Ying’s Presentation: Stop Explaining Black Box Models for High Stakes Decisions

Complex Tasks

1. High Stake Decision Making Problem
    a. black box models aren’t clear about explanations
    b. Explainable: understand how model works
    c. Accuracy vs Interpretability
2. Trade off between accuracy and interpretability for structured data
3. Explanations not faithful to what the original model computes
    a. Explanation does not mimic calculations made by the original model
    b. Example of COMPAS recidivism model: predict criminals
4. Does not provide enough detail to understand the original model
    a. Leave out too many details and information
    b. Saliency Maps: not how relevant information is being used
5. Risk Assessments
    a. calibrate information into these models
6. Human errors: largely complicated black box models
    a. it’s difficult to troubleshoots

Issues with Interpretable Models

1. Companies can’t make profit out of transparent models
    a. prevent from being reverse engineered
    b. argument: transparency would improve the quality of the system
2. Efforts to construct: computation and domain expertise
    a. solve application-specific constraint problems
    b. accountability
3. Responsible ML Governance
    a. NO black box models deployed if there exist interpretable model with the same level of performance
    b. Organization report accuracy of interpretable modeling methods
4. Algorithmic Challenges in Interpretable ML
    a. Logical Models
    b. Decision Trees
    c. Computationally hard problem to optimize problems to solve
    d. Exploration of search space
5. Construct Optimal Sparse Scoring Systems
6. Define Interpretability for Specific Domains and Create Methods
    a. Accurate Interpretable Models
    b. Rashmon Set: set of reasonably accurate prediction models

Conclusion:

1. Shift focus from assumption black box model is necessary for accurate predictions

2. Encourage policy makers to accept black box models blindly
3. Poor decisions throughout high stake areads
4. Black Box Models:
    - i. Private
    - ii. Interpretable

Discussions:

1. Black Box models:
    - a. huge parameters space
    - b. should it be avoided to avoid misleading biases?
2. Equality and Fairness:
    - a. Maybe not the model fault, it’s the data (real world) fault that is bias?
    - b. Should have the same outcome
    - c. Find Balance in the decision making process
        - i. who held being responsible if model is being used
    - d. Interpretability vs Explainability
        - i. Non-standardized terms in the two
        - ii. Olah’s paper → explainability?
            1. Taking an Interpretable Framework for Black Box Models
            2. Contrast with Olah’s paper: Interactive Visualization for interpretability
        - iii. Maybe don’t exclude the black box models, it may be not interpretable now but with continuous research it’ll be distilled and become more interpretable in the future?
    - e. Assessing Fairness:
        - i. External: What the features affect the output
        - ii. Sensitive Attribute: decorrelate the attributions