因果関係に基づく公平・高精度な機械学習の実現

近原 鷹一

NTT コミュニケーション科学基礎研究所

近年の機械学習技術の目覚ましい発展に伴い，銀行における融資承認・企業における人材採用・児童の虐待検知・罪人の釈放など，個人に対して意思決定を行う問題を機械学習予測によって実現する取組みが増えつつある．機械学習予測を導入することで意思決定の精度向上や人的コストの低減などが見込めるが，このような意思決定はその多くが人間の人生に重大な社会的影響を与えるものであり，それゆえに意思決定が性別・人種・宗教・障害・性的指向といったセンシティブな特徴に関して公平であるか否かを懸念する声が高まっている．

本講演では，数ある公平性の定義の中で最も強力とされる，変数間の因果関係に基づく公平性尺度を取り上げ，この尺度を満たす予測モデルを学習する手法を，講演者らの最近の研究 [1] を含めて紹介する．本講演は以下の 3 部で展開する．

1　機械学習と公平性 (Machine Learning and Fairness)

第 1 部では公平性を志向した機械学習の問題設定を概説する．本講演では教師あり学習の設定について考えるが，その多くは公平性制約を課しながら予測誤差を最小化する，制約付き/ペナルティ付き最適化問題として定式化される．

このような問題の定式化には，どのような予測が公平でどのような予測が差別的であるかを表す公平性尺度を，計算機が理解できる形に落とし込む必要があるが，そのためには法律 (そしてその背後にある法哲学) による差別の定義を鑑みる必要がある．その一例として，米国公民権法が定義する「disparate impact (差別的効果)」について取り上げ，これを確率変数間の独立性として定式化した公平性尺度，demographic parity (民主的公平性) について述べる．

このような変数間の相関関係に基づく公平性尺度では，センシティブ特徴と決定結果の間の因果関係を考慮できないという問題があり，特に両者の間の相関関係が交絡変数に大きく影響される場合には，差別的な因果効果が大きいにも関わらず公平であると判定してしまう危険がある [2]．これは，「差別は因果関係に基づいて判断する必要がある」と規定する多くの法律の姿勢と相反するものである．

2 因果効果入門 (Introduction to Causal Effects)

第 2 部では変数間の因果関係に基づく公平性尺度を定式化するうえで必要となる，統計的因果効果推定の諸概念について述べる．変数間の因果関係の強さを表す因果効果は，potential outcome (潜在結果) と呼ばれる確率変数の差として定義され，この potential outcome は真のデータの生成過程を表すモデルである structural equation model (SEM; 構造方程式モデル) [3] で定義される．

因果効果は因果グラフ上の各経路ごとに分解でき，NDE(直接因果効果)・NIE(間接因果効果)，その一般化である path-specific causal effect (PSE; 経路特異的因果効果) として定義される．これらは mediation analysis (媒介分析) で用いられるが，精度と公平性のトレードオフを達成するための公平性尺度の定式化に役立つ．

3 因果関係に基づく公平な予測モデルの学習 (Learning Fair Predictive Models based on Causality)

第 3 部では因果関係に基づいて公平な予測モデルを学習する方法について紹介する．因果効果に基づく既存の公平性尺度 [4, 5] のうち，path-specific counterfactual fairness (PC-fairness) [5] は，個人レベルの公平性を保証し，かつ因果グラフ上の不公平な経路を指定可能である．しかし，因果効果推定の難しさのために，既存の学習手法 [6] では「データが扱いやすい関数形から生じている（SEM の関数形が additive noise model (ANM; 加法的ノイズモデル) で表される）」という仮定を満たしていなければ PC-fairness を保証することができない．この問題を解決する学習手法として，講演者らの最近の研究 [1] を取り上げる．

参考文献

[1] Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. in AISTATS, 2021.

[2] Karima Makhlouf, Zhioua Sami, and Palamidessi Catuscia. Survey on causal-based machine learning fairness notions. in arXiv, 2020.

[3] Judea Pearl. Causality. Cambridge university press, 2009.

[4] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. in AAAI, 2018.

[5] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-fairness: a unified framework for measuring causality-based fairness. in NeurIPS, 2019.

[6] Silvia Chiappa and Thomas PS Gillam. Path-specific counterfactual fairness. in AAAI, 2019.