True to the model or true to the data?

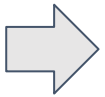
July 17, 2020

[1]Hugh Chen*, **[1]Joseph D. Janizek***, [2]Scott Lundberg, [1]Su-In Lee
[1]University of Washington
[2]Microsoft Research
*(Equal contribution)

What are Shapley values?

Example of a coalitional game:

S	v(S)
{}	0
{Ava}	1
{Ben}	1
{Cat}	1
{Ava,Ben}	2
{Ben,Cat}	2
{Ava,Cat}	2
{Ava,Ben,Cat}	3



$$\phi_{Ava}(v) = w_1(v(\{Ava\}) - v(\emptyset)) + w_2\big(v(\{Ava, Ben\}) - v(\{Ben\})\big) + w_3\big(v(\{Ava, Cat\}) - v(\{Cat\})\big) + w_3\big(v(\{Ava, Ben, Cat\}) - v(\{Ben, Cat\})\big)$$

Shapley value of i | Weight $W(|S|, |N|)$

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}\big(v(S \cup \{i\}) - v(S)\big)$$

All subsets (w/o i) | Profit individual i adds

Two common approaches for Shapley value feature attribution

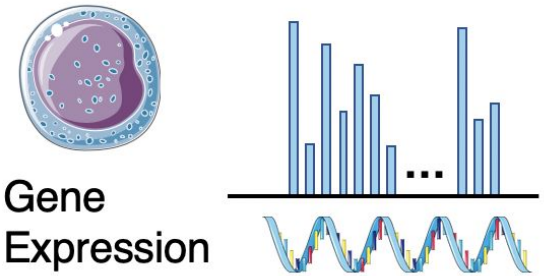
Observational conditional expectation	$v(S) = \mathbb{E}_D[f(x)\|x_S]$	Hard to compute	Stays on-manifold	Spreads credit among correlated features
Interventional conditional expectation	$v(S) = \mathbb{E}_D[f(x)\|do(x_S)]$	Easy to compute	May go off-manifold	Only gives credit to features used by the model

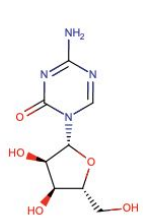
How do we resolve this problem?

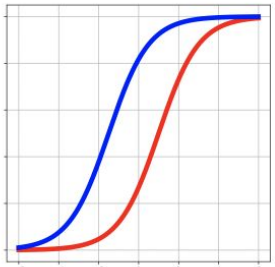
- Is one approach preferable *in general*?
- Do the differences in these approaches represent an *insurmountable difficulty* in the use of Shapley values for feature attribution?
- This problem may go away when instead of trying to find an answer for feature attribution *in general*, we try to find an answer for *specific applications* of feature attribution

Biological hypothesis generation

- Paired gene expression and drug response data
- Goal: learn which genes drive drug response

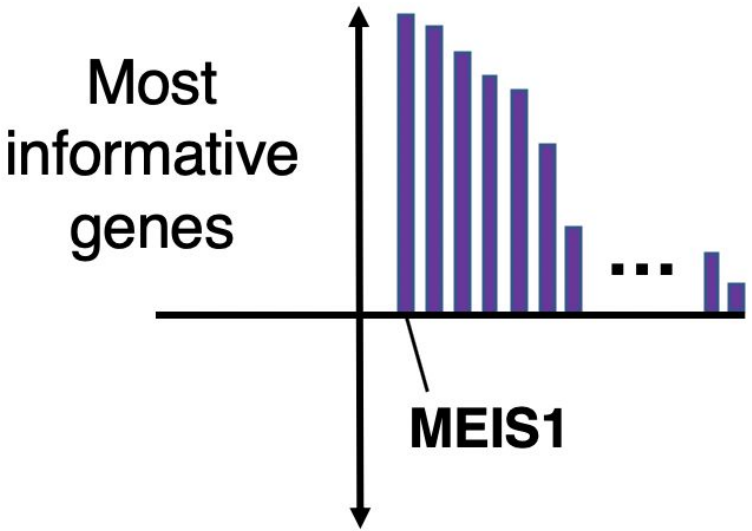






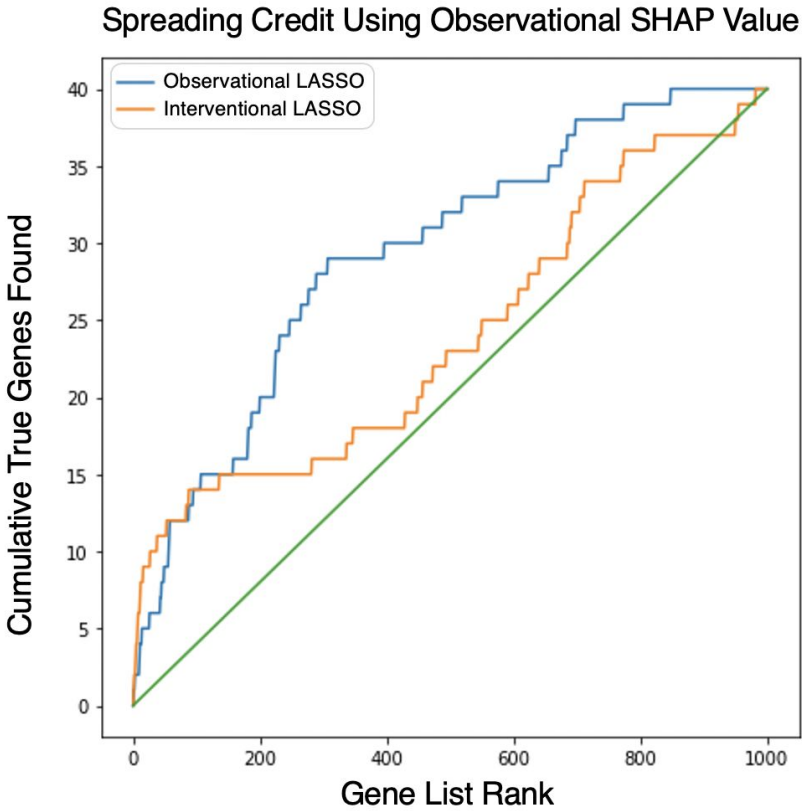
Biological hypothesis generation

- Paired gene expression and drug response data
- Goal: learn which genes drive drug response
- Train a model to predict drug response from gene expression
- Rank genes by their average magnitude Shapley value across all samples to generate list of candidates for experimental testing
- In this case, we care about the true data generating process, *not the specific model we have trained*



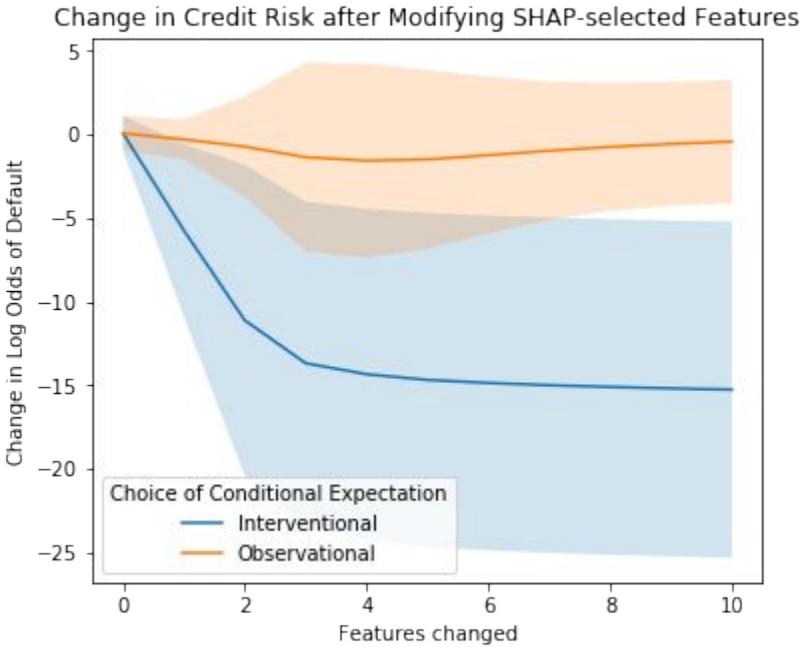
Biological hypothesis generation

- Paired gene expression and drug response data
- Goal: learn which genes drive drug response
- Train a model to predict drug response from gene expression
- Rank genes by their average magnitude Shapley value across all samples to generate list of candidates for experimental testing
- In this case, we care about the true data generating process, *not the specific model we have trained*
- Ranking genes by Observational Conditional Expectation Shapley values finds more true genes
- True to the data



Model explanation

- Model predicting loan applicant's risk of default
- Goal: learn which features most impact the output of *this particular model*
- True to the model
- Allowed each hypothetical applicant to change the features identified as most important by Interventional or Observational shapley values
- We see that the interventional Shapley values help the applicants decrease their log odds of default much more than the observational Shapley values



See our paper or contact us for more details!

hughchen.github.io
jjanizek.github.io

https://arxiv.org/abs/2006.16234