

Conditionally Interpretable Super Learner

Yannet Interian
University of San Francisco

In collaboration with Efstathios D Gennatas (UCSF), Timothy D Solberg (UCSF),
Mark Van der Laan (UC Berkeley), Gilmer Valdes (UCSF)

Outline

- Interpretability of Machine Learning algorithms
 - Why interpretability?
 - Interpretable Models
 - Type of interpretable models
- Conditionally Interpretable Super Learner
- Preliminary Results

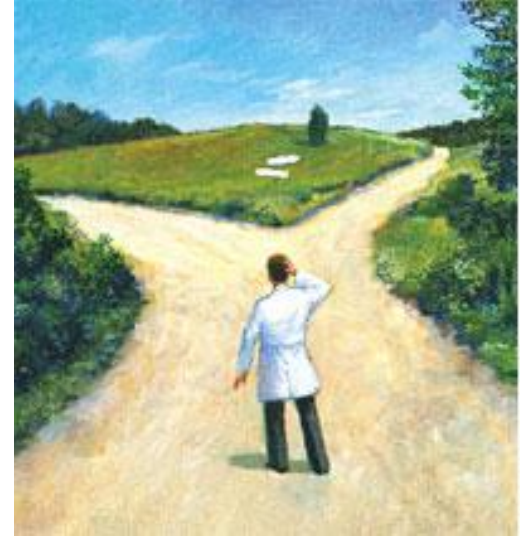
What is interpretability? Here is one definition

A model is interpretable if physicians are able to precisely describe using clinical language how the model makes predictions for every possible patient in a way that they could contest or agree with the prediction.

Why Interpretability? Acceptance

Physicians rated the ability to explain decisions as the most highly desirable feature of a decision-assisting system

[Teach and Shortliffe, 1981].



Why interpretability?

Limitation of observational data

- Case study to predict the probability of death (POD) for patients with pneumonia
- High-risk patients would be admitted to the hospital while low-risk patients were treated as outpatient

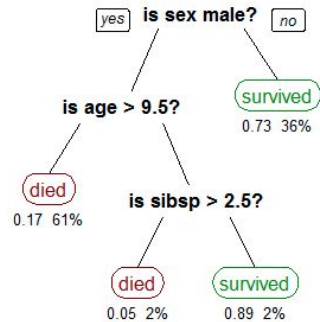
Asthma = > lower risk of dying from Pneumonia

Type of interpretable models

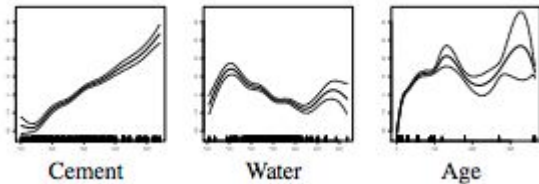
Simple models

Linear models

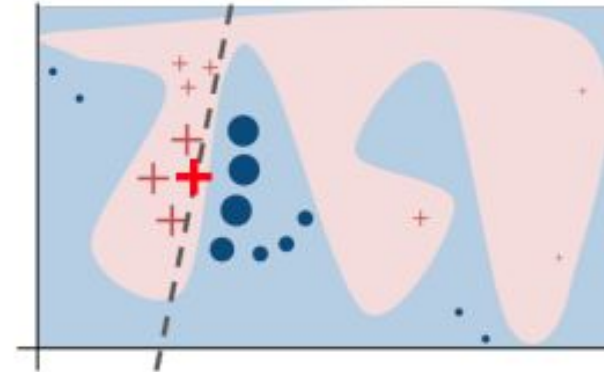
Trees



Generalized additive models



Post-hoc explanations



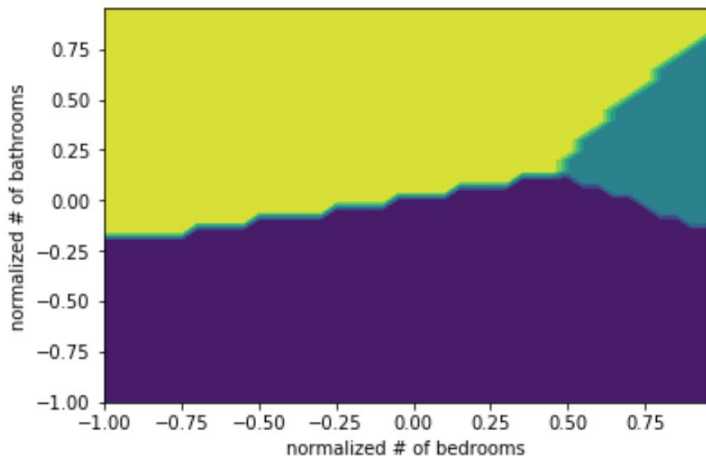
LIME Ribero et al 2016

Pieces of Conditionally Interpretable Super Learner

- **Oracle:** a model (possibly non interpretable) that partitions the feature space into groups
- **Base models:** Interpretable models
 - Lasso, Ridge, Elastic Net
 - Decision trees with `max_depth = 3,4,5`
 - Each group is modeled by a base model

Example

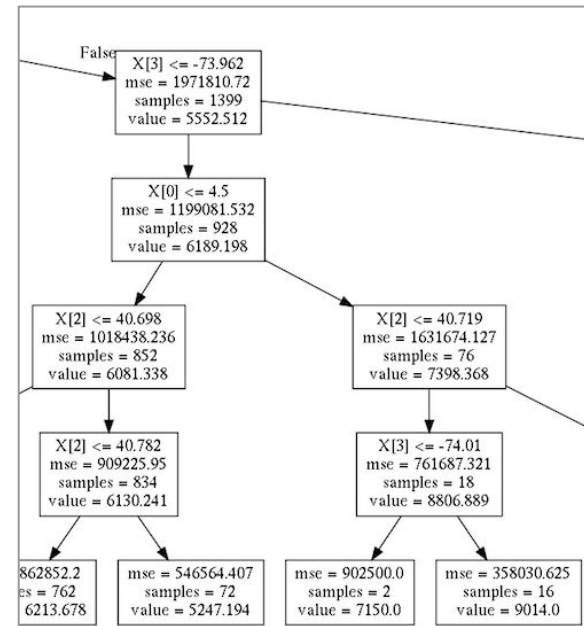
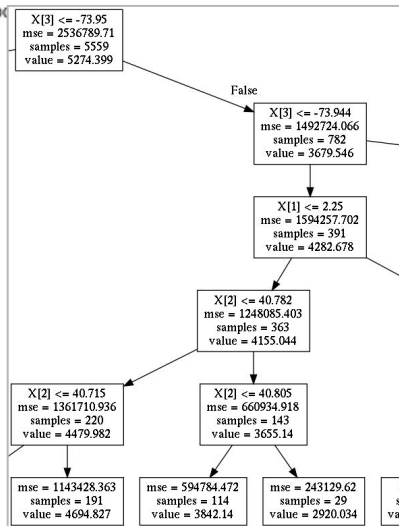
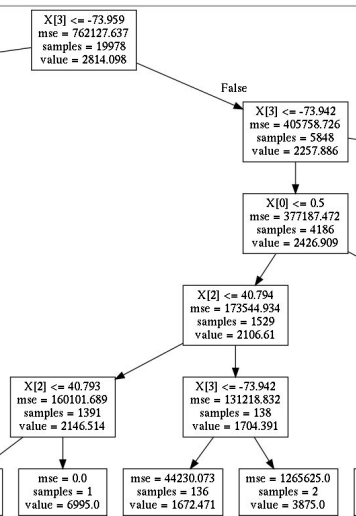
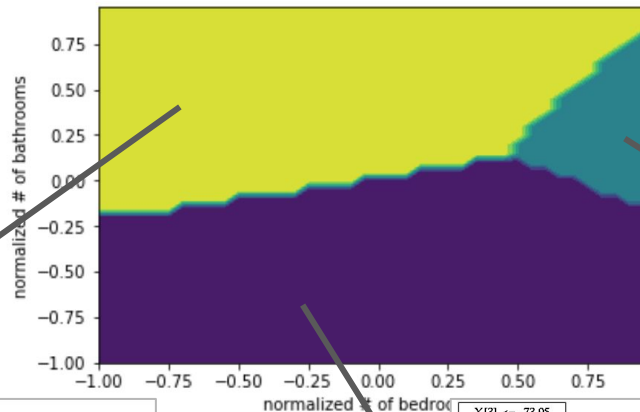
	bedrooms	bathrooms	latitude	longitude	price
0	3	1.5	40.7145	-73.9425	3000
1	2	1.0	40.7947	-73.9667	5465
2	1	1.0	40.7388	-74.0018	2850
3	1	1.0	40.7539	-73.9677	3275
4	4	1.0	40.8241	-73.9493	3350



- Oracle divides the space into 3 pieces (Neural Network).
- Each piece is modeled by as Decision Tree of depth 5
- R^2 of this model 0.67
- R^2 of best Random Forest with 1000 trees 0.71
- One tree has R^2 of 0.62

Example

	bedrooms	bathrooms	latitude	longitude	price
0	3	1.5	40.7145	-73.9425	3000
1	2	1.0	40.7947	-73.9667	5465
2	1	1.0	40.7388	-74.0018	2850
3	1	1.0	40.7539	-73.9677	3275
4	4	1.0	40.8241	-73.9493	3350

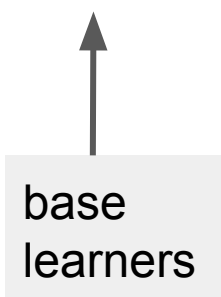


Optimization function

$$\mathcal{L}(SL) = \sum_{i=1}^N \sum_{k=1}^K 1(o(x_i) = k) L(y_i | M_k(x_i))$$



oracle



base
learners

Find oracle and base learners that minimize this loss

Loss function for Oracle

$$\begin{aligned}\mathcal{L}(SL) &= \sum_{i=1}^N \sum_{k=1}^K 1(o(x_i) = k) L(y_i | M_k(x_i)) \\ &= \sum_{i=1}^N \sum_{k=1}^K w_{ik} 1(o(x_i) \neq k)\end{aligned}$$

- Given K fixed based learners you can compute weights w_{ik} .
- The oracle is solving a weighted multi-class classification problem.

Algorithm

- Fit K base learners on a random subset of data
- Loop
 - Fit `oracle` to an “extended” dataset using loss from the K learners
 - Assign each observation to one of the K learners
 - In this step we may lose some learners
 - Refit each learner on the assigned points

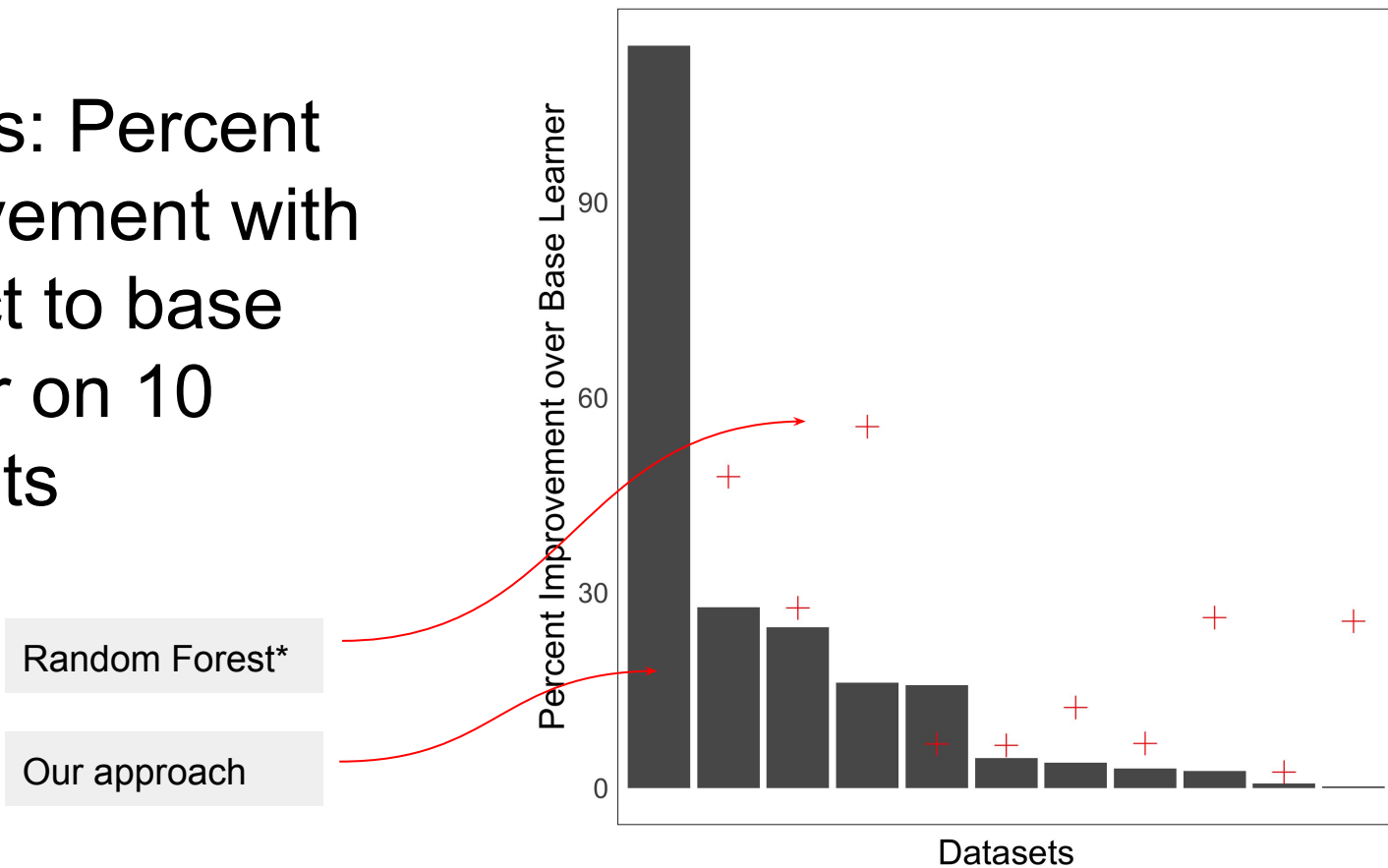
Initial results -- datasets

Selected from pmlb (Penn machine learning benchmarks repository)

- At least 2000 samples
- Non trivial (Random Forest has better R^2 than base models)
- Real world data (not synthetic)
- Regression

dataset	num_samples	num_features
1199_BNG_echoMonths	17496	9
1201_BNG_breastTumor	116640	9
1595_poker	1025010	10
201_pol	15000	48
218_house_8L	22784	8
225_puma8NH	8192	8
294_satellite_image	6435	36
537_houses	20640	8
564_fried	40768	10
573_cpu_act	8192	21
574_house_16H	22784	16

Results: Percent Improvement with respect to base learner on 10 datasets



*Random Forest uses 1000 trees hyper-parameter search (max-depth)

Summary

- Introduced Conditionally Interpretable Super Learner
 - Globally interpretable model
- Introduced a two-stage algorithm to fit the model
- Showed some preliminary results