Explaining Black-Box Machine Learning Predictions

Sameer Singh
University of California, Irvine

work with Marco T. Ribeiro and Carlos Guestrin
Machine Learning is Everywhere...
Classification: Wolf or a Husky?

Adopt or not?

Machine Learning Model

Wolf!
Classification: Wolf or a Husky?

Adopt or not?
Classification: Wolf or a Husky?

Only 1 mistake!
Visual Question Answering

Is there a moustache in the picture?

> Yes

What is the moustache made of?

> Banana
Essentially black-boxes!

- **Trust**: How can we trust the predictions are correct?
- **Predict**: How can we understand and predict the behavior?
- **Improve**: How do we improve it to prevent potential mistakes?
Classification: Wolf or a Husky?

Only 1 mistake!

We’ve built a snow detector...
The Man Who Accidentally Adopted a Wolf Pup

It did not go well.

By A.J. McCarthy
Visual Question Answering

What is the moustache made of?
> Banana

What are the eyes made of?
> Bananas

What?
> Banana

What is?
> Banana
From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion. If you’d like to know more, send me a note.
I would like to apply for a loan.

Here is my information.

Sorry, your request has been denied

Why? What were the reasons?

Currently

Cannot explain.. [0.25,-4.5,3.5,-10.4,...]
How did we get here?
What is Machine Learning?

Historical Input Data

Historical Output Data

Machine Learning

New Inputs

Classifier

Predict Output

Complex Math
Should I give out a loan?
Get Historical Data
You can interpret it...
- Both have a positive effect
- \( X1 > X2 \)

\[
\text{if: } 10X1 + X2 - 5 > 0
\]

\[
\text{otherwise}
\]

\[
10X1 + X2 - 5 > 0
\]

\[
\text{otherwise}
\]
Decision trees

You can interpret it...
- X2 is irrelevant if X1 < 0.5
- Otherwise X2 is enough
Looking at the structure

Trust
How can we trust the predictions are correct?
Test whether the structure agrees with our intuitions.

Predict
How can we understand and predict the behavior?
Structure tells us exactly what will happen on any data.

Improve
How do we improve it to prevent potential mistakes?
Structure tells you where the error is, thus how to fix it.
Arrival of Big Data
Big Data: Applications of ML
Big Data: More Complexity
Big Data: More Complexity
Big Data: More Complexity
The peaks, or more tightly arrayed lines, pinpoint areas of highest risk with more precision than the regression line can provide.
Big Data: More Dimensions

- Savings
- Income
- Loan Amount
- Profession
- Age
- Marital Status
- Credit scores
- Past defaults
- Recent defaults

This *easily* goes to hundreds
- Images: thousands
- Text: tens of thousands
- Video: millions
- ... and so on
Complex Surfaces + Lots of dimensions

Black-boxes!
Accuracy vs Interpretability

Accuracy

Real-world use case

Research on "interpretable models"

10X1 + X2 - 5 > 0

x1 > 0.5

x2 > 0.5

millions of weights, complex features

156x83 millions of weights, complex features

475x282 Real-world use case

650x263 Research on “interpretable models”

94x419 Accuracy vs Interpretability

240x318 Accuracy

681x78 10X1 + X2

757x78 - 5 > 0

723x170 X1 > 0.5

739x136 X2 > 0.5
Deep Learning

Focus on accuracy!

Human-level

Real-world use case

Research on “interpretable models”

Accuracy

Interpretability
Looking at the structure

- **Trust**: How can we trust the predictions are correct?
  - Test whether the structure agrees with our intuitions.

- **Predict**: How can we understand and predict the behavior?
  - Structure tells us exactly what will happen on any data.

- **Improve**: How do we improve it to prevent potential mistakes?
  - Structure tells you where the error is, thus how to fix it.
Explaining Predictions

The LIME Algorithm
Applying for a Loan

I would like to apply for a loan.

Here is my information.

Sorry, your request has been denied

Why? What were the reasons?

My Research

35% of my decision is based on your race, 20% on your income, and 15% on your savings.
Being Model-Agnostic...

No assumptions about the internal structure...

Data \rightarrow f(x) \rightarrow Decision

Explain any existing, or future, model
LIME: Explain Any Classifier!

- Interpretability
- Accuracy

Real-world use case

Make everything interpretable!
From: Keith Richards
Subject: Christianity is the answer
NTTP Posting-Host: x.x.com

I think Christianity is the one true religion. If you’d like to know more, send me a note.
Being Model-Agnostic...

“Global” explanation is too complicated
Being Model-Agnostic...

“Global” explanation is too complicated
Being Model-Agnostic...

“Global” explanation is too complicated

Explanation is an interpretable model, that is locally accurate
Example – Image Classification

Original Image \( P(\text{labrador}) = 0.21 \)

<table>
<thead>
<tr>
<th>Perturbed Instances</th>
<th>( P(\text{Labrador}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.34</td>
</tr>
</tbody>
</table>

Locally weighted regression

Explanation
Google’s Object Detector

\[ P(\text{dog}) = 0.21 \]

\[ P(\text{guitar}) = 0.24 \]

\[ P(\text{guitar}) = 0.32 \]
Classification: Wolf or a Husky?

Only 1 mistake!
Neural Network Explanations

We’ve built a great snow detector...
Visual QA

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the mustache made of?</td>
<td>banana</td>
</tr>
<tr>
<td>How many bananas are in the picture?</td>
<td>2</td>
</tr>
<tr>
<td>English</td>
<td>Portuguese</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>This is the question we must address</td>
<td>Esta é a questão que temos que enfrentar.</td>
</tr>
</tbody>
</table>
Neural Machine Translation

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is the question we must address</td>
<td>Esta é a questão que temos que enfrentar.</td>
</tr>
<tr>
<td>This is the problem we must address</td>
<td>Este é o problema que temos que enfrentar.</td>
</tr>
</tbody>
</table>
### Neural Machine Translation

<table>
<thead>
<tr>
<th>English</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is the <strong>question</strong> we must address</td>
<td>Esta é a questão que temos que enfrentar.</td>
</tr>
<tr>
<td>This is the <strong>problem</strong> we must address</td>
<td>Este é o problema que temos que enfrentar.</td>
</tr>
<tr>
<td>This is what we must address</td>
<td>É <em>isso</em> que temos de enfrentar.</td>
</tr>
</tbody>
</table>
Salary Prediction

### Features and Values

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>$37 &lt; Age ≤ 48</td>
</tr>
<tr>
<td>Workclass</td>
<td>Private</td>
</tr>
<tr>
<td>Education</td>
<td>≤ High School</td>
</tr>
<tr>
<td>Marital Status</td>
<td>Married</td>
</tr>
<tr>
<td>Occupation</td>
<td>Craft-repair</td>
</tr>
<tr>
<td>Relationship</td>
<td>Husband</td>
</tr>
<tr>
<td>Race</td>
<td>Black</td>
</tr>
<tr>
<td>Sex</td>
<td>Male</td>
</tr>
<tr>
<td>Capital Gain</td>
<td>0</td>
</tr>
<tr>
<td>Capital Loss</td>
<td>0</td>
</tr>
<tr>
<td>Hours per week</td>
<td>≤ 40</td>
</tr>
<tr>
<td>Country</td>
<td>United States</td>
</tr>
</tbody>
</table>

### Salary Prediction

- **Salary**
  - **>$50K**: 29%
  - **≤$50K**: 71%

**Decision Rule**

IF Education ≤ High School  
Then Predict Salary ≤ $50K
Salary Prediction

28 < Age ≤ 37
Workclass = Private
Education = Doctorate
Marital Status = Married
Occupation = Professional
Relationship = Husband
Race = White
Sex = Male
Capital Gain = None
Capital Loss = None
Hours per week > 45.00
Country = United-States

IF Married and Education = Doctorate
Then Predict Salary > 50K

Data from the US Census
“Global” Behavior

What about explaining the rest of the model?
Explaining Global behavior

LIME explains a single prediction
local behavior for a single instance

Can’t examine all explanations
Instead pick $k$ explanations to show to the user

Representative
Should summarize the model’s global behavior

Diverse
Should not be redundant in their descriptions

Single explanation
Are they useful?
Quantitative Evaluation

- **Understand** what ML is doing
- **Compare** different ML algorithms
- **Improve** the existing model
- **Predict** how ML will behave
Quantitative Evaluation

Understand what ML is doing

Compare different ML algorithms

Improve the existing model

Predict how ML will behave
Understanding Behavior

We’ve built a great snow detector...
Understanding Behavior

Question 1
Would you trust this model?

Question 2
What is the classifier is doing?
Did they notice it?

<table>
<thead>
<tr>
<th></th>
<th>Before explanations</th>
<th>After explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Didn't trust the model</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td>&quot;Snow insight&quot;</td>
<td>50%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Quantitative Evaluation

- **Understand** what ML is doing
- **Compare** different ML algorithms
- **Improve** the existing model
- **Predict** how ML will behave
Comparing Classifiers

Classifier 1
Change the model
Different data
Different parameters
Different "features"
...

Classifier 2

Accuracy?
Look at Examples?
Deploy and Check?
“I have a gut feeling..”
Explanations?
Comparing Classifiers

Original Image  “Bad” Classifier  “Good” Classifier
I think Christianity is the one true religion. If you'd like to know more, send me a note.

After looking at the explanation, we shouldn't trust the model!
It seems to be picking up on more reasonable things.. good!
UI for Comparing Classifiers
Comparing Models

89% of users identify the more trustworthy model

If we picked based on accuracy, we would get it wrong.

% picked better model

Guessing | LIME-random | LIME-global

40 | 70 | 100
Quantitative Evaluation

Understand what ML is doing

Compare different ML algorithms

Improve the existing model

Predict how ML will behave
Improving Classifiers

- Classifier
  - Generate Explanations
    - Compute Accuracy
  - Suggest Changes
    - Lay people
  - Show Explanations
    - Explanations
UI for fixing bad classifiers

Example #5 of 10

True Class: Atheism

Words that the algorithm considers important:

<table>
<thead>
<tr>
<th>Word</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host</td>
<td></td>
</tr>
<tr>
<td>Posting</td>
<td></td>
</tr>
<tr>
<td>NNTP</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td></td>
</tr>
<tr>
<td>New</td>
<td></td>
</tr>
<tr>
<td>Thanks</td>
<td></td>
</tr>
<tr>
<td>anyone</td>
<td></td>
</tr>
<tr>
<td>email</td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
</tr>
<tr>
<td>has</td>
<td></td>
</tr>
</tbody>
</table>

Bar length indicates importance, and color indicates to which topic: Christianity (green) or Atheism (pink).

Please click on the words (right next to the bars) that you think the algorithm is using incorrectly, because they are not important to distinguish between Atheism and Christianity. They should be red and crossed off after you click them.

Document

From: johnchad@triton.unm.edu (jchadvie)
Subject: Another request for Darwin fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish. This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Thanks,

john chadwick
johnchad@triton.unm.edu
or
Fixing bad classifiers

![Graph showing the accuracy on hidden set with different training methods.
- Train by lay people.
- Train by us (without explanations).
- Train on original data.

Accuracy on hidden set is plotted against training rounds (w/o feedback, 1 round, 2 rounds, 3 rounds).

Accuracy values:
- 0.5
- 0.575
- 0.65
- 0.725
- 0.8]
Quantitative Evaluation

- **Understand** what ML is doing
- **Compare** different ML algorithms
- **Improve** the existing model
- **Predict** how ML will behave
Predicting Behavior

Data

Classifier

Predictions & Explanations

New Data

Predictions

Show to user

User guesses what the classifier would do on new data

Compare Accuracy
User Studies: Precision

How good are users guesses on unseen instances?

<table>
<thead>
<tr>
<th></th>
<th>VQA 1</th>
<th>VQA 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Explanation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>With Explanations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
User Studies: Time

How long do users take to make their guesses?

- No Explanation
  - VQA 1
  - VQA 2

- With Explanations
  - VQA 1
  - VQA 2
Explanations are important!

- **Trust**: How can we trust the predictions are correct?
- **Predict**: How can we understand and predict the behavior?
- **Improve**: How do we improve it to prevent potential mistakes?

Model-Agnostic Explanations
Model Agnostic Explanations

“Why should I trust you?”
Explaining the predictions of any classifier
Ribeiro, Singh, Guestrin, KDD 2016
github.com/marcotcr/lime

Thanks!

sameer@uci.edu
sameersingh.org