MATHEMATICAL
SCIENCES
April 2024 Number 730

特集／データサイエンスと数理モデル

巻頭言

持 橋 大 地

本誌の 2019 年 6 月号で東京大学の山西健司先生を中心に, 特集『データサイエンスの数理』が組まれてから 5 年ほど経ち, 新たに本特集『データサイエンスと数理モデル 諸問題の発見, その解決へのアプローチ』が組まれることになりました.

この間に「データサイエンス」は広く一般に知られる言葉になり, 多くの大学で「データサイエンス学部」が設置, あるいは設置予定になっています. 企業においても, データサイエンスを扱う部署が新たに置かれることが増え, 自然言語処理[1)]および機械学習を専門とする筆者も様々な企業のデータサイエンティスト, あるいは同等の職務の皆様と共同研究を行ってきました.

1. データサイエンスとは

そもそも「データサイエンス」とは何でしょうか. 明確な定義はありませんが, 前回の特集で山西先生が述べられているように,

- データに合わせた **(新しい) モデル化**を行い,
- データを通じて新たな知見を引き出すために, **説明性**を大事にすること

が中心的な概念になっていると考えられます. これは統計学や機械学習を含みつつも, 目指すところが若干異なっています. たとえば, データをブラックボックスの深層学習に入れて, 理由はわからないが予測はできる, というような「モデル」はデータサイエンスとはいえないでしょう. むしろ必要に応じて既存のモデルを使い, あるいは新しくモデルを開発してデータから知見を引き出し, 問題に関する洞察を深めることが目的となります.

よって必然的に,「データサイエンス」は新しい分野, あるいは新しい種類のデータへの取り組みを指す言葉ということになります. 既存の問題であれば, すでに各分野で方法は確立されているはずだからです. 逆に充分に進めば, 開発した方法論は各分野に吸収されて「データサイエンス」とは呼ばれなくなるでしょう. それでも現在この標語が使われているのは, どの分野にも共通な考え方や方法が適用できるからに他なりません.

2. 数理モデルの意義

データサイエンスにおいて, 数理モデルが中心的な役割を果たすことは, 今さら強調するまでもないでしょう. それでは, あらためて, なぜ数理モデルが必要なのでしょうか. それは, それぞれの数理モデルが世界を見る一つの「枠組み」, あるいは一種の科学的な「言語」[*1)]であるからだと筆者は考えています. すなわち, ある「言語」で見えないものが, 別の「言語」では見えることがあるからで

*1) ここでいう「言語」とは, 世界観を表現するデザイン言語のような, 一般的な意味で用いています.

す．本特集の清先生の記事のように, 回帰モデルでは見えないものが新しい相関 (コピュラ) を使えばとらえられる, ということがあります．最適化では解けなかった問題が統計モデルにすると解ける, あるいはその逆もあるでしょう．このとき, 必要に応じて枠組を新しく作る場合もあります．本特集の中村先生・松原先生の記事のように, 文化現象あるいは物理データについて物理方程式と統計モデルを接続することはその好例です．

さらに数理モデルには, 人間の直感を超えられるという特徴があります．筆者も言語を扱う者として, 簡単な例で思い描いていたものを遥かに超える面白い例が, モデルを走らせると見つかることを何度も経験しています．数理モデルは, 情報理論の韓太舜先生が述べられた[2]ように「時には実践を一挙に凌駕する, はるか彼方の地平に我々を着地させる」ことができるものです．その度合いは場合によって異なりますが, データサイエンスの最も大きな夢はここにあると言っていいでしょう．

3. データサイエンスの広がり

最初に述べたように, データサイエンスを規定する特徴は新しい分野, あるいは新しいデータを扱うことです．このとき, 「データ」とは何かも, 今までと大きく異なってくるべきだと考えられます．従来の機械学習における固定された「データセット」, あるいは統計学で多く暗黙に仮定されてしまっている表形式のデータ以外のものも「データ」になりえます．情報化の現在, データサイエンスに使えるデータは無数にあり, その活用方法を発見することも分野の課題の一つでしょう．

よって「データサイエンス」は従来の理工学だけでなく, 人文・社会科学や生物学・医学といった分野にも広がっていくはずです．筆者も所属組織において人文学オープンデータ共同利用センター (CODH) を兼務していますが, どの場所においても数理モデル, およびそのための教育は必須になり, その幅は今後ますます広がっていくと考えられます．

4. 本特集の構成

こうした観点から, 本特集では理工学に止まらず, さまざまな分野から, データサイエンスにとって重要な研究をご紹介します．

統計学の清先生 (東京大学) には, 異種の変数間の相関をとらえられる最小情報従属モデルについて, 従来とは別の切り口から解説していただきました．また, 理論物理の出身で音楽情報科学を研究されてきた中村先生 (京都大学) には, 筆者も協力して現在取り組まれている, 芸術の進化モデルについて解説していただきました．経済学の北川・木戸先生 (ブラウン大学・京都大学) には, EBPM とよばれるデータに基づく政策立案を支える, 政策決定の数理モデルについてご紹介いただき, 「計量・数理政治学会」を最近設立された福元先生 (学習院大学) には, 計量モデルと理論モデルの融合について政治学を例に議論していただいています．松原先生 (大阪大学) には, 深層学習と微分方程式を融合する幾何的深層学習について説明していただきました．日野・有竹先生 (統数研・一橋大学) には, 実問題で重要になるテスト時の分野の変化に, 最適輸送で対応する研究について書いていただきました．神経科学が専門で R コミュニティでも活躍されている三村先生 (統数研) には, 生物データにおける線形混合モデルの重要性を R 実装も交じえて説明していただきました．持橋 (統数研) は, 従来の教科書にないようなガウス過程や相互情報量を使った自然言語処理について解説しています．

実際のデータサイエンスの現場はこれらの最先端の話題がすぐに適用できるとは限りませんが, これらの論考が, 読者の皆様にとってデータサイエンス分野の未来を照らす探照灯になっていただくことを願っています．

参考文献

1) 中谷秀洋, 持橋大地 (特集担当). 岩波データサイエンス Vol.2 特集：統計的自然言語処理—ことばを扱う機械. 岩波書店, 2016.
2) 韓太舜. たかが学問, されど学問－情報理論に魅せられて－ [電気通信大学最終講義]. http://www.quest.lab.uec.ac.jp/han-farewell-lecture.pdf.

（もちはし だいち，統計数理研究所）