An Introduction to Honest_ML - Why Confidence Intervals Matter

By Eric Schles

Contact

Twitter - @EricSchles

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

What are statistical models?

- Distributions & Data
- Descriptive Statistics as a first approximation
- Linear Regression
- Tree Based Models
- Neural Networks

Distributions & Data

Introduction to distributions and Data:

https://github.com/EricSchles/datascience_book/blob/master/1/Statistical%20Tests%20-%2001.ipynb

Claim

In fact, descriptive statistics are so powerful you can simulate data knowing only the descriptive statistics and the underlying distribution that are faithful representations of the data generating process:

https://github.com/EricSchles/honest_ml/blob/main/conference_talks/Specifying%20multiple%20parameters.ipynb

Linear Regression

https://github.com/EricSchles/datascience_book/blob/master/2/An%20Introduction%20to%20Regression%20-%2003.ipynb

- MLE
- Linear regression picture

Trees

Tree based models answer a different question - What if you just forget about trying to figure out the underlying distribution and just put data into buckets. Does this work?

https://github.com/EricSchles/datascience_book/blob/master/4/An%20Introduction%20To%20Information%20Theory%20-%2005.ipynb

- Algorithm
- Trees picture

(Yes, sometimes)

Neural Networks

https://github.com/EricSchles/datascience_book/blob/master/5/An%20Introduction%20to%20Neural%20Networks%20-%2007.ipynb

Vanilla Neural Networks:

Neural Networks answer a similar question to linear regression, except with backpropagation, which is model stacking in the forward pass and parameter update in the backpass. Additionally a non-linearity is (often) added on top of the model at each step.

Structure aware Neural Networks:

- Convolutional Neural Networks - add a learned kernel of 'correlations' (sort of correlations because a convolution is a cross correlation flipped about the y axis)
- LSTM Recurrent Neural Networks - add a notion of time dependence in the data, account for state and then prune to the 'best' representation (based on the metric)
- Transforms - sample the data _very_ intelligently during training

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