Who's Responsible? Jointly Quantifying the Contribution of the Learning Algorithm and Data

Gal Yona, Amirata Ghorbani & James Zou

Data Shapley [GZ19]

Specifies four natural conditions for an equitable data valuation \( \varphi: \{x_1, \ldots, x_n\} \rightarrow \mathbb{R}^n\):
1. Null player receives zero value.
2. Symmetric players receive equal value.
3. Sum of values is \( \varphi(A) \).
4. Linearity.

A classic result from cooperative game theory (Shapley33): there exists a unique \( \varphi \) that satisfies these properties! But: data-centric (doesn't take into account A)...

This work: Extended Shapley [YGZ19]

Fix a benchmark algorithm B and add the algorithm A as an "additional" \( n + 1 \) player.

Specify five natural conditions for an equitable data-algorithm valuation \( \varphi: \{A, x_1, \ldots, x_n\} \rightarrow \mathbb{R}^{n+1}\):
1. Null data receives zero value; if A is identical to B, algo receives zero value.
2. Symmetric players receive equal value.
3. Sum of values is \( \varphi(A) \).
4. Linearity.

Existence of a unique solution extends! These are the Extended Shapley values.

Q: recognizing the interaction between existing biases in data and different (potentially subtle) modeling choices, can we disentangle their effect on the overall performance?

Joint data-algorithm valuation problem

A: A reduction to the data valuation problem (recently studied e.g. in [GZ19], [ADS19], [JDW+19]). Specifically the approach in [GZ19].

#2: Allocating responsibility for unfairness

- Training data: 1000 images from LFW-A dataset (imbalanced: 21% female, 5% black).
- Performance measure: maximal accuracy gap among groups (WM, WF, BM, BF) on the balanced PPB dataset.
- Algorithm A: Logistic regression applied to 128-dimensional feature vectors obtained by passing the images through a ResNet-V1 pre-trained on CelebA.
- Benchmark B: a constant classifier (perfectly fair: \( \varphi(B) = 0 \)).

\[ \varphi_A(D) = 22.9 \text{ (WM-BF)} \]

\[ \varphi_B = 22.1 \]

contact: gal.yona@gmail.com