Data Science Evaluation Workshop

Evaluation Break Group

Breakout Agenda

- Challenges
  - Workflow
  - Ground Truth
  - Other
- Evaluation
  - Measures of Effectiveness
  - Workflow Needs
  - Ground Truth
  - Other
- The Future

Domains

- Environment
- Health
- Defense
- Citizen Science
- Finance
- Security and Communication
- NLP

While the desire is for general characteristics, the details were a devil.

Data Quality

- Alignment when fusing data
- Provenance
  - Model results used as "data"
- Missing data, anomalous data
- Accuracy
  - Fitness for use
  - Granularity and sensitivity of measurement
- Temporal aspects
- Knowledge of operational definition and collection instrumentation
- Contrast with experimental data collection
- Tools for looking at the data collected via monitoring programs

Ground Truth

- What is ground truth? How is it established?
  - Dependent upon fitness for use
  - Needs to be assessed in the context of use.
- Human judgements
  - how can you use this to create ground truth in a large data.
  - What is meant by Truth when it is formed by opinions? No objective reality
- Using models of reasonableness
- Just need to be better than other models – need to compare and validate
- Selecting a source of ground truth when many and competing sources might exist. Ex., Legal issues with respect to establishing ground truth as in financial industry
- Calibration given various sources of input – Measurement reliability and validity
- Conceptual definition of the dependent variable and its operationalization

Measures of Effectiveness

- Run raw data through the system and compare computed answer to the actual data

- Information Entropy

- "Properly" weighting false positives and false negatives when assessing the model

- This is cast at a meta-level – need to know the task. Need to understand the use case.

- Conflicting Measures of Effectiveness – financial advising, different segments have different objectives

Workflow

- Obtaining data from multiple sources and their alignment to focus on the problem to be addressed
  - Schema alignment

- Assumptions and knowledge about the data creation or collection process (.e.g, measurement methodology, sampling)
  - Repurposing of data

- How to anonymize data (co-reference problem)? How to work with big data? Generating synthetic data.

The Future

- Need platforms for sharing data and align across dataset
- Better ability to use of simulation and synthetic data to establish ground truth
- Curate data through expert inputs