Local and Global Model Interpretability via Backward Selection and Clustering

Brandon Carter, Jonas Mueller,
Siddhartha Jain, and David Gifford

bcarter@csail.mit.edu

MIT Computer Science & Artificial Intelligence Laboratory

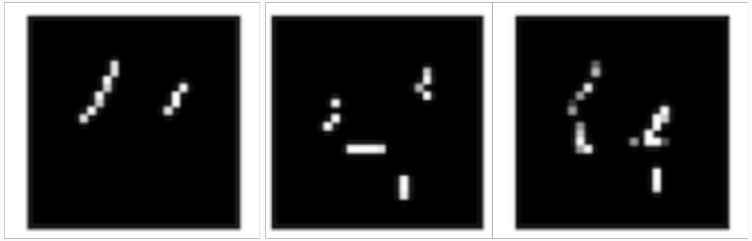
Why Interpretability?

- Adoption of neural networks and nonparametric methods has led to:
  - Large increase in predictive capabilities
  - Complex and poorly-understood black-box models
- Imperative that certain model decisions can be interpretably rationalized
  - Ex: loan-application screening, recidivism prediction, medical diagnoses
- Interpretability is also crucial in scientific applications, where goal is to identify general underlying principles from accurate predictive models

Sufficient Input Subsets

- One simple rationale for ***why*** a black-box decision is reached is a sparse subset of the input features whose values form the basis for the decision
- We propose the ***sufficient input subset*** (SIS), a minimal feature subset whose values alone suffice for the model to reach the same decision (even without information about the rest of the features' values)

4



4 4 4

SIS Help Us Understand Misclassifications

Misclassifications

5 (6)

5 (0)

Adversarial Perturbations

9 (9)

9 (4)

Formal Definitions

- Black-box model that maps inputs $\mathbf{x} \in \mathcal{X}$ via a function $f : \mathcal{X} \to \mathbb{R}$
- Each input has indexable features $\mathbf{x} = [x_1, \ldots, x_p]$ with each $x_i \in \mathbb{R}^d$
- A **SIS** is a subset of the input features $S \subseteq [p]$ (along with their values)
- Presume decision of interest is based on $f(\mathbf{x}) \geq \tau$ (pre-specified threshold)
- Our goal is to find a **complete** collection of **minimal-cardinality subsets** of features S, each satisfying $f(\mathbf{x}_S) \geq \tau$
- $\mathbf{x}_S$ = input where values of features outside of S have been masked

Algorithm

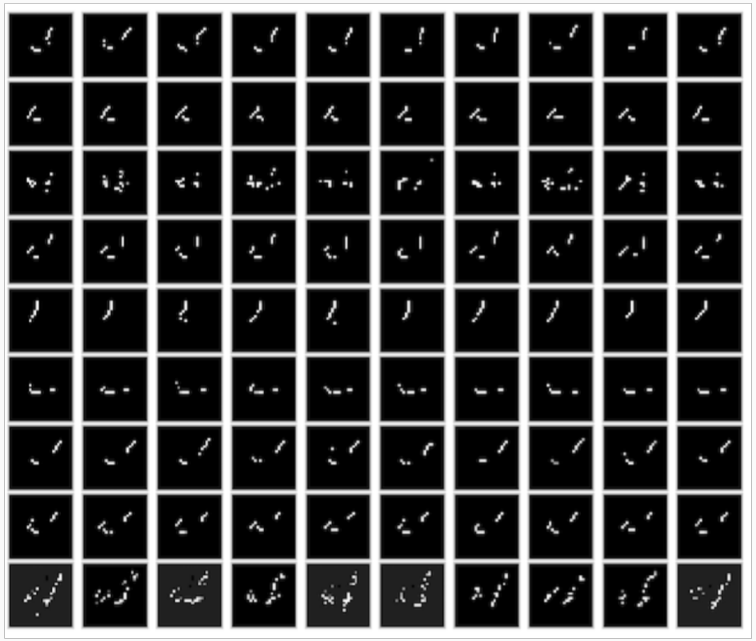
- From a particular input: we extract **SIS-collection** of disjoint feature subsets, each of which alone suffices to reach the same model decision
- Aim to quickly identify each sufficient subset of minimal cardinality via backward selection (preserves interaction between features)
- Aim to identify all such subsets (under disjointness constraint)
- We mask features outside of SIS via their average value (mean-imputation)
- Compared to existing interpretability techniques, SIS is **faithful to any type of model** (sufficiency of SIS is guaranteed), and does **not** require: gradients, additional training, or an auxiliary explanation model

SIS Clustered for General Insights

- Identifying the input patterns that justify a decision across many examples helps us better understand the general operating principles of a model
- We cluster all SIS identified across a large number of examples that received the same model decision (DBSCAN)
- Insights revealed by our SIS-clustering can be used to compare the global operating behavior of different models

SIS Clustering Shows CNN/MLP Differences

Cluster	% CNN SIS
C_1	100%
C_2	100%
C_3	5%
C_4	100%
C_5	100%
C_6	100%
C_7	100%
C_8	100%
C_9	0%



Applying SIS to Natural Language

- We use a dataset of beer reviews from BeerAdvocate [McAuley et al. 2012]
- Different LSTM networks are trained to predict user-provided numerical ratings of aspects like **aroma**, **appearance**, and **palate**

LSTMs Learn Aspect-Specific Features

on tap at the brewpub december 27 2010 pours a dark brown color with a good tan head that leaves behind a bit of lacing and sticks around for awhile the nose is really nice and chocolatey really love the level they 've used under that a bit of roasted malt but this was mostly about the chocolate the taste is n't quite as nice though the chocolate notes really still stand out the feel was quite nice with a full body pretty viscous for what it is drinks quite well i 'm a big fan

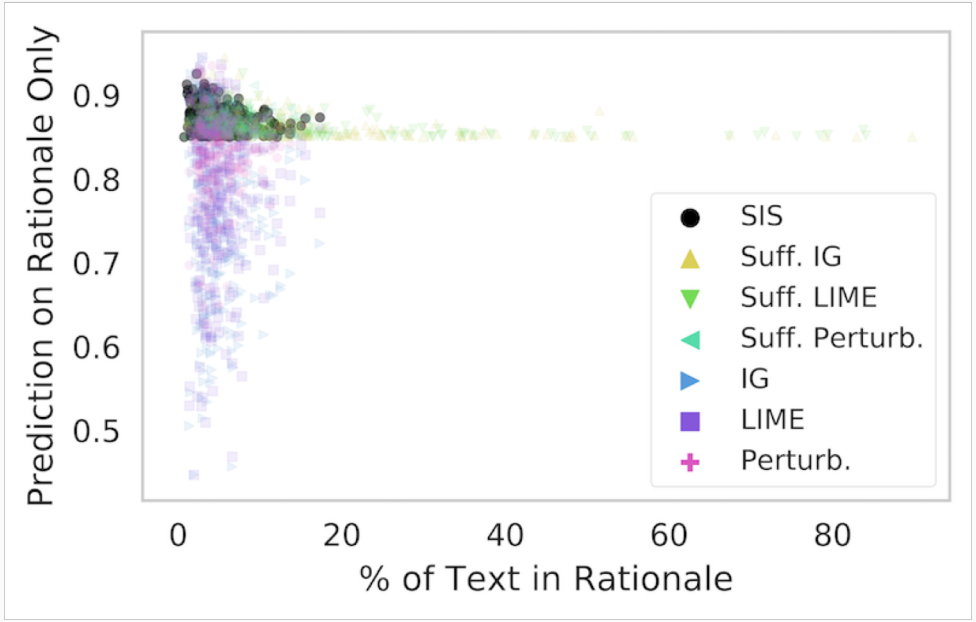
Appearance Aroma Palate

Multiple SIS in Aroma Review

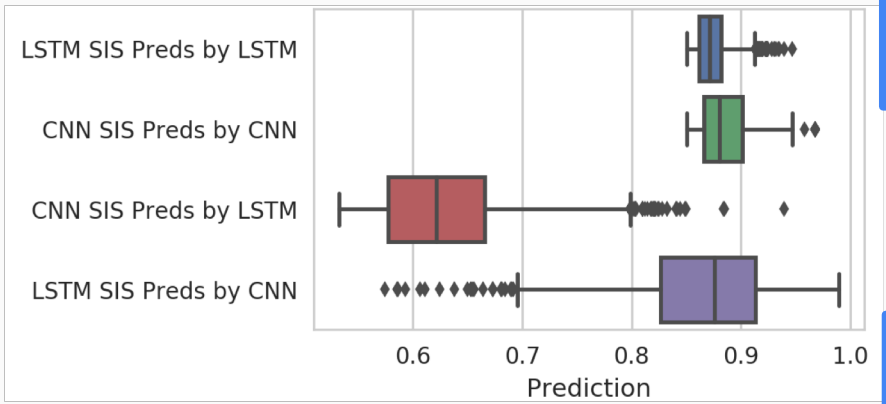
on tap at a the pour is a dark amber color bordering on mahogany with a finger 's worth of slightly off white head s wow the nose on this beer is phenomenal tons of vanilla bourbon maple syrup brown sugar caramel and toffee provide a wonderful sweetness some dark fruit notes and chocolate fill in the background of the aroma t the flavor is similarly impressive lots of sweet rich vanilla bourbon and oak accompanied by toffee caramel brown sugar and maple syrup the finish is all that prevents this from a perfect score as there is a bit of alcohol and heat but there are some nice hints of chocolate m the mouthfeel is smooth creamy rich and full bodied a light but nearly perfect level of carbonation d i was told this beer was good but i had to see for myself this is one of if not the best barrel aged barleywines i 've come across i might go back again soon to have some more

Aroma SIS 1 Aroma SIS 2 Aroma SIS 3

SIS Produces Minimal Sufficient Subsets



SIS Clustering Shows LSTM/CNN Differences



Clu.	% LSTM	SIS #1	SIS #2	SIS #3	SIS #4
C1	0%	delicious	-	-	-
C2	0%	very nice	-	-	-
C3	20%	rich chocolate	very rich	chocolate complex	smells rich
C4	33%	oak chocolate	chocolate raisins raisins oak bourbon	chocolate oak	raisins chocolate
C5	70%	complex aroma	aroma complex peaches complex	aroma complex interesting cherries	aroma complex

SIS Resources

Full paper in arXiv:
https://arxiv.org/abs/1810.03805

Code for paper and analysis:
https://github.com/b-carter/SufficientInputSubsets

Code for open-source SIS library and tutorial:
https://github.com/google-research/google-research/tree/master/sufficient_input_subsets