

# Measuring and optimizing machine learning interpretability

Christoph Molnar, Giuseppe Casalicchio, Bernd Bischl

LMU Munich

DagStat 2019-03-21

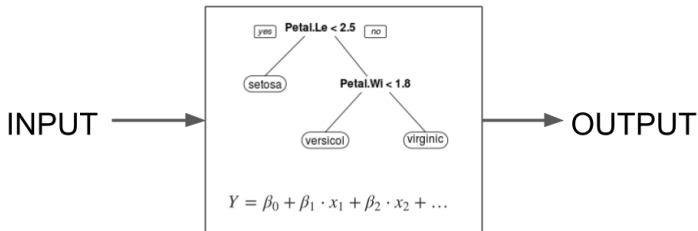
# Black Box Problem



# Problem: Tradeoff interpretability and performance

Use only interpretable models (e.g. linear model)!

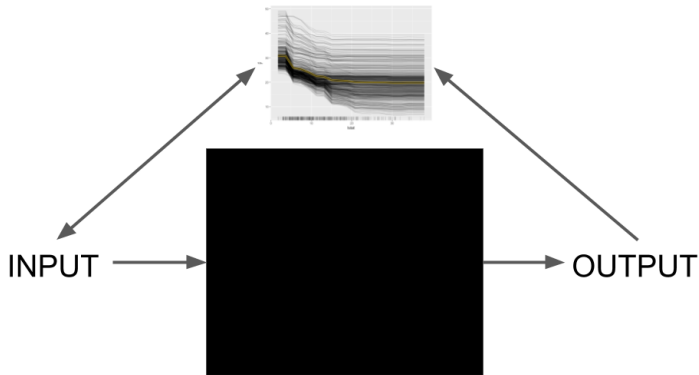
⇒ Pay with loss in performance.



## Problem: Tradeoff interpretability and performance

Just use post-hoc interpretation methods (e.g. feature importance)!

⇒ Works better for less complex models.



What's interpretability anyway?



# We propose measures of model complexity

Measure model complexity in a model-agnostic way: number of features, interaction strength, main effect complexity

⇒ Allows model comparison

⇒ Allows direct optimization for interpretability

⇒ Makes claims of interpretability more explicit (“Model A uses less features than B and has less interactions”)

# Functional Decomposition

$$f(x) = \underbrace{f_0}_{\text{Intercept}} + \underbrace{\sum_{j=1}^p f_j(x_j)}_{\text{1st order effects}} + \underbrace{\sum_{j \neq k}^p f_{jk}(x_j, x_k)}_{\text{2nd order effects}} + \dots + \underbrace{f_{1,\dots,p}(x_1, \dots, x_p)}_{\text{p-th order effect}}$$

# Functional Decomposition

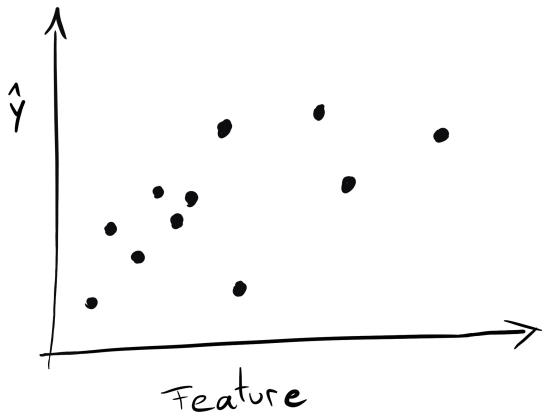
$$f(x) = \overbrace{f_0}^{\text{Intercept}} + \overbrace{\sum_{j=1}^p f_j(x_j)}^{\text{1st order effects}} + \overbrace{\sum_{S \subseteq \{1, \dots, p\}, |S| \geq 2} f_S(x_S)}^{\text{Higher order effects}}$$



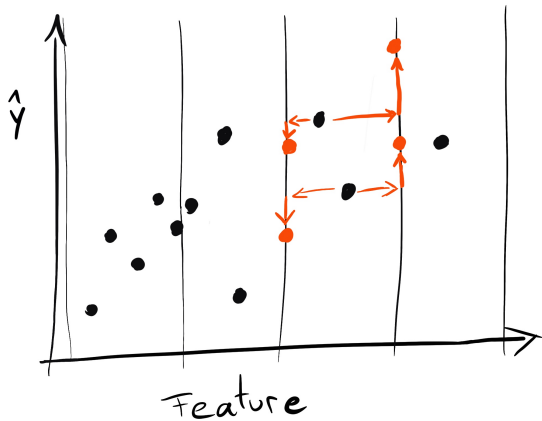
# Functional Decomposition

$$f(x) = f_0 + \underbrace{\sum_{j=1}^p \overbrace{f_j(x_j)}^{\text{How complex?}} + \overbrace{IA(x)}^{\text{How much interaction?}}}_{\text{How many feature used?}}$$

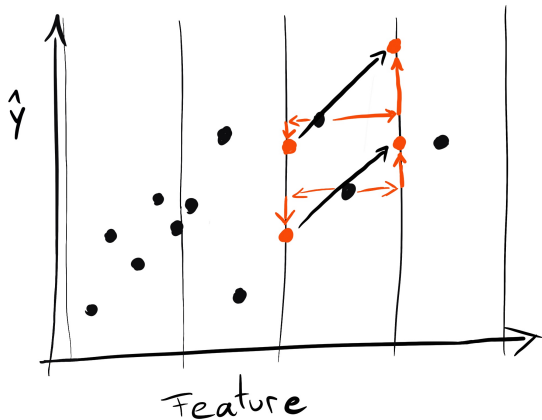
# Accumulated Local Effects (ALE)



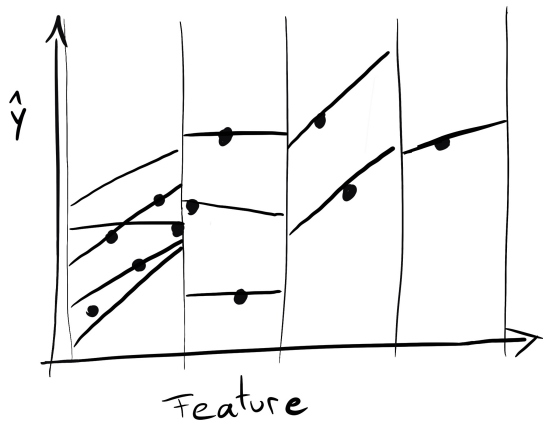
# Accumulated Local Effects (ALE)



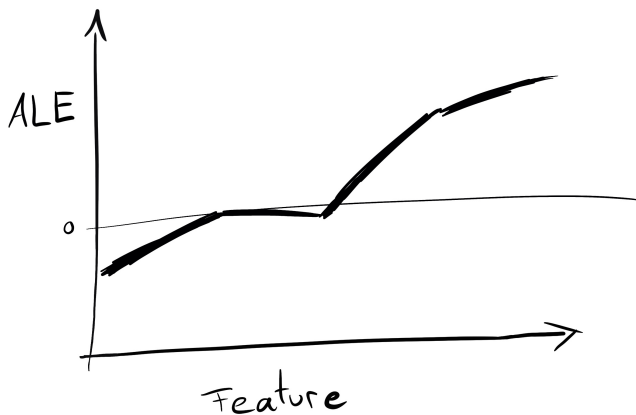
# Accumulated Local Effects (ALE)



# Accumulated Local Effects (ALE)

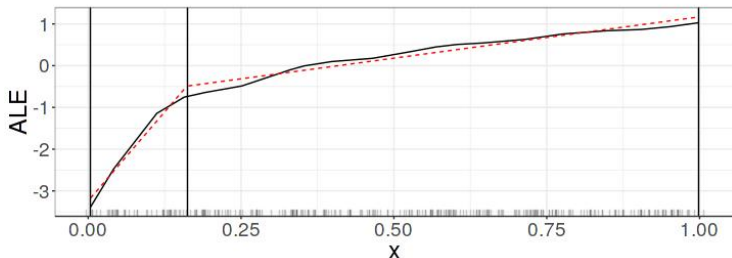


# Accumulated Local Effects (ALE)



# Main effect complexity

- ▶ Approximate ALE plot with linear segments
- ▶ Count number of non-zero coefficients
- ▶ Average over all features, weight with variance



# Interaction Strength

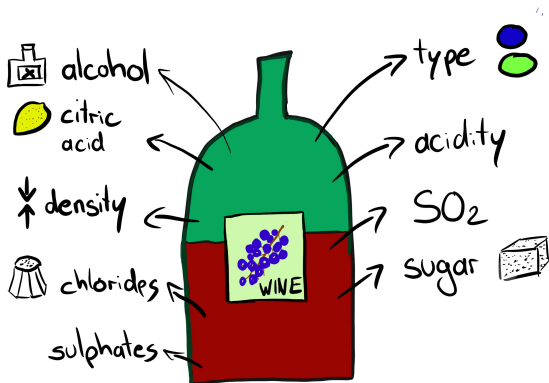
Measure main effect model with proportion of error explained:

$$\text{Interaction Strength} = \frac{\mathbb{E}(L(\hat{f}, f_0 + \sum_{j=1}^p f_j))}{\mathbb{E}(L(\hat{f}, c))}$$

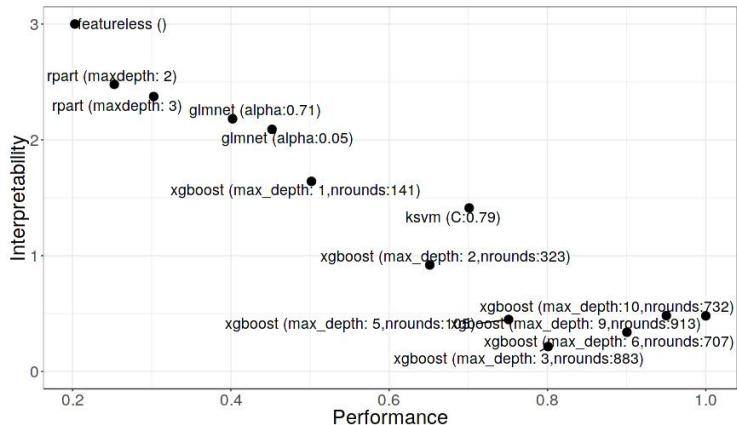


## Application: Multi-Objective Optimization













































- ▶ Predict wine quality from physicochemical properties
- ▶ Minimize MAE, number of features, interaction strength, main effect complexity
- ▶ Search across different model classes and hyperparameter settings



# Application: Multi-Objective Optimization



# Application: Multi-Objective Optimization

	xgboost (max_depth: 9,nrounds:913)	ksvm (C:0.79)	xgboost (max_depth: 1,nrounds:141)	rpart (maxdepth: 3)
MAE	0.47	0.53	0.57	0.61
NF	11	11	11	3
IA	0.64	0.27	0.02	0.14
MEC	4.12	1.99	2.90	1.81
fixed.acidity				
volatile.acidity				
citric.acid				
residual.sugar				
chlorides				
free.sulfur.dioxide				
total.sulfur.dioxide				
density				
pH				
sulphates				
alcohol				

**Table 3.** A selection of 4 models from the Pareto optimal set. From left to right, the models with best MAE, best MAE when  $MEC \leq 2$ , best MAE when  $IA \leq 0.1$ , best MAE with  $NF \leq 7$ .

# Interested in interpretable machine learning?

More on interpretable machine learning in my book  
<http://christophm.github.io/interpretable-ml-book/>.

