離散最適化ユニット
Discrete Optimization Unit

前原貴憲
Takanori Maehara

ミッション

離散構造論・離散最適化手法に基づく
機械学習・人工知能研究に活用可能な
理論・手法を開発する

コアメンバー

- 前原貴憲 (UL)
- 波多野大督
- Jean-Francois Baffier

2019年度主要成果

- 不公平なものを公平と見せかけるサンプリング
  [Fukuchi, Hara, Maehara, AAAI'20r]
- 群ラベル付き最短路問題
  [Yamaguchi, SODA'20]
- 組合せバンディットに対するリグレット改善
  [Hatano, Fukunaga+, NeurIPS'19]]
- 観測の背後に組合せ構造があるかの統計的検定
  [Ishihata, Maehara, IJCAI'19]

上記を含む全 14 本の論文誌・国際会議発表

不公平なものを公平と見せかけるサンプリング

公平な分類問題

入力：通常の分類問題と同じ．ただし特徴量にセンシティブ属性（性別・人種・他）が含まれる
出力：センシティブ属性に依存しない分類

最近の機械学習のトレンドの1つ．様々な手法が提案されてきた．

我々の問題意識

・公平な分類アルゴリズムを使っていることをどのように（社会に向けて）説明すればよいか？

不公平な分類アルゴリズムを使っている人が，自分は公平だ，と嘘を付ける方法は説明にならない
⇒ どのような設定なら嘘をつけるか？

公平性の証拠と対応する嘘

・公平性指標の数値を公開する
… 数値を改竄されると検証困難
・モデル or 全入出力を公開する
… プライバシーなど考えることが多い
・一部の入出力（ランダムサンプル）を公開する
… 一見良さそう．似たことは実用的によく行われている（代表的な人を少数人もってくるなど）
… 本当にランダムサンプルなら理論保証がつく
… サンプルをごまかされる危険性は？

問題設定 (Stealthily-Biased Sampling)

不公平な分類結果から以下を満たすサンプルを取れ：
1).公平性指標は充足する
2) ランダムサンプルと見分けがつかない

基本的な嘘のつき方

入試試験の合格者が男性100人，女性10人のとき：

男女同数ずつ合格者・不合格者をサンプルして出力すると，出力結果は明らかに公平になる

しかし，これは元の分布とサンプルの乖離が大きいので検出可能（受験者統計・在籍者統計を見ればよい）

主結果：嘘をつくためのアルゴリズム：

minimize	Wasserstein(元の分布, サンプル)
subject to	サンプルは公平，サンプル数は指定

ランダムサンプルとの「見分けのつかなさ」をWasserstein 距離を用いてモデル化し，最適化問題で定式化．この問題は**最小費用流問題**に帰着して多項式時間で解ける（およそ数千サンプル程度まで効率的にとれる）．

主結果：嘘の判別困難性

ランダムサンプルと上記アルゴリズムで得られるサンプルに「サンプルが従う分布が同じかどうかの検定(Kolmogorov-Smirnov test)」をかけて嘘を見抜くことを考える

定理：Wassetstein距離を用いてKS検定の判別可能性の上界が書ける．つまりWasserstein距離を小さくするとKS検定での判別可能性が減る．

結論・今後の課題

・ランダムサンプルを示す方法では公平性のアピールは（よほど極端でない限り）できない
・モデル自体を公開するしかない？