● 包 含 特定助教

Han BAO (Assistant Professor)

研 究 課 題： 仕様検証可能な機械学習
（Verifiable machine learning）
専 門 分 野： 統計的機械学習（Statistical Machine Learning）
受 入 先 部 局： 情報学研究科（Graduate School of Informatics）
前職の機関名： 東京大学大学院 情報理工学系研究科
（Graduate School of Information Science and Technology, The University of Tokyo）

現代社会では計算能力や計算資源の発達によって膨大な量のデータが収集可能になり、科学や意思決定においてデータ駆動型の知識発見はますます重要になりつつあります。統計的機械学習とはそのような推論を支える技術の一つであり、統計的手法を援用した帰納推論を現代的な計算機上で実現する分野です。

機械学習はゲーム AI や自動運転など多様な領域で成功を収めていますが、帰納推論はその正しさを保証することが容易ではなく、このことが安全性を重要視する領域で機械学習を用いる妨げとなっています。

私はこれまで、収集したデータに大きな偏りが見られる場合や、第三者によるデータの改竄を想定する場合において、私たちが構築したアルゴリズムが期待する性能を達成できるかを検証する理論的枠組みを研究してきました。これを更に一般化し、利用者が推論に対して期待する「仕様」を予め明示し、構築したアルゴリズムが要請を満たすか検証を行う枠組みを目指しています。

As modern computational powers and resources enable us to collect a vast amount of data, data-driven knowledge discovery has become more and more important in science and decision making. Statistical machine learning is one of the key components to form the basis of such inference, by implementing statistics-based inductive inference on modern computers.

While machine learning has been successful in various domains such as game AI and autonomous driving, it is not straightforward to guarantee the performance of inductive inference, which often hinders machine learning from real-world applications in risk-sensitive domains.

I have been working on establishing a theoretical framework to verify whether our inference algorithm can achieve the expected performance under the situations where the collected data is largely skewed, or we need to defend against a third person's data perturbation. Ultimately, I aim to establish a framework to verify whether a learning algorithm satisfies properties specified by users.

機械学習予測の性能保証の難しさ

機械学習は統計的手法に基づく帰納推論、すなわち有限の観測を元にその背後に潜む法則性を導き出す方法論です。予測がどの程度正しいか、確からしいかを調べることは学問としての機械学習分野において大切な要素の一つであり、統計的学習理論という一大分野が広がっています。統計的推論は有限の観測を一般化するため、未知のデータに対する推論の正しさは蓋然的にしか保証できませんが、これまでの研究の蓄積によって、与えられたデータ量によって予測誤差を定量化できるようになってきました。これが汎化の学習理論です。汎化理論によれば予測モデルの複雑度に対して十分なデータ量があれば、汎化誤差（未知のデータに対する予測誤差）は限りなく小さくできることが知られています。これが昨今のビッグデータブームをはじめとする、巨大なデータを用いた機械学習の実応用における成功の理由の一つであると言えます。

ところが、機械学習の応用が進むにつれて、今まで想定されていなかった予測モデルの挙動が顕在化してきました。例えば、モデルに入力されるデータに対して人間が気づかないような微小な細工を加えることで、第三者が予測結果を恣意的に操作できてしまうという現象（敵対的攻撃）が知られています。また、モデルの学習時に用いたデータに内在していたバイアスに起

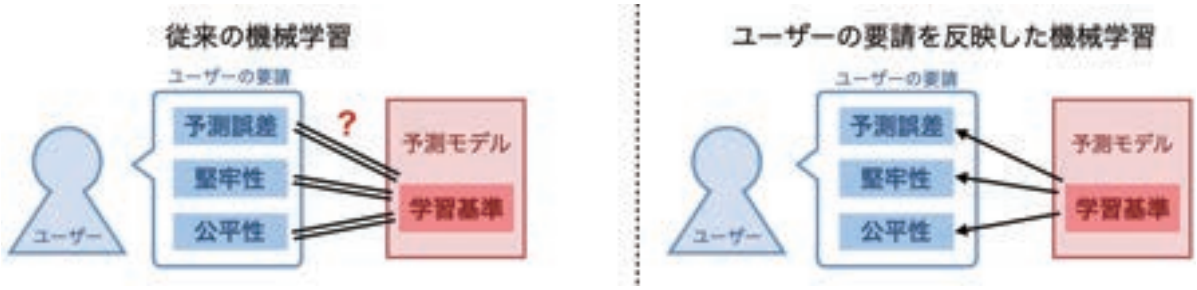

図 1　予測モデルに用いる学習基準とユーザー要請とのギャップ.

因して、犯罪率予測プログラムが人種に基づいて予測を行ってしまうような公平性の問題が報告されています。このように、ひたすらデータを収集して予測モデルに学習させればよい、というわけにはいかなくなってきたのが機械学習の直面している問題です。

学習と評価のギャップを乗り越える

なぜ私たちは機械学習の予期しなかった挙動に直面しているのでしょうか。その大きな鍵として私が注目しているのが、学習基準と評価基準のギャップです。一般的な機械学習の枠組みは、収集したデータを用いて予測性能が高くなるようなモデルを作る「学習」のフェーズと、学習したモデルを使って実際に現象を予測し、想定通りに予測できているか定量化する「評価」のフェーズにわけられます。例えば、画像分類を行いたいのであれば対象の画像をできるだけ多く集めて分類モデルを学習させて、分類正答率で性能を評価するのが自然でしょう。ところが、分類正答率が最大になるモデルを直接探すのは容易ではないため、モデルの学習時には分類正答率を元にした別の基準を使って学習を行うのが一般的です。そのため、本来私たちが期待している性能（評価基準）と、実際に学習したモデルが最適化している目的関数（学習基準）にはギャップがあるのです。さらに、評価基準を定める際に敵対的攻撃や公平性を考慮されていなければ、学習したモデルがこういった問題を解決することは非常に困難です。

私の研究では学習と評価のギャップに着目し、与えられた評価基準に対して私たちの学習基準が十分であるかを検証してきました。例えば、予測の偽陽性・偽陰性を適切にコントロールするための学習基準の設計指針を提案したり、敵対的攻撃に対して一般的な多くの学習基準が堅牢でないことを発見してきました。この流れを推し進めることによって、ユーザーが定めた評価基準に対して適切な学習基準を設計する指針を与えていき、機械学習の信頼性を向上させたいと考えています。

人間社会に馴染む人工知能を目指して

ディープラーニングの台頭により機械学習の実世界応用が急速に進むにつれて、機械学習はますますブラックボックス化する傾向にあります。私は、学習と評価のギャップを通してユーザーと予測モデルの「対話」を促し、相互に理解可能な技術へと昇華させることを目指しています。このことがひいては社会に受け入れられやすい人工知能の基盤となることを信じています。

参考文献

[1] Bao, H. & Sugiyama, M. Calibrated Surrogate Maximization of Linear-fractional Utility in Binary Classification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS2020)*, PMLR 108:2337-2347, 2020.

[2] Bao, H., Scott, C., & Sugiyama, M. Calibrated Surrogate Losses for Adversarially Robust Classification. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT2020)*, PMLR 125:408-451, 2020.