Why Should I Trust You?

Explaining the Predictions of Any Classifier

Casper Vestergaard Kristensen Alexander Munch-Hansen

November 17, 2019

Aarhus University

Outline

1. Meta information

2. Article
   - The LIME framework
   - Explaining Predictions
   - Explaining Models

3. Experiments

4. Conclusion

5. Recap

Meta information

Authors

- Marco Tulio Ribeiro
- Sameer Singh
- Carlos Guestrin

Publishing

- Conference Paper, Research
- KDD ’16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
    - A premier interdisciplinary conference, brings together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data.
    - Sigkdd has the highest h5 index of any conference involving databases or data in general
    - Highly trusted source

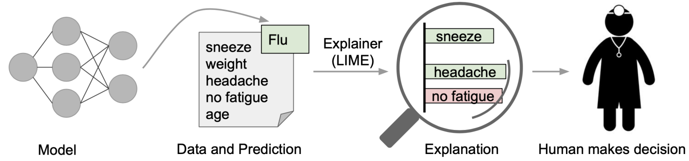
Article

Problem definition

- People often use Machine Learning models for predictions
- Blindly trusting a prediction can lead to poor decision making
- We seek to understand the reasons behind predictions
    - As well as the model doing the predictions

Problem definition

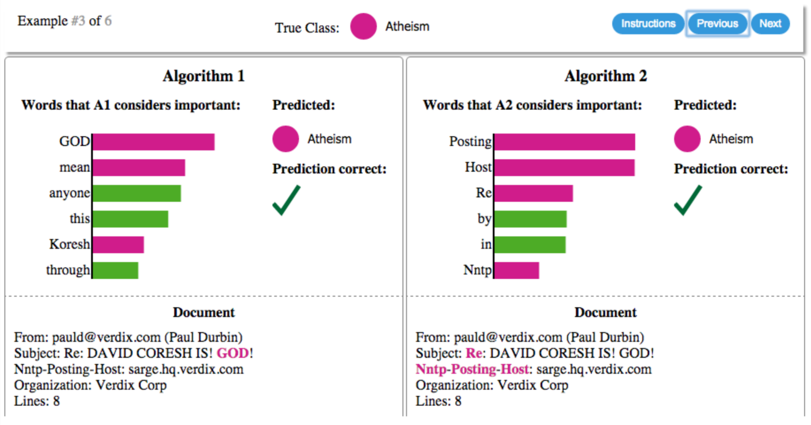
- People often use Machine Learning models for predictions
- Blindly trusting a prediction can lead to poor decision making
- We seek to understand the reasons behind predictions
    - As well as the model doing the predictions



Previous Solutions

- Relying on accuracy based on validation set
- Gestalt
- Modeltracker
    - Help users navigate individual instances.
    - Complementary to LIME in terms of explaining models, since they do not address the problem of explaining individual predictions.
    - The our submodular pick procedure of LIME can be incorporated in such tools to aid users in navigating larger datasets.
- Recognizing the utility of explanations in assessing trust, many have proposed using interpretable models
    - May generalize poorly, if data can’t be explained in few dimensions
    - So interpretability, in these cases, comes at the cost of flexibility, accuracy, or efficiency

A look into two predictions



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD**!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

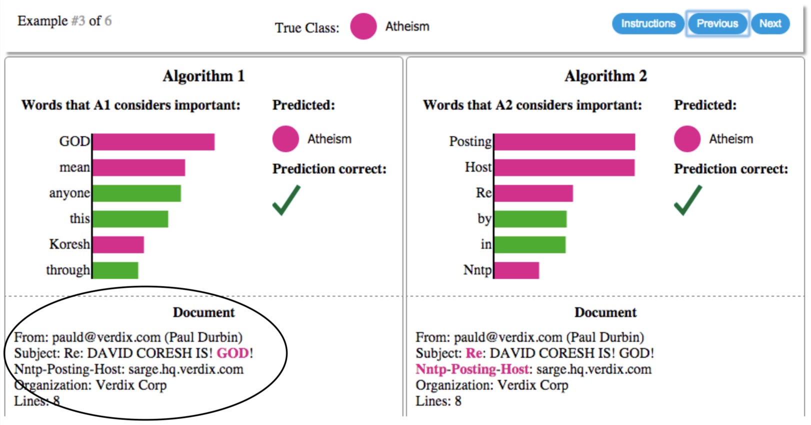
Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: **Re**: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

A look into two predictions



Example #3 of 6

True Class: ● Atheism

Algorithm 1

Words that A1 considers important:

Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD**!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

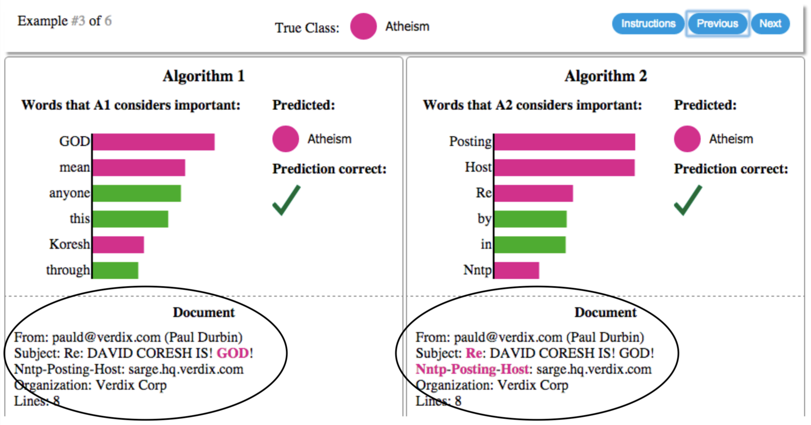
Predicted: ● Atheism

Prediction correct: ✓

Document

From: pauld@verdix.com (Paul Durbin)
Subject: **Re**: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

A look into two predictions



Document

From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD**!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Document

From: pauld@verdix.com (Paul Durbin)
Subject: **Re**: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

LIME

- The algorithm created
- Explains the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an *interpretable* model.

Properties of a good explanation

- It should be *intepretable*:
    - They must provide qualitative understanding between the input variables and the response
    - They must take into account the users limitations
    - Use a representation understandable to humans
    - Could be a binary vector indicating presence or absence of a word
    - Could be a binary vector indicating presence of absence of super-pixels in an image
- It should have *fidelity*:
    - Essentially means the model should be faithful.
    - Local fidelity does not imply global fidelity
    - The explanation should aim to correspond to how the model behaves in the vicinity of the instance being predicted
- It should be *model-agnostic*:
    - The explanation should be blind to what model is underneath

The Fidelity-Interpretability Trade-off

We want a simple explanation, still capable of displaying fidelity

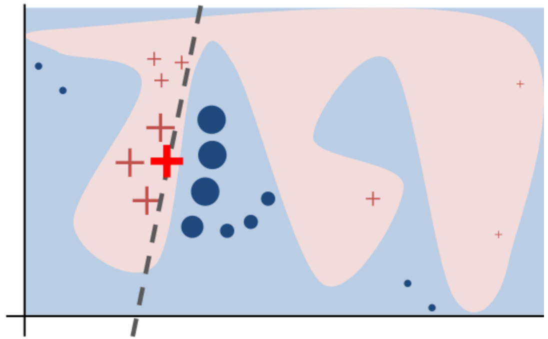
- Let an explanation be defined as a model $g \in \{0, 1\}^{d'} \in G$, where G is a class of *potentially interpretable* models
    - Linear models, decision trees
    - g is a vector showing presence or absence of *interpretable components*
- $\Omega(g)$ explains the *complexity* of an explanation g
    - Could be height of a decision tree or number of non-zero weights of a linear model
- The model we try to explain is $f : \mathbb{R}^d \to \mathbb{R}$
    - In classification, $f(x)$ is the probability or binary indicator that x belongs to a certain class
- $\pi_x(z)$ is a proximity measure between instance z and x and defines the locality around x
- $\mathcal{L}(f, g, \pi_x)$ defines how *unfaithful* g is in approximating f in the locality around π_x.
- Ensuring both *interpretability* and *local fidelity*, we minimize $\mathcal{L}$ while having $\Omega(g)$ be low as well

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

Sampling for Local Exploration

Goal: Minimizing $\mathcal{L}(f, g, \pi_x)$ without making assumptions on f

- For a sample x, we need to draw samples around x
- Accomplished by drawing non-zero elements of x, resulting in perturbed samples z'
- Given $z' \in \{0, 1\}^{d'}$, we compute un-pertubed $z \in R^d$, $f(z)$, so we have a label for z'.



More definitions

- $G =$ Class of linear models: $g(z') = w_g \cdot z'$
- $L =$ The locally weighted square loss
- $\pi_x(z) = \exp(-D(x,z)^2/\sigma^2)$
    - An exponential kernel function based on some distance function D (could be L2 distance for images)
- Thus; $L(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2$

Explaining an individual prediction

Algorithm 1: Sparse Linear Explanations using LIME

```
Require: Classifier $f$, Number of samples $N$
Require: Instance $x$, and its intepretable version $x'$
Require: Similarity kernel $\pi_x$, Length of explanation $K$
    $\mathcal{Z} \leftarrow \{\}$
    for $i \in \{1, 2, 3, \ldots, N\}$ 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
    $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ $\triangleright$ with $z'_i$ as features, $f(z)$ as target
    return $w$
```

Explaining models

Idea: We give a global understanding of the model by explaining a set of individual instances

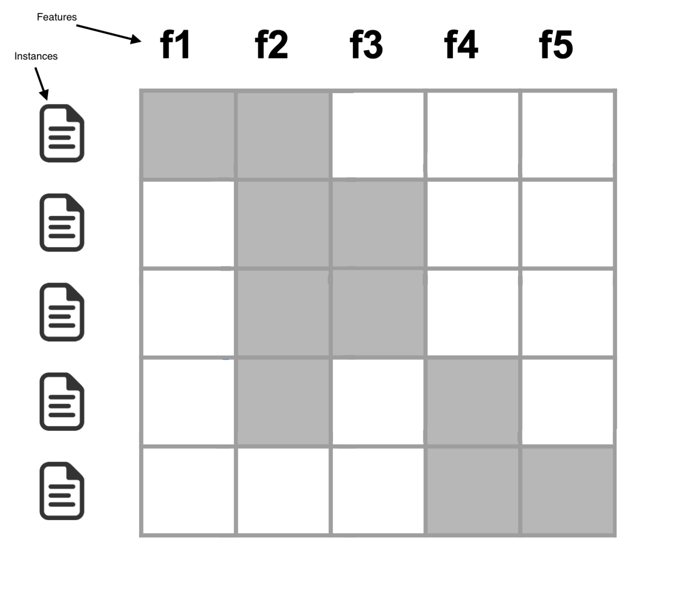
- Still model agnositc (since the indiviudal explanations are)
- Instances need to be selected in a clever way, as people won't have time to look through all explanations
- Some definitions
  - Time/patience of humans is explained by a budget *B* which denotes number of explanations a human will sit through.
  - Given a set of instances **X**, we define the *pick step* as the task of selecting *B* instances for the user to inspect.

The pick step

The task of selecting *B* instances for the user to inspect

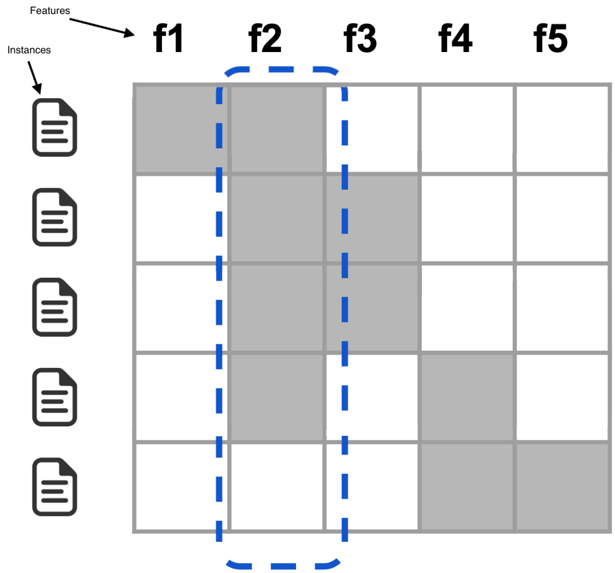
- Not dependent on the existence of explanations
- So it should not assist users in selecting instances themselves
- Looking at raw data is not enough to understand predicitions and get insights
- Should take into account the explanations that accompany each prediction
- Should pick a diverse, representative set of explanations to show the user, so non-redundant explanations that represent how the model behaves globally.

Picking instances



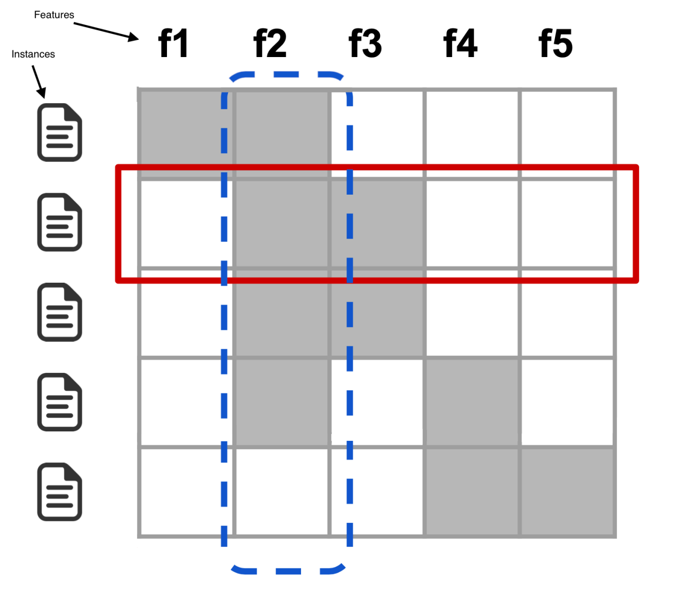
	f1	f2	f3	f4	f5
	▨	▨			
		▨	▨		
		▨	▨		
		▨		▨	
				▨	▨

Picking instances

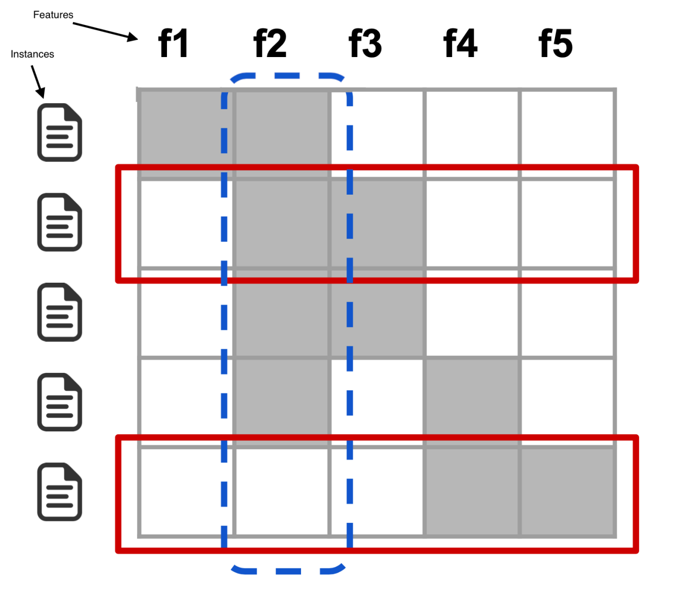


	f1	f2	f3	f4	f5
📄	■	■			
📄		■	■		
📄		■	■		
📄		■		■	
📄				■	■

Picking instances



Picking instances



Definitions

- $I_j = \sqrt{\sum_{i=1}^{n} W_{ij}}$
- $c(V, W, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : W_{ij} > 0]} \, I_j$
- $Pick(W, I) = \operatorname*{argmax}_{V, |V| \leq B} c(V, W, I)$

Explanation of algorithm 2

- Given explanations for set of instances X, $\left(|X| = n\right)$. Construct $n \times d'$ *explanation matrix* W

Submodular Picks

Algorithm 2: Submodular pick (SP) algorithm

```
Require: Instances $X$, Budget $B$
forall $x_i \in X$ do
    $W_i \leftarrow$ explain$(x_i, x'_i)$                ▷ Using Algorithm 1
end
for $j \in \{1 \ldots d'$ do
    $I_j \leftarrow \sqrt{\sum_{i=1}^{n} |W_{ij}|}$      ▷ Compute feature importances
end
$V \leftarrow \{\}$
while $|V| < B$ do                                       ▷ Greedy optimisation of Eq 4
    $V \leftarrow V \cup \text{argmax}_i\ c(V \cup \{i\}, W, i)$
end
return $V$
```

Experiments

Experiments

Interested in three questions:

- Are the explanations faithful to the model?
- Can the explanations aid users in ascertaining trust in the predictions
- Are the explanations useful for evaluating the model as a whole

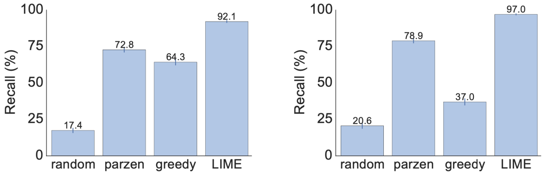
Setup

- Use two datasets, *books* and *DVDs*, both of 2000 instances.
    - Task is to classify reviews as *positive* or *negative*
- Decision Trees (**DT**), Logistic Regression (**LR**), Nearest Neighbours (**NN**), and SVMs with RBF kernel (**SVM**), all used BoW as features, are trained.
    - Also train random forest (**RF**) with 1000 trees.
- Each dataset used for training will consist of 1600 instances and 400 will be used for testing.
- Explanations of **LIME** is compared with **parzen**
    - **parzen** approximates black box classifier globally and explains individual predictions by taking the gradient of the prediction probability function.
    - Both are also compared to a greedy method where features are picked by removing most contributing ones until prediction change, as well as a random procedure.
    - $K = 10$ for the experiments

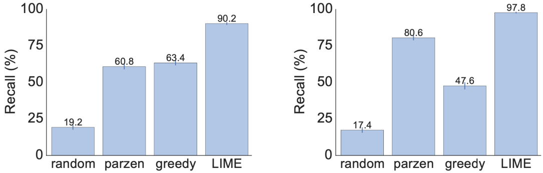
Faithfulness

- Faithfulness of explanations is measured on classifiers that are interpretable, **LR** and **DT**.
    - Both are trained s.t. the max no. of features which they can find is 10, so features found by these are the *gold standard* of features, in regards to which features are important.
- For each prediction on the test set, explanations are produced and the fraction of the gold features found, is computed.

Faithfulness



(a) Sparse LR (b) Decision Tree



(a) Sparse LR (b) Decision Tree

Should I trust this prediction?

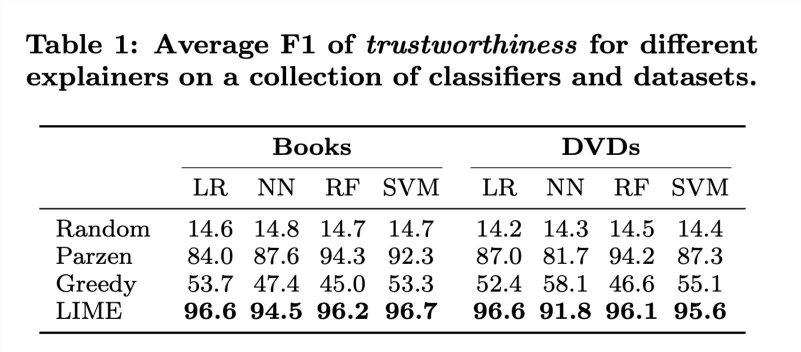


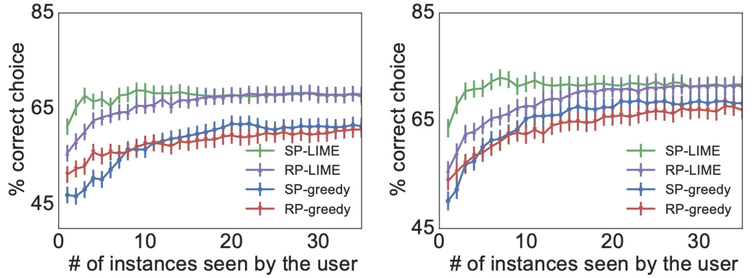
Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	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**

Can I trust this model?

- Evaluate if explanations can be used for model selection
- They add 10 artificially âĂIJnoisyâĂİ features s.t.
  - Each artificial feature appears in 10% of the examples in one class, and 20% of the other in the training/validation data.
  - While on the test instances, each artificial feature appears in 10% of the examples in each class.
- Results in models both using actual informative features, but also ones creating random correlations.
- Pairs of competing classifiers are computed by repeatedly training pairs of random forests with 30 trees until their validation accuracy is within 0.1% of each other, but their test accuracy differs by at least 5%.

Can I trust this model?



(a) Books dataset (b) DVDs dataset

Can humans pick the best classifier?

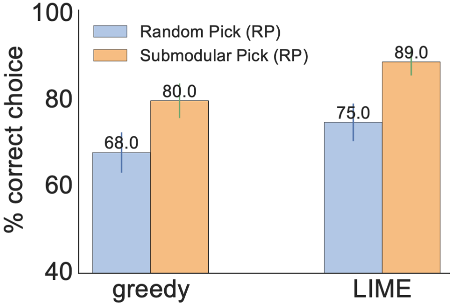
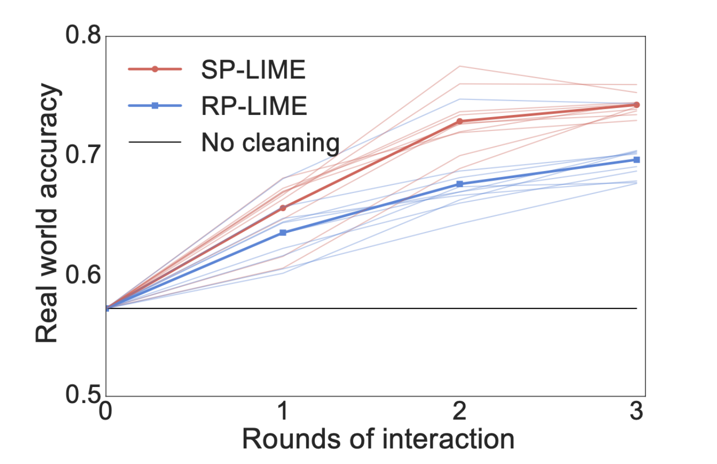
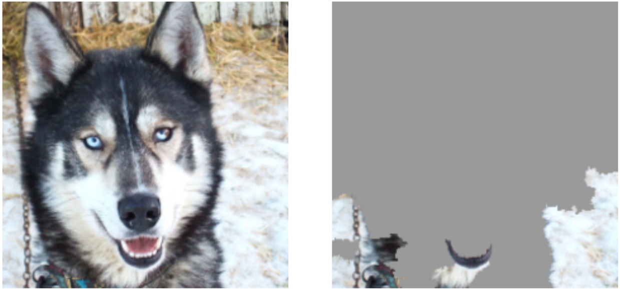


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

Can non-experts improve a classifier?



Can we learn something from the explanations?

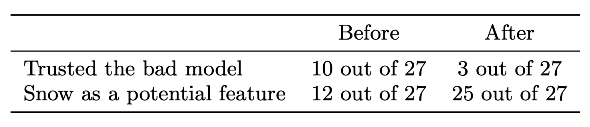


(a) Husky classified as wolf

(b) Explanation

Can we learn something from the explanations?

- Present 10 predictions without explanations
  - 2 are miss-predictions with a husky in snow and a wolf without snow, the rest are correct
- Ask three questions:
  1. Do you trust this algorithm to generalize?
  2. Why?
  3. How do you think the algorithm distinguishes?
- Results shown in table, before and after having seen the explanations.



	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Conclusion

Conclusion

- They argue that trust is crucial for effective human interaction with machine learning systems
- Explaining individual predictions is important in assessing trust
- They proposed LIME, a modular and extensible ap- proach to faithfully explain the predictions of any model in an interpretable manner
- They introduced SP-LIME, a method to select representative and non-redundant predictions, providing a global view of the model to users.
- Experiments demonstrated that explanations are useful for a variety of models in trust-related tasks in the text and image domains

Future work

- They use only sparse linear models as explanations, our framework supports the exploration of a variety of explanation families, such as DTs.
    - This estimate of faithfulness can also be used for selecting an appropriate family of explanations from a set of multiple interpretable model classes, thus adapting to the given dataset and the classifier.
- One issue that they do not mention in this work was how to perform the pick step for images.
- They would like to investigate potential uses in speech, video, and medical domains, as well as recommendation systems.
- They would like to explore theoretical properties (such as the appropriate number of samples) and computational optimizations (such as using parallelization and GPU processing)

Recap

Recap