 3, 3, 4\}$ の組合せの場合である.

ここで　$R = R_1 + R_2$ を考えてこの最小化を行っている

評価関数3

子どもたちは, さらなるの評価基準も導入している.
すなわち, サツマイモの形の多様性である.

評価関数３

基準3.　各クラスのサツマイモの形は多様な方が良いとする

各クラスのイモの形の多様度を v_j とする.
ただし, v_j の値が大きい程, 多様性が高いことを意味する.

$$R_3 = -\sum_{j=1}^{4} v_j$$

となる.

最終的に　$R = \lambda_1 R_1 + \lambda_2 R_2 + \lambda_3 R_3$　を評価関数と考えて最小化している

サツマイモの最適解

A，B 案以外の解の例

今回の 基準１と基準２に基づく最適解には，A 案，B案以外にも

{110, 230, 470}, {150, 280, 390}, {180, 310, 330}, {120, 120, 230, 350},

{120, 230, 470}, {150, 280, 390}, {180, 310, 330}, {110, 120, 230, 350}

等々他にも多数ある.

子どもたちの例では，可能な最適解のすべての場合が尽くされて
検討されているわけではないが，現実問題として解いた結果
としては，目的を十分に達成しているであろう.

最適化とデータ解析

最適化と統計的データ解析

統計的データ解析において評価関数の最適化は基本

特に，トレード・オフの関係にある評価関数の和の
最適化は重要

例えば 回帰分析における正則化 （過学習を抑える手法）

$R = MSE$ (平均２乗誤差) + λ (L^2 正則化項)

Ridge 回帰

$R = MSE$ (平均２乗誤差) + λ (L^1 正則化項)

Lasso 回帰

モデル選択

情報量基準

もっと一般に，モデル選択の情報量基準もトレードオフの関係にある関数の和の最適化である

$$I = -2\,MLL\ \text{(最大対数尤度)} + 2\,k\ \text{(モデルパラメタ数)}$$

中学校での実践発表を受けて

中学校での実践事例

振り返りの重要性

生徒自身にとって, 自らの活動の評価だけでなく,

教員側にとっても
生徒の考え方や捉え方,
生徒の気づきに 気づく機会となる.

今回の実践報告の中に、非常に興味のある振り返り記述がある

生徒の振り返りシート

振り返りシート　ロイロノートから

1　全数調査と標本調査の特徴・違い

①私は、標本は全体によく似ているもので、多少の誤差はあれど、そこまでではないと思っていた。もちろん標本平均の傾向として、間違っているものではなかったが、時々1時間程度の標本の差があって、無作為抽出をしたからといって、正確な値になるのではないのだなと思った。しかし、調査がとても早く終わったので、全数調査よりも使いやすい手段だなと思った。	はじめは、全数調査に比べて標本調査は誤差が大きく、信頼性に欠けると思っていましたが、皆で取った標本の平均を出して箱ひげ図にしてみると、ほとんどの人の睡眠時間が小さな範囲内に収まっていたので、信頼してもいいのかなと思いました。また、全数調査が不可能なこと（水質調査など）や、短時間で大まかな推定ができればいいものもあるため、標本調査はとても重要な調査方法の一つだということがわかりました。

誤差についての考えが変わった記述

(生徒のロイロノートより)

生徒の振り返りシート

4　標本平均

標本平均と母平均はそこまで違いは出ないけど、そもそもの抽出した標本が偏っていたらあまり意味がないものとなってしまう。
また 標本の大きさが 多いほうが標本平均の正確性が増す。

10個とった時の数が、全数調査と離れているのでやっぱり、標本調査ではズレが出てくると思った。ずれてしまった原因は標本の数値の差が大きくて外れ値を取ってしまったときに、差が出てきてしまった。なので、値の数の差があまりない数では、標本調査があまり誤差がなく、調べることができると思った。

偏らないで標本を選ぶこと、標本の大きさについての記述

標本の範囲が小さいと誤差が少なくできることがわかった記述

(生徒のロイロノートより)

点推定から区間推定

区間推定に向けて

・標本平均値の信頼性について気が付いている生徒がいる

・標本数が多くなると, 信頼度(正確性)が増すと主張している

=> 生徒は, 標本平均値だけではなく, 何らかの信頼度を
込めて推定することが重要と気づいている.

中学校での範囲ではないと思われるが、
信頼度を込めた, 区間推定の話に発展できるか？

点推定から区間推定

tー分布の代わりに

通常は、標本（母分散は未知）からの区間推定には t -分布を用いて信頼区間を構成する
　　ー＞　中学生には、この方法は無理であろう

何か他の方法はあるか？

ブートストラップ法（1）

経験分布に基づきリサンプリングを行い, 大量のリサンプリング標本より、ある統計量の分布を推定し,推定値の区間 推定や誤差評価を行う.

1979年　Efron により提唱され理論的整備がなされ，以降,
計算機環境の大幅な発展により, 広く用いられるようになった

本来は、標本平均のように統計量の分布が容易に分かる場合ではなく, 分布を求めるのが困難な場合(順序統計など)に威力を発揮する.

ブートストラップ法

ブートストラップ法 (2)

ブートストラップ法は, RANDBETWEEN関数の整数乱数だけを用いて実行可能なので, 表計算のコピー＆ペースト機能が使えれば表計算アプリ上でも行える.

手順は比較的容易ではあるが, 理論的な正当性を説明するのは中学生には困難.

区間推定の概念は重要と思われるので, 他に何か良い手法があればと思われる.

大学(学部)での取組

～ お茶の水女子大での例～

大学での取組

大学(学部)の数理・データサイエンス教育

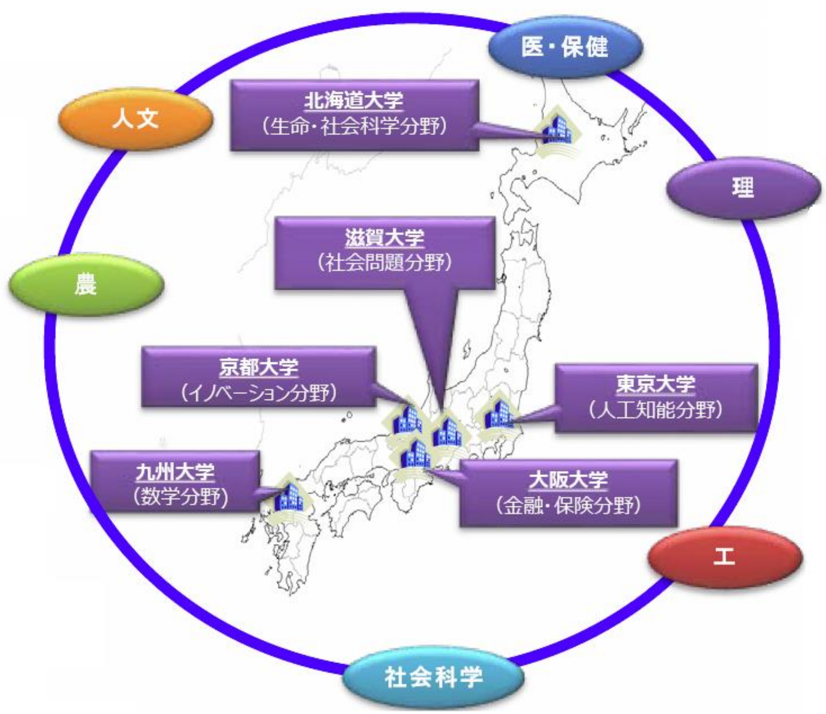
デジタル時代の「読み・書き・そろばん」である「数理・データサイエンス・AI」の基礎などの必要な力を全ての国民が育み、あらゆる分野で人材が活躍する環境を構築する必要がある

大学(学部)教育においては、「数理・データサイエンス(統計学)」は、高等教育における基本的素養として、文系・理系を問わず、すべての学生が修得すべきものであると位置付けている

数理・データサイエンス教育の全国展開

大学（学部）の数理・データサイエンス教育

6大学を拠点として設置し、全学的な数理・データサイエンス教育を先行的に実施するとともに、拠点大学で形成するコンソーシアムにおいて、標準カリキュラム・教材を開発



お茶大での取組

文理融合AI･データサイエンスセンター

2019年度より「文理融合AI･データサイエンスセンター」を設置し、東京大学を拠点校とするコンソーシアムの協力校として活動を開始

全学部学生が実社会や実生活において直結するデータサイエンスを修得するために、その思考方法と基礎を身につけてもらうことを目標としている

カリキュラムマップ


実践知の涵養
卒業研究でのデータサイエンスの実践
自然科学系
実験計画策定
データ処理
人文社会科学系
史資料分析
社会調査
専門知の涵養
シミュレーション関連科目
情報処理科目
統計分析科目
教養知の涵養
プログラミング演習
コンピュータ演習
情報処理演習
文理融合データサイエンス
・全学部の1，2年生を対象
・データサイエンスの重要性を浸透
・e-learning教材の作成と情報発信
・数理情報教育の体系化
・履修認定書の発行

全学データサイエンス学際カリキュラム

現在 お茶の水女子大学の「全学データサイエンス学際カリキュラム」は

文部科学省「数理・データサイエンス・AI教育プログラム
（リテラシーレベル）」 として認定

MDASH
Literacy
Approved Program for Mathematics,
Data science and AI Smart Higher Education

数理・データサイエンス・AI
教育プログラム認定制度
リテラシーレベル

認定有効期限：令和8年3月31日まで

https://www.cf.ocha.ac.jp/datascience/j/menu/curriculum/index.html