D-48

教師あり学習における中立化指標の f-ダイバージェンスによる統一的理解

福地 一斗(筑波大), 佐久間 淳(筑波大)

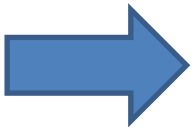
- 教師あり学習による予測から発生する差別/不公平/偏見を排除したい

$$\text{採用} = f(\text{性別}, \text{資格}, \text{住所}, \text{学歴})$$

性別と採用の強い相関 = 性別により採用を決定 **差別!**

教師あり学習

損失最小化: $\min_{f \in \mathcal{F}} R_n(f)$



中立化教師あり学習

罰則 : $\min_{f \in \mathcal{F}} R_n(f) + \eta N(f)$

制約 : $\min_{f \in \mathcal{F}, N(f) \leq \eta} R_n(f)$

相関/独立性を測る中立性指標$N(f)$によって罰則/制約を設ける

貢献

- 様々な中立性指標のf-ダイバージェンスによる統一的評価
  - CVスコア[Calders10], statistical parity[Dwork12],相互情報量 (KLダイバージェンス) [Kamishima12],共分散中立性リスク [福地14]
- 一般的な中立性指標の汎化性能を導出

MDL