R と機械学習による社会分析：教師付き学習と因果推論/格差研究への応用

日時： 2024 年 3 月 8 日（金） 10:30～17:00
場所： オンライン開催（詳細は別途ご案内）
料金： 一般 5,000 円、学生 2,500 円
講師： 川田恵介（東京大学）
定員： 35 名 ※変更の可能性あり

■本コースの内容

本コースでは、これまで機械学習の手法に触れたことがない学生・研究者を念頭に、サーベイデータへの活用方法を紹介します。数式などは極力用いず、直感的なコンセプトの紹介と R による実装に注力します。

機械学習への関心は学際的に高まっており、実務者からも大きな期待が寄せられています。しかしながら社会科学における伝統的なデータ分析法とは、異なるルーツを有しており、実践への活用はハードルが高いと考えられがちです。実際には、伝統的な手法との融合が進む中で、多くのコンセプトが共有されており、これまでの分析を補完する便利なツールとしての活用法が多く提案されています。特に推定結果が、推定の前提となる統計モデルの定式化に強く依存してしまう問題の軽減が期待されています。このような問題は古くから指摘されてきましたが、既存の分析の多くが、実質的に目をつぶってきました。機械学習の活用は、この古くて厄介な問題について、実用的かつ包括的な解決策を提示できます。

具体的には、(1) 教師付き学習 (ランダムフォレストや OLS との Stacking モデルなど)を用いて、条件付き平均値関数を"学習"する方法、(2) 社会集団を特徴づけるパラメータ (格差や因果効果など) 推論への応用、を紹介します。(1)は教師付き学習のそもそも関心である”予測モデル"構築にそのまま応用できます。(2)は、伝統的なセミパラメトリック推定と合わせて活用することで、信頼区間の計算などより信頼できる推定が可能となります。このため社会科学が伝統的な関心としてきた研究課題について、機械学習の活用が可能になります。

さらに実際のデータを用いて、R を用いて実装(Super Learner/grf パッケージ等を使用)する方法も解説します。こちらも今まで R に触れたことがない受講生を想定します。機械学習や R は、学び始めるハードルが高い、と考えられがちです。また学際的に発展した手法の宿命として、手法を学んだものの実践段階で混乱している研究者も散見されます。しかしながらしっかりと内容を絞って順序立てて習得すれば、そのハードルは他の手法や言語と比べて決して難しいものではありません。

なお「R による格差と因果効果の推定: 識別の対比と推定方法の共有」と合わせて受講していただくことで、因果効果や格差の定義をしっかり踏まえた上で、機械学習を活用することが可能になります。幅広い学生・研究者の参加をお待ちしております!!!

■次のような方におすすめです
・初めて機械学習を学ぶ方
・機械学習の格差研究・因果推論への応用に興味がある方
・しっかりとした根拠に基づいた定量的な分析を行いたいが、数理的な議論が苦手な方

■注意事項
・どなたでも参加可能です。
・ただし、SSJ データアーカイブのデータを利用した講義の場合、利用したデータを 3 月中に削除して頂く必要がございます。
・大学または公的研究機関所属の研究者・学生（学部生も可）、SSJ データアーカイブへデータを寄託されている民間研究機関の方は、その後、研究目的で SSJ データアーカイブより申請して頂くことで利用可能です。
・R・R studio・必要パッケージのインストールを事前に済ませてください。登録・インストール方法を紹介した動画を事前に配布します。

■本コースの日程
・予測モデル・社会集団を特徴づけるパラメータ・条件付き平均値・過剰適合の整理
・教師付き学習: 回帰木系統と線形モデル系統
・平均差・条件付き平均差の推定
・条件付き平均差のノンパラメトリック推定
*進度によって内容が若干変わることがあります。