Interpretable Machine Learning: Definitions, Methods, Applications

Presenter: Arshdeep Sekhon

https://qdata.github.io/deep2Read

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-As, Bin Yu

July 2019

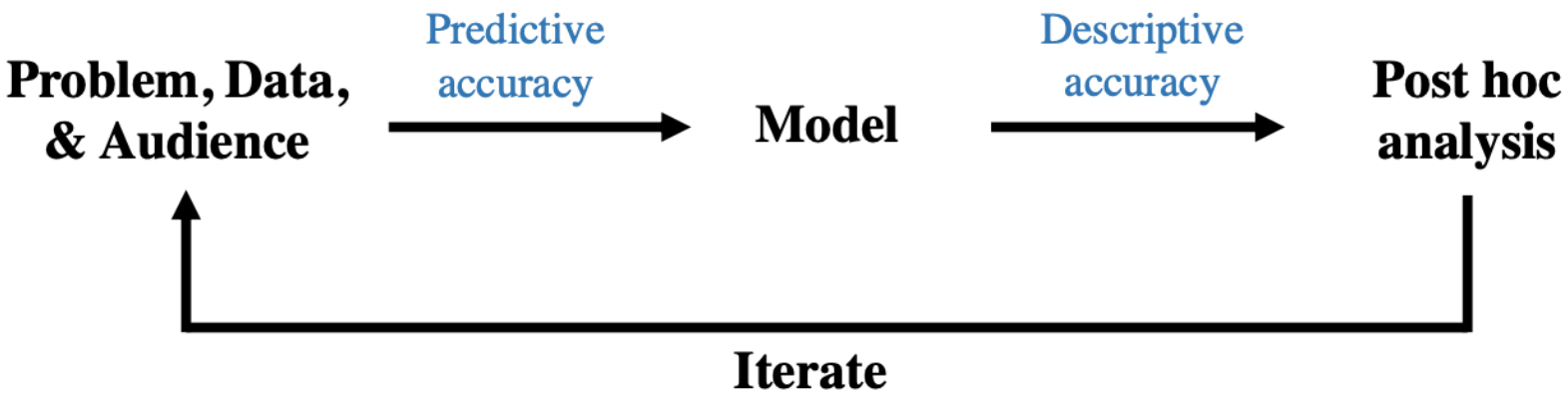
Introduction

- Interpretation methods relationships as part of the data-life cycle
- A framework to evaluate Interpretable Models: P(redictive) D(escriptive) R(elevancy) Framework

Interpretability

- Definition: Use of ML models to extract relevant information about domain relationships in data
- relevant information: insight into a domain problem for an audience
- Related to Causal Inference
  - change in one variable causes change in another
  - Extract "associative" relationships in ML models and check if causal
- Stability:
  - Stable results to small perturbations to data/model/initializations

Interpretability in Data Life Cycle

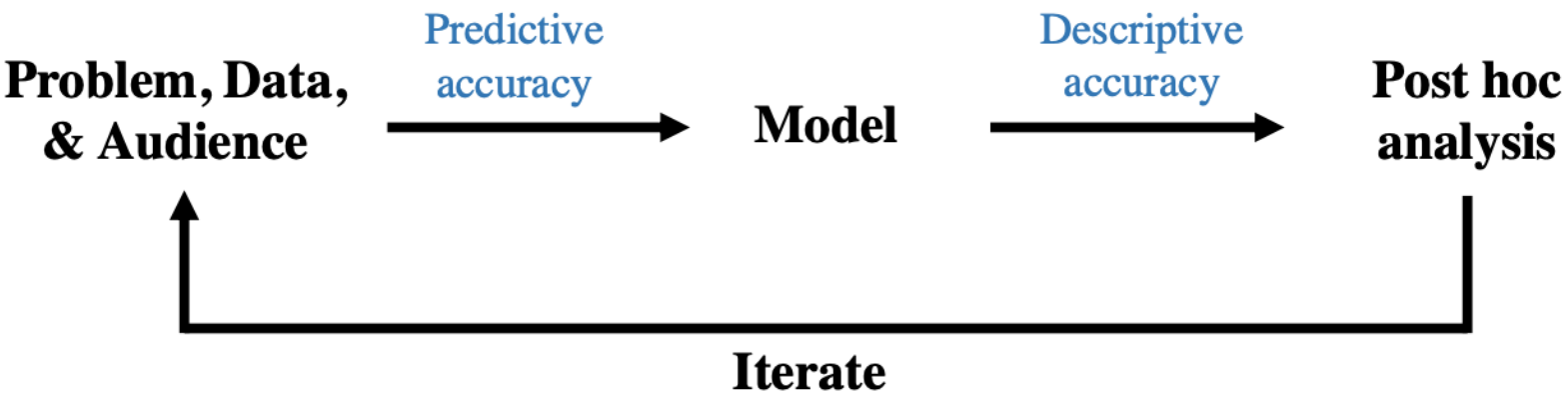


- Interpretation at Modeling Stage
- Interpretation at Post Hoc stage

PDR framework: How to select an interpretation model

- Accuracy
  - Predictive Accuracy: Model should be faithful to the underlying process, bad predictive accuracy other things learnt from the model likely to be incorrect
  - Descriptive Accuracy: degree to which model captures relationships learnt by the model
- Relevancy : should give domain user the relevant knowledge
- Model Based Interpretation affects both predictive as well as descriptive accuracy, whereas post-hoc methods only affect descriptive accuracy

Categorization of Interpetability Models



- Model Based
- Post Hoc Analysis

Model Based Interpretability

- Simpler and hence lower predictive accuracy
- Types of Model Based Interpretability
  - Sparsity
  - Simulatability
  - Modularity
  - Domain Based Feature Engineering and Model Based Feature Engineering

Model Based Interpretability: Sparsity

- limiting the number of non zero parameters
- for linear models, the non zero parameters are meaningfully related to the prediction, but hsould be stable
- In general, a good way because
  - fewer parameters: improves accuracy
  - prior knowledge regarding sparse assumption: more relvancy
  - increases prediction as well as descriptive accuracy
- particularly useful for high dimensional features, identify key ones to visualize easily

Model Based Interpretability: Simulatability, Modularity

- Simulatability: Decision Trees: Model can simulate and reason about its decision making process
- Modularity: A portion of the model can be interpreted, for example a weak one is attention model

Model Based Interpretability: Domain Based and Model Based Feature Engineering

- Domain Based: more good and informative features make relationship learning easier
- These features increase relevancy: more meaningful to an audience

Model Based Interpretability: Domain Based and Model Based Feature Engineering

- Domain Based: more good and informative features make relationship learning easier
- These features increase relevancy: more meaningful to an audience
- Model Based Feature Engineering: unsupervised learning to get better features, for example, dimensionality reduction

Post Hoc Interpretability

- analyze trained model without changing model
- Two types
  - Dataset Level: global general relationships learnt by model
  - Prediction Level: feature importance scores and interactions

Post Hoc Interpretability: Dataset Level

- Feature Importance Scores
- Statistical Feature Importance: Confidence Values(p values) on learnt coefficient
- But not always good: imply association but not causation
- Example:Initial reports by Harvard's Office of Institutional Research used logistic regression to model the probability of admission using different features of an applicant's profile, including their race: found that the coefficient associated with being Asian (and not low income) had a coefficient of -0.418 with a significant p-value (< 0.001). This negative coefficient suggested that being Asian had a significant negative association with admission probability. In contrast, the expert report supporting Harvard University finds that by accounting for certain other variables, the effect of race on Asian students acceptance is no longer significant.

Post Hoc Interpretability: Dataset Level

- visualizing filters, regularization paths, etc
- Analyzing trends and outliers in prediction: for example, checking residual plots for linear regression

Post Hoc Interpretability: Prediction Level

- how individual predictions are made by a model, can be aggregated to give dataset level interpretations
- For example, Feature Importance Scores
  -

Open Problems in Interpretation

- No consistent ways to evaluate: PDR, computational cost,etc
- Measuring Descriptive Accuracy: no consistent method, most papers show hand picked examples, one way is showing a simulation model, specify a generative procedure and show near perfect generalization accuracy
- Measuring relevancy score: Good ways are amazon turk and using model interpreations directly in experiment design