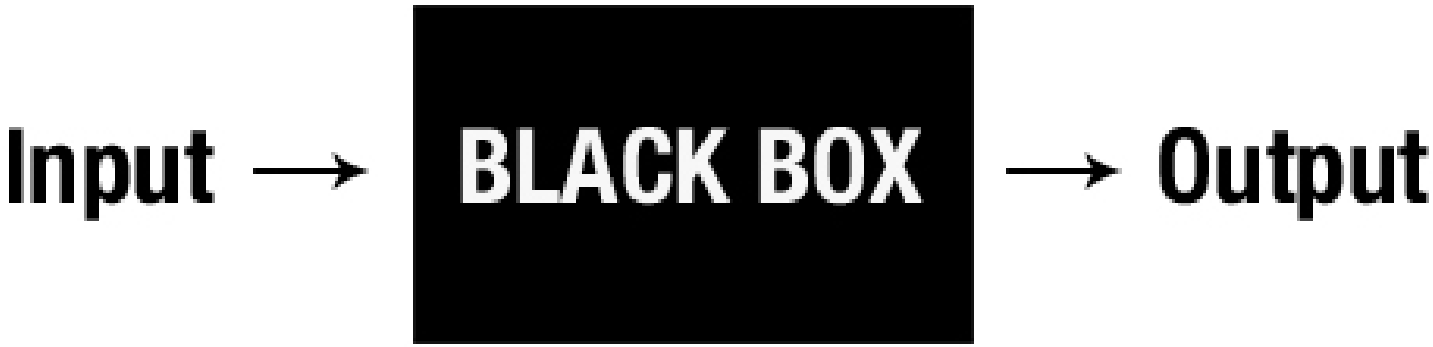
Explainable/Interpretable Machine Learning

Debashis Ghosh

February 27, 2018

One cartoon for deep learning

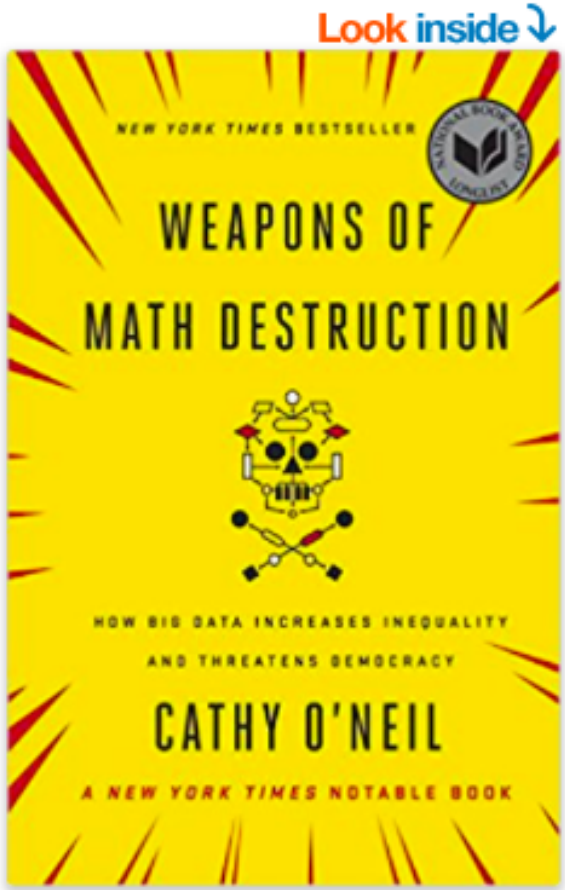


Implications

- It is hard to understand what is going on
- Training involves lots of parameters that need learning/tuning.
- Scientists/clinicians are less likely to use them in practice because they don't know what is going on

More subtle implications

- "The irony is that the more we design Artificial Intelligence technology that successfully mimics humans, the more that AI is learning in a way that we do, with all of our biases and limitations."
- See



This is being regulated!

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 16 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee ([1]),

Having regard to the opinion of the Committee of the Regions ([2]),

Acting in accordance with the ordinary legislative procedure ([3]),

Whereas:

(1) The protection of natural persons in relation to the processing of personal data is a fundamental right. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provide that everyone has the right to the protection of personal data concerning him or her.

(2) The principles of, and rules on the protection of natural persons with regard to the processing of their personal data should, whatever their

GDPR

- From `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation`
- "Citizens have rights to question and fight significant decisions that affect them that have been made on a solely algorithmic basis."
- Citizens have a "right to explanation".
- Points to the need for *transparency* in algorithmic prediction

LIME framework

- Ribeiro et al. (2016), "Why should I trust you"
- Introduces Locally Interpretable Model (Agnostic) Explainer
- Idea: assume that neighborhood around *predictions* are locally linear or admit simple representations

LIME: Conceptual overview

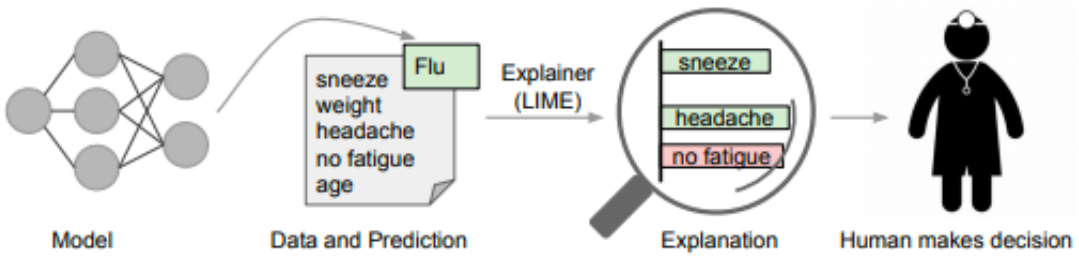


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

LIME: Conceptual overview (cont'd.)

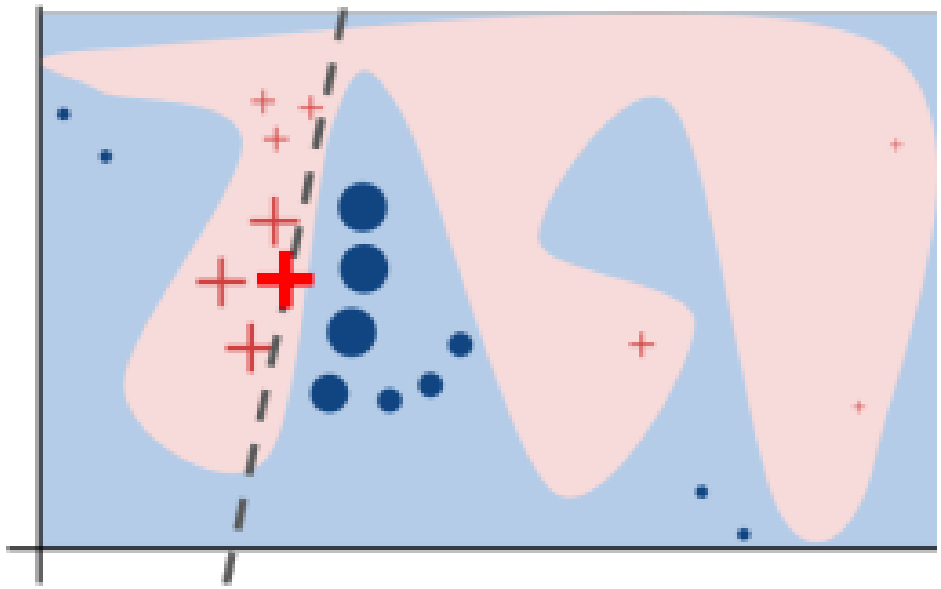


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Desirable properties in LIME

- **Interpretability:** provide qualitative understanding between input variables and the response (user-dependent)
- **local fidelity:** explanation corresponds to how the model behaves in a local neighborhood of the prediction
- **global perspective:** Identify examples/observations/instances that are representative of the full model.
- **model agnostic:** explainer should be able to explain any model (treat the original model as being black-box)

LIME: the details

- Optimization problem:

$$\xi(x) = \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

where

$$f = \text{classification rule/function}$$

$$\mathcal{G} = \text{space of functions}$$

$$\pi_x = \pi_x(z) = \text{proximity measure from an instance } z \text{ to } x$$

$$\mathcal{L} = \text{measures how unfaithful g is to approximating f locally}$$

LIME: the details

- Optimization problem #1:

$$\xi(x) = \min_{w_g} \sum_{z,z' \in \mathcal{Z}} \exp(-D(x,z)^2/\sigma^2)(f(z) - w_g \cdot z')^2 + \infty I(\|w_g\|_0 > K)$$

where

$$D = \text{distance function between points in } Z$$

$$\|w_g\|_0 = L_0 \text{ norm}$$

$$w_g = \text{weight vector}$$

K: number of nonzero terms in the model

Generic algorithm

Algorithm 1 Sparse Linear Explanations using LIME

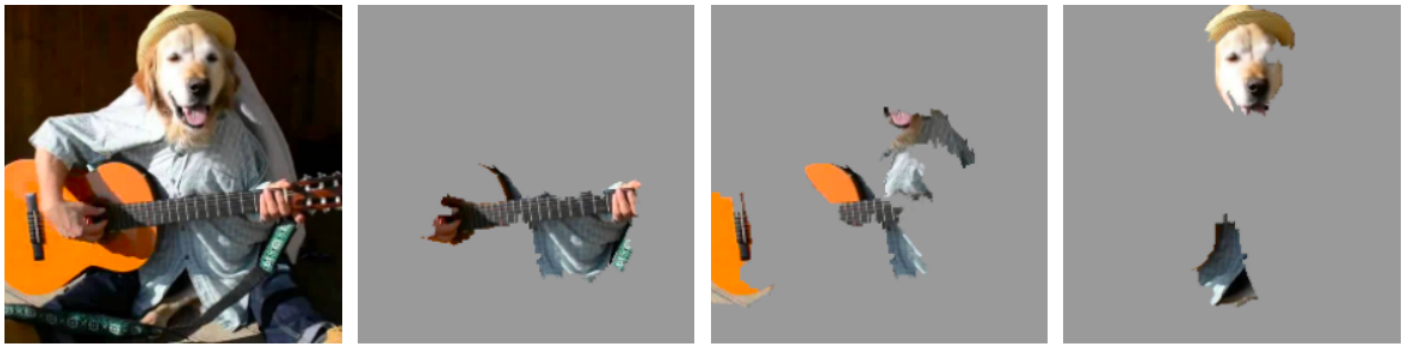
Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

Generic algorithm (cont'd.)

- Key: interpretable version of the instance
- This has to be specified by the user and will be problem-dependent
- For text data, use bag of words as the interpretable version
- For image data, use grouped clusters of pixels (superpixels)
- Example: R code (**lime** package)

GoogleNet Example



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Submodular picking

- **Goal:** pick observations that 'cover' the space but are not redundant
- Define a coverage function that enumerates important features and solve as an optimization problem find the subset of samples that maximizes the coverage
- The problem in NP-hard
- Authors propose a greedy approach

Submodular picking

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B

for all $x_i \in X$ **do**

$\mathcal{W}_i \leftarrow$ **explain**(x_i, x_i') ▷ Using Algorithm 1

end for

for $j \in \{1 \dots d'\}$ **do**

$I_j \leftarrow \sqrt{\sum_{i=1}^{n} |\mathcal{W}_{ij}|}$ ▷ Compute feature importances

end for

$V \leftarrow \{\}$

while $|V| < B$ **do** ▷ Greedy optimization of Eq (4)

$V \leftarrow V \cup \text{argmax}_i \, c(V \cup \{i\}, \mathcal{W}, I)$

end while

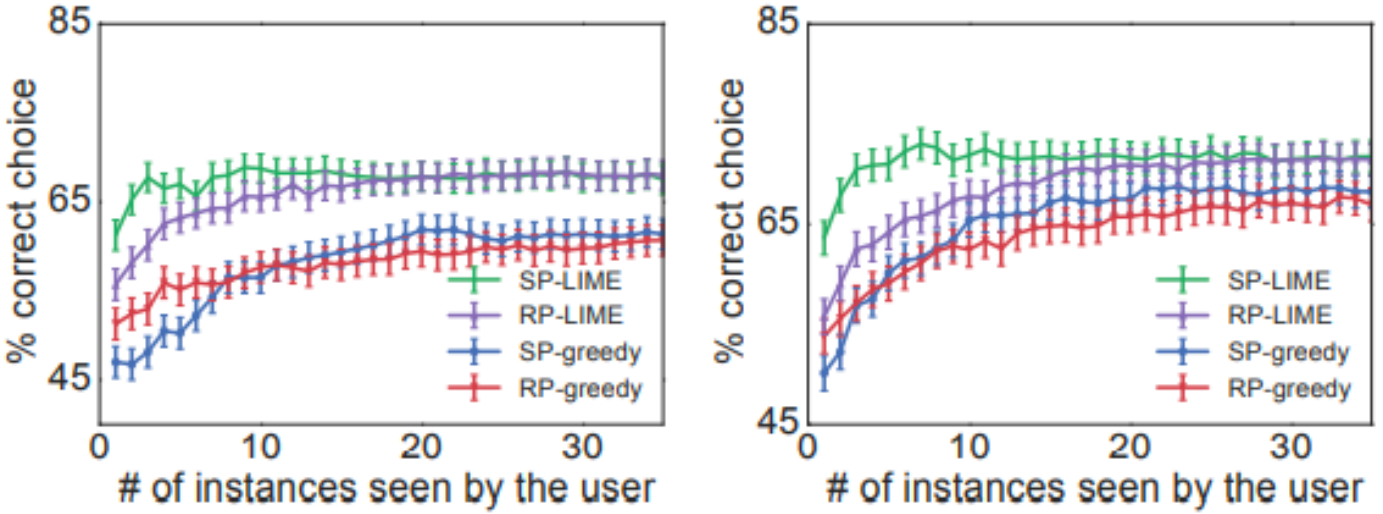
return V

Simulation results - #1

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

Simulation results - #2



(a) Books dataset (b) DVDs dataset

Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

Simulation results - #3

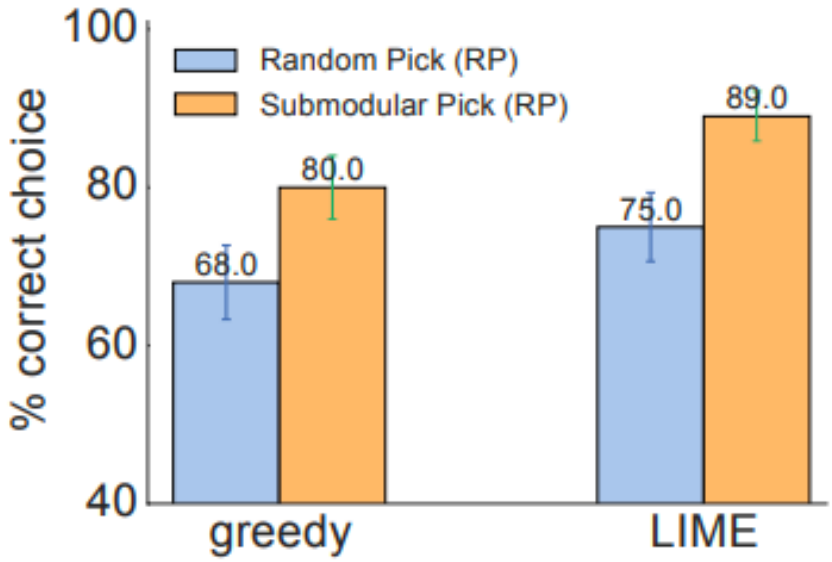


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Comments

- The modelling here is of predictions, so it is a type of 'meta-'modelling approach
- This very much relies on the linearity in the local neighborhood
- Algorithm is quite fast

Falling Rule Lists

- An alternative to LIME
- Descendant of Inductive Logic Programming as well as classification trees/recursive partitioning
- Goal: develop interpretable rules that can be formulated as IF/ELSE statements with monotonic probabilities

Example

Falling Rule Lists

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density $\geq$ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age $\geq$ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Table 1: Falling rule list for mammographic mass dataset.

Falling Rule Lists

- Components:
  1. Size of list
  2. If clauses (boolean functions on feature space)
  3. risk scores (satisfy monotonicity constraints)

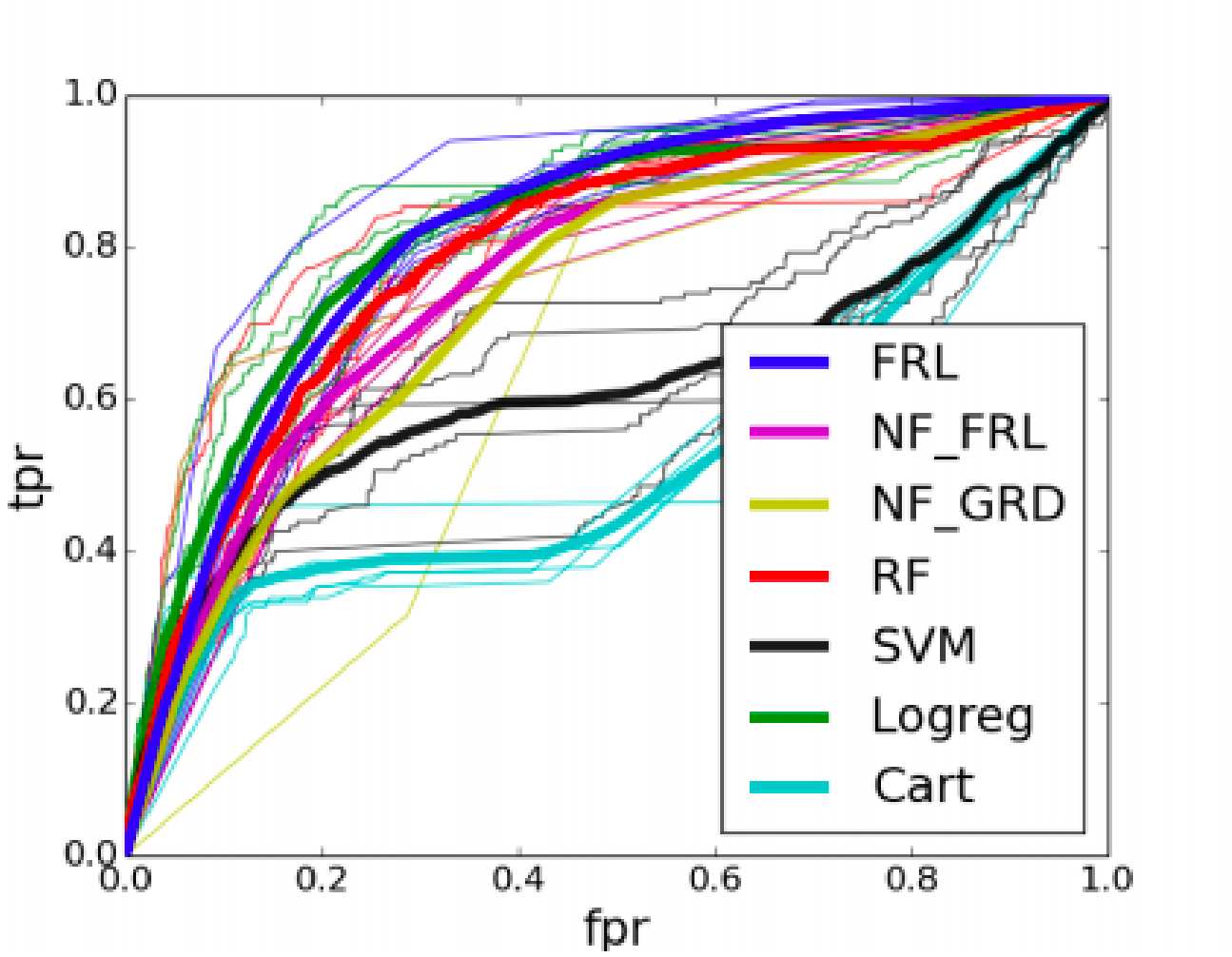
Falling Rule Lists (cont'd.)

- Can write down a likelihood based on a Bernoulli model with probability depending on risk score
- Authors use a prior on the Boolean functions and the risk scores
- Perform a reparametrization to deal with monotonicity constraints

Falling Rule Lists (cont'd.)

- Bayesian inference based on posterior of $(L, \{c_l(\cdot)\}, K,$ risk scores$)$
- Computationally intensive steps

Example


1.0
0.8
0.6
0.4
0.2
0.0
tpr
0.0
0.2
0.4
0.6
0.8
1.0
fpr
FRL
NF_FRL
NF_GRD
RF
SVM
Logreg
Cart

Conclusion

- Interpretable machine learning is an important field for making black box algorithms more understandable
- Treat the predictions as given and model those in terms of interpretable components (LIME) or start with constructing a priori interpretable rules (falling rule lists)
- This work appears to have some links to generalized degrees of freedom in statistics (Ye, 1998, JASA)