Explainability

Intrinsic explainable/Self-explanatory

- Explainable models
  - Linear models
  - Decision trees
  - Rule-based decision sets
- Explainable mechanisms — Attention
- Leverage intrinsic explainable models/ mechanisms
  - Add constraints to increase explainability
    - Sparsity
    - Monotonicity
    - Disentanglement
  - Surrogate model — LIME — GraphLIME

Instance-level

- Gradient
  - CAM, Grad-CAM
  - Sensitivity Analysis(SA), BP

  Originally for images, later directly applied to graphs as well
- Perturbation
  - Score function
    - Ablation (occlusion)
    - Bivariate association
    - Shapley value — SHAP
  - Selector (how to perturb)
    - Search
      - MCTS — SubgraphX
      - Greedy selection — ZORRO
    - Mask learning
      - Model predict mask
      - L2X

      Different mask generation method and objective, e.g. MI for L2X

Model-level

- Feature permutation
- Input optimization (data synthesis)
  - Saliency map (images)

    Realistic regularizers
    - Total variation norm
    - Prior from generative model: VAE or GAN
  - Graph generation via RL — XGNN

Data-level

MCI

Appications

- Model validation
- Knowledge discovery
  - Graph pattern mining
  - Ex. pneumonia risk prediction
  - Ex. toxic molecule identification (node-level)

Challenges

- Design
  - Artifacts
    - Non-realistic
    - Adversarial
  - Faithfulness — Uncertainty
    - Uncertainty of estimating Shapley value
    - Uncertainty across multiple models
  - Human-friendly
    - Contrastive — Ex. x is better than y
    - Selective — Ex. top 3 reasons
- Evaluation
  - No standard criterion
    - Application-grounded
    - Human-grounded
    - Functionally-grounded

    Metrics
  - No benchmark
  - How to evaluate faithfulness