Using HTC to Understand and Communicate Appropriate Uses of Cross-Validation in Psychological Science

Hannah Moshontz and Sarah Sant'Ana

Outline

- Background
- Our study
- HTC use
- Progress and lessons

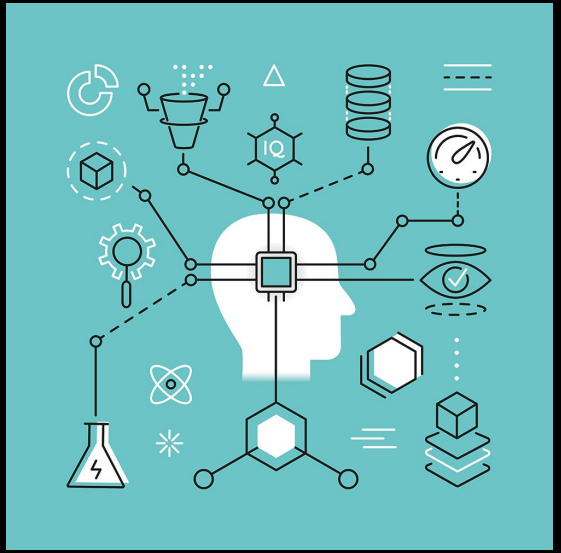
Background

Explanation vs prediction

Incorporation of machine learning techniques has allowed psychology to shift its focus

Can consider large amount of variables (and their interactions), complex data structures and relationships

Important clinical outcomes



Machine Classification and Analysis of Suicide-Related Communication on Twitter

Pete Burnap
School of Computer Science and Informatics
Cardiff University
burnapp@cardiff.ac.uk

Gualtiero Colombo
School of Computer Science and Informatics
Cardiff University
colombog@cardiff.ac.uk

Jonathan Scourfield
School of Social Sciences
Cardiff University
scourfield@cardiff.ac.uk

Computer-based personality judgments are more accurate than those made by humans

Wu Youyou, Michal Kosinski, and David Stillwell

Identifying Autism from Neural Representations of Social Interactions: Neurocognitive Markers of Autism

Marcel Adam Just, Vladimir L. Cherkassky, Augusto Buchweitz, Timothy A. Keller, Tom M. Mitchell

Facebook language predicts depression in medical records

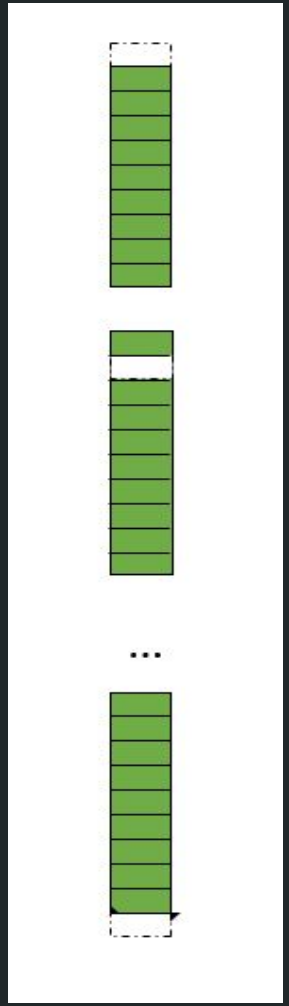
Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz

The Problem

Cross validation (CV) : A data resampling technique to assess how well results of a model generalize to new data

The problem: Psychologists use single loop CV to both select the best model and evaluate model performance

While CV estimates model performance in internal test sets, we bias results by only choosing the top performing model. **Estimates of model performance obtained with this method produce artificially lower estimates of error.**



Current study

Empirically demonstrate the bias that can occur during single loop CV across a variety of contexts

Provide examples of alternative CV techniques and demonstrate that this bias can be greatly reduced

Method

Data simulated to be totally random (i.e., the models ‘should’ do no better than chance)

Multiple contexts likely affect degree of bias

Type of CV	Classification Algorithm	Sample size	Number of variables	Ratio positive cases
Single loop k fold	GLM net	100	100	50% (balanced)
Single loop with validation	Random forest	500	1,000	10% (skewed)
Nested CV	SVM	1,000	10,000	
		10,000		

1000 simulations of each context

Workload

- Within each simulation, there are hundreds of models being fit
- Run time for one trial = 10 minutes to a week
- Time to create simulated data = up to an hour
- Multiplying processes by 1000 and doing for each study context combination
- To date, we have used over 1.5 million computing hours and expect to use 2.5 million more!

Our Strategy

- Work in progress
- Understand system and optimize resources
- UW and OS grids
- More members submitting jobs

Difficulties and Troubleshooting

- Workflow organization
- Optimization across study context
- Cross validation code errors
- First flocking attempt errors
- We are now troubleshooting masters!

General strategies

- Prevent errors through automation
- Careful file naming
- Use CHTC resources

The impact of CHTC

- Support team from CHTC has been our most valuable resource
- We could not do this kind of work without CHTC -- it would take years!
- Ability to confirm sound methodology and advance better practices