データサイエンス

key word

- ダイバージェンス
- 情報幾何
- 分散共分散行列
- リスクの漸近展開

椎名　洋
You Shiina

データサイエンス学部
教授

【プロフィール】
・1986 年 3 月東京大学法学部卒業
・1992 年 3 月東京大学経済学研究科博士課程単位取得満期退学
・2004 年 2 月経済学博士（東京大学）
・1992 年 4 月信州大学経済学部講師
・1995 年 4 月信州大学経済学部助教授
・2004 年 4 月信州大学経済学部教授
・2020 年 4 月滋賀大学データサイエンス学部教授

【主な社会的活動】
・日本統計学会・日本数学会
・日本統計学会誌（和文）編集委員（2007－2009）
・応用統計学会学会誌編集委員（2014-2016）
・ Japanese Journal of Statistics and Data Science Associate Editor （2017-2019）

【主な著書・論文】
・「データサイエンスのための数学」（共著）講談社、2019
・「統計検定 1 級対応 統計学」（共著）東京図書、2013
・’Estimation of a continuous distribution on the real line by discretization methods’ Metrika, 2018.

【代表的な研究テーマ】

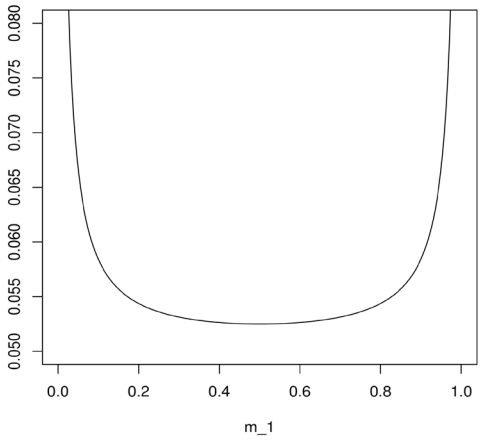
□ 情報幾何による推定の効率性の把握

課題解決に役立つシーズの説明

データ解析の様々な分野で、効率的な予測とそこから生じるリスクの評価は、非常に重要な課題である。分かりやすい例として、表が出る確率が p であるコインを 10 回投げる実験を考える。表が出た割合によって、p を推定する（最尤推定量）時に、どれくらい本当の値と推定値が（平均的に）ずれるかを表にしたものが下記の図である。横軸が p（ラベルは m_1）、縦軸がずれをカルバックライブラーダイバージェンスと呼ばれるもので測っている。p が 0.8 より大きくなる（或いは、0.2 より小さくなる）あたりから、急速にずれが大きくなっているのがわかる。特に、p が 0.9（あるいは 0.1）を超えた際のずれの増加は、著しい。結論としては、1 回に 10 回しか生じないような現象に関する確率を、10 回の試行で判断することが非常に危険であることが分かる、逆に言えばもっと多くの試行、すなわちサンプルの収集が必要であることを示している。この例で重要なのは、本当の値 p とその推定値のずれを直接的に二乗誤差のような損失関数で計測しているのではなく、p によってきまる確率分布と推定値によって決まる確率分布のずれから生まれるリスクを測っている点である。

この例のような単純な場合も含めて、データ解析の際に使われている道具の良し悪し（効率性）について、抽象的な比較ではなく具体的な値として把握されていない事柄が多々ある。損失関数として、多くの統計・機械学習の分野で二乗損失が標準的なものとして使われているが、母集団のパラメーターの変換に関して不変ではないので、最終的なリスクの「絶対的な」評価に使うには不向きである。ダイバージェンスに基づくリスクの比較は、パラメーターの変換に関して不変なので、さまざまな統計的な推定・学習の比較を可能にしてくれる。また、判別機械学習では、予測の正解率が非常によく使われる指標であるが、ダイバージェンスによる分布間のずれの評価は、二つの分布から生まれたデータをどの程度の正解率で判別できるかにもつながるので、二乗損失よりも適用範囲が広い。

以上のようなフレームワークの中で、現在重点的に研究しているテーマは、離散型の分布の漸近的なリスク解析である。離散型分布は特定のモデルに依存しない点や、実務上の多くの計測が実際には離散型にならざるを得ない点で、重要な研究対象であると思われる。情報幾何は、ダイバージェンスから自然に導かれる統計分布に関する幾何的な構造を把握するのに有用な道具であるが、離散型の分布のもつ幾何的な構造が分析対象として扱いやすいことが分かっている。これを使ってよりよい推定の方法を見つけることは、実際のデータ解析の現場において、大きく統計的なリスクを減少させる可能性があり、非常に幅広い分野に適用することができるのではないかと考えている。



企業・自治体へのメッセージ

既存のデータ解析手法の適用に比べて、新しい分析手法の開発は時間もかかるうえに、成功の確率も低いと思われますが、逆に成功した時は大きなリスク・コストの削減につながる可能性があります。そのような分析手法の開発にお役に立てれば幸いです。