An Introduction to Honest_ML - Why Confidence Intervals Matter

By Eric Schles

About

Principal Applied Research Scientist @ Johns Hopkins University Hospital & Adjunct Professor

2nd Year PhD Student @ CUNY Graduate Center

Contact

Twitter - @EricSchles

Linkedin - https://linkedin.com/in/ericschles

Shameless Plug: looking for New Masters DS students @ CUNY Graduate Center

https://www.eventbrite.com/e/open-house-for-masters-program-advanced-certificate-in-data-science-tickets-419128583877

Issues With Statistical Models

- Goals of Statistical Modeling
  - Generalization
  - Representing the data well
- How do you tell whether a sample is representative of your population?
- The issue with point statistics in metrics
  - Random seed problems
    - Side effects
  - Model specific issues

Goals Of Statistical Modeling

- Representation - how well does the model represent the data you have available?

- Generalization - how does the model perform on out of sample data?

Representativeness of a sample?

Considerations:

- Underlying distribution
- Sampling process
- Size of population (approximately)

Problems with Statistical models

- The issue with point statistics in metrics
  - Random seed problems
    - Side effects
  - Model specific issues

Evidence:

https://github.com/EricSchles/randomness_experiments/blob/master/scikit-learn-experiments-breast_cancer.ipynb

- Varying seed in train test split
- Varying seed but keeping it constant in train test split

Honest_ml

- Library motivation
  - Problems this solves
- Introductory example
- Why everything should have confidence intervals

demo

https://github.com/EricSchles/honest_ml/blob/main/examples/simple_example.ipynb

https://github.com/EricSchles/honest_ml/blob/main/examples/pipeline_example.ipynb

https://github.com/EricSchles/honest_ml/blob/main/examples/confusion_matrix_testing.ipynb

https://github.com/EricSchles/honest_ml/blob/main/examples/experiments.ipynb

https://honest-ml.readthedocs.io/en/latest/