SecHack365

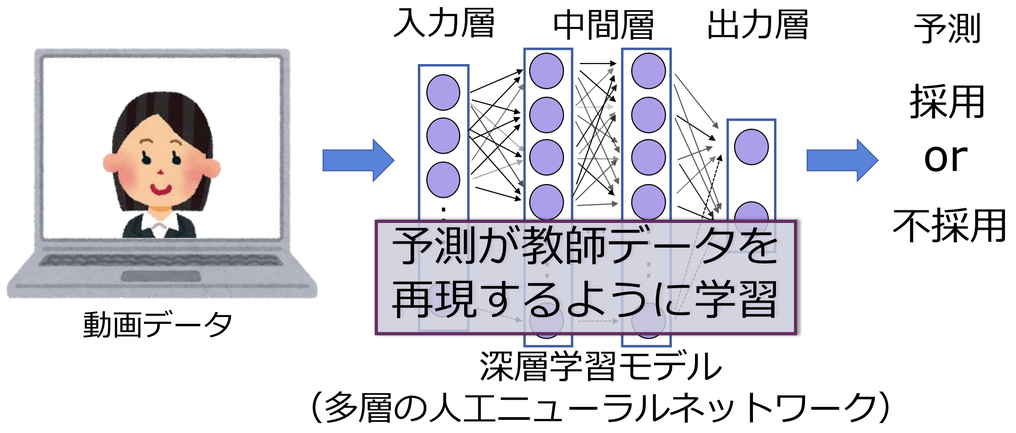
FairTorch ～深層学習の公平性ライブラリ～

思索駆動コース 福地成彦

機械学習・深層学習

2010代以降、深層学習を筆頭とする機械学習(=AI)が情報システムに組み込まれ、社会に浸透しつつある。

例) 新卒採用の動画面接の機械学習による自動化



深層学習の実装にはTensorFlowやPyTorchなどのフレームワークが用いられることが多い。

人事や与信などの機械学習システムの判断が人間の一生を左右する可能性もある。
もし、機械学習が特定の集団を差別していたとすると、
私たちは安心して情報システムに接することができない。
→**機械学習による差別はAI時代の情報システムのリスク**

機械学習・深層学習の公平性
= AI時代の**サイバーセキュリティー**

機械学習の公平性

公平性の定義：公平性を学習させるには定義が必要。

用語

- **センシティブな特徴量**：性別や人種などの判断の材料にするべきではない個人の属性
- **集団公平性**: センシティブな特徴量の集団ごとの統計に基づいた公平性の定義

集団公平性の中の2つの定義

- **人口統計公平性**: 各センシティブな特徴量の集団ごとの陽性率が等しい。
  例) 男女で採用AIの合格率が等しい。
- **均等オッズ**: 教師ラベルに該当する集団のうちで各センシティブな特徴量の集団ごとの陽性率が等しい。
  例) 採用に値する人材のうち男女で合格率が等しい。

現状の問題・課題

- 人材採用の機械学習モデルで男女差別が報告されている。
- 公平性の問題への対処のため、AIの倫理・公平性の担当部署を置く企業もある。
- 前述のどの公平性を深層学習に採用するべきかという倫理的な議論があるが、それ以前に簡単に公平性を実装する方法すらない。
  →既存の深層学習のコードを少し変更するだけで深層学習の判断を公平にできるライブラリを作りたい。

FairTorch

FairTorch: 深層学習フレームワークPyTorchで実装された深層学習モデルに公平性を導入するためのライブラリ。
PythonとPyTorchを用いて2つの公平性を実現するための3つのアルゴリズムを実装した。

FairTorch
↓ **推論結果を公平にする損失関数** ←
PyTorch で実装された深層学習モデル

公平性	アルゴリズム
• 人口統計的公平性 • 均等オッズ	• 不等式制約(L1ノルム) • 不等式制約(L2ノルム) • 敵対的学習

2 × 3 =6種類の設定から、
ユーザーが要件やデータの性質に合わせて選択可能。

実装のイメージ: 既存コードに最短2行追加するだけ！

```
dp_loss = DemographicParityLoss(sensitive_classes=[0, 1], alpha=100)
 ...(中略)...
y_pred = model(x)
loss = binary_cross_entropy_loss(y_pred, y) + dp_loss(x, y_pred, sensitive_features)
```

← 公平性の損失関数の初期化。3つのアルゴリズムで共通の引数をとる。

← 公平性の損失関数の加算。

実験: FairTorchで人口統計的公平性は改善するのか？

データ: UCI adult dataset (米国の1994年の国勢調査の一部のデータ)
入力変数: 性別、年齢、学歴などの個人情報
目的変数: 年収が5万ドル以上(陽性)か 5万ドル未満(陰性)か
モデル: 2層全結合ニューラルネットワーク
比較条件: ① **FairTorch**なし, ② **FairTorch**(不等式制約), ③ **FairTorch** (敵対的学習)

表. **FairTorch**の有無の条件ごとの男女の陽性率

条件/陽性率	男性(%)	女性(%)	男女差(%)
①FairTorchなし	20.5	14.0	**6.5**
②不等式制約	4.3	2.3	**2.0**
③敵対的学習	8.7	4.5	**4.1**

FairTorchありの場合(②, ③)では男女の陽性率の**差**が縮小
→**人口統計的公平性**が改善された

今後の展望と課題

- **FairTorch**を機械学習の社会実装をしているエンジニアに使ってもらえるように、ドキュメント等の整備や広報、コミュニティ活動を行う。
- 安定的に公平性を深層学習モデルに導入できるようアルゴリズムの改善(バッチサイズの依存性の排除など)をする。

NICT 国立研究開発法人 情報通信研究機構