Data Systems Group

Institute of Computer Science, University of Tartu

XAI-Spring2020 Project Plan
5th March 2020
Abdul Wahab

Scope: A **C**omprehensive **Q**ualitative and **Q**uantitative **E**valuation of **M**odel-**A**gnostic **L**ocal **I**nterpretability **M**ethods.

Phase 1: Benchmarking of Tabular Interpretability methods on 3 BioMedical Tabular Datasets	
Task	**Comments**
1. Consolidating Metrics + Implementing New Metrics	Consolidating all existing and relevant metrics for interpretability + implementing andy new metrics that suit the Tabular Dataset type Interpretability.
2. Dataset and Model setup	Gathering the relevant datasets and setting up the model and interpretability methods on these datasets.
2. CLEAR Benchmarking	Understanding the approach behind this approach method and doing the benchmarking experiments
4. BreakDown Benchmarking	// //
5. D-LIME Benchmarking	// //
6. Counterfactual Explanations Guided by Prototypes	// //
Phase 2: Benchmarking of Text Interpreatbility Methods on 3 Datasets	
7. Consolidating Metrics + Implementing New Metrics	Consolidating all existing and relevant metrics for interpretability + implementing andy new metrics that suit the Tabular Dataset type Interpretability.
8. Dataset and Model setup	Gathering the relevant datasets and setting up the model and interpretability methods on these datasets.

9. POLAR	Understanding the approach behind this approach method and doing the benchmarking experiments
8. Text Dataset Interpretability Method 2	// //
9. Text Dataset Interpretability Method 3	// //
Phase 3: Benchmarking of Above Methods on Images	
7. Consolidating Metrics + Implementing New Metrics	Consolidating all existing and relevant metrics for interpretability + implementing andy new metrics that suit the Tabular Dataset type Interpretability.
8. Dataset and Model setup	Gathering the relevant datasets and setting up the model and interpretability methods on these datasets.
10. Counterfactual Explanations Guided by Prototypes	Understanding the approach behind this approach method and doing the benchmarking experiments
11. Contrastive Explanation (Foil Trees)	// //
12. Convex Density Constraints for Computing Plausible Counterfactual Explanations	// //