Explaining Explainable AI

Marco Tulio Ribeiro (Microsoft Research)
ML according to ML101

What’s missing?
ML in real life: humans matter

Which model / features should I use?
Is my model any good?

Should I trust / deploy this model?
Is it racist?

Why did it say that?
Can I trust this?

Data → statistics →
ML Person → Business / Product person → End-user
GDPR!!
**Explain**

[ikˈsplān] 🔊

VFRB

*make (an idea, situation, or problem) clear to someone by describing it in more detail or revealing relevant facts or ideas.*
# Cheat sheet

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Explanation Type</th>
<th>Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are explanations for?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Purpose of explanation

1. Augment humans

How do I diagnose disease X?

Should I buy this book?

Should I grant bail to John?

Why is Mary suddenly bleeding?

What is wrong with Mary?
Purpose of explanation

General

Specific
Purpose of explanation

2. Help humans evaluate the AI

Can I trust this AI?

Is the AI racist?

GDPR compliance! Explain every prediction!

General

Specific
Purpose of explanation
Purpose of explanation

3. Help humans improve the AI

What kind of data should I get?

Do I need a new architecture?

How do I fix this?
Cheat sheet

1. Purpose
- What are explanations for?
  - Augment humans
  - Evaluate the AI
  - Improve the AI
- What questions do you have?

2. Explanation Type
- general
- specific

3. Technique
Example: predicting income

Census Data

<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>

\[ f(x) \]

AI

Salary

Less than 50K: 71%
More than 50K: 29%
Explanation(s)

- Feature Importance
- Sliced Statistics
- Whole model
- Partial Dependency Plot
- Anchor
- Linear model
- Counterfactual

General

Specific
Explanation(s)

Which one is the best?

Feature importance

Sliced Statistics

General

Influence

Whole model

Partial Dependency Plot

Anchor

Linear model

Counterfactual

What are important predictors of income?

Bill G

What do I make more money?

How do I make my model fair?
Cheat sheet

① Purpose

- What are explanations for?
  - Augment humans
  - Evaluate the AI
  - Improve the AI

- What questions do you have?
  - general
  - specific

② Explanation Type

- Linear model
- Partial Dependency Plot
- Counterfactual
- Sliced Statistics
- Anchor

③ Technique

Whole model
Technique: how to get explanations

Can I please have Interpretable ML?

I'm comfortable with

Interpretable model

Sure, let's train a linear model!

Black box explanations

Sure, bring your model and we'll explain it

Changed my mind, I want

Interpretable model

OK, let's train a different model

Interpretable model

More accurate

Fast

Changed my mind, I want

Interpretable model

More accurate

Fast
Technique: how to get explanations

Can I trust the AI to diagnose patients alone?

Interpretable model

Sure! Give me a few more years :)

- Exact
- Fast

Black box explanations

Let me give you Rich’s contact info...

- More accurate
- Flexible
Accuracy vs Interpretability

- Accuracy: millions of weights, complex features
- Interpretability

Real-world use case:
- $X_1 > 5$
- $X_2 > 5$
- $10X_1 + X_2 - 5 > 0$

Research on interpretable models
Deep Learning / Ensembles

Accuracy

Interpretability

Focus on accuracy!

Human-level

Real-world use case
Being Model-Agnostic...

- Ignore the internal structure
- Explain any existing, *or future* model
- Compare any 2 models to each other
- Adapt explanation to target user

\[ X \rightarrow f(x) \rightarrow + \]
Model-Agnostic: Explain Any Classifier!

Make everything interpretable!
Interpretable model

Accuracy vs. Interpretability

Avoid shortcuts, ‘real’ accuracy
Cheat sheet

1 Purpose

- What are explanations for?
  - Augment humans
  - Evaluate the AI
  - Improve the AI

- What questions do you have?
  - General
  - Specific

2 Explanation Type

- Linear model
- Partial Dependency Plot
- Anchor
- Counterfactual
- Sliced Statistics
- Whole model

3 Technique

- Interpretable models
- Black box
LIME

"Why Should I Trust You?": Explaining the Predictions of Any Classifier
Ribeiro et al, KDD 2016
Explaining individual predictions

“Global” explanation is too complicated
Being Local...

“Global” explanation is too complicated
Being Local…

Explaination is locally faithful

“Global” explanation is too complicated
What an explanation looks like

Model seems good, let's look at some explanations.

Atheism Christianity

RBF SVM

Test set Accuracy

94%

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.

Prediction Probabilities

<table>
<thead>
<tr>
<th>Category</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>atheism</td>
<td>82%</td>
</tr>
<tr>
<td>christianity</td>
<td>18%</td>
</tr>
</tbody>
</table>

11% of training, always in atheism.
Local Interpretable Model-Agnostic Explanations

1. Sample points around $x$
2. Get predictions from complex model
3. Weight samples according to distance to $x$
4. Learn simple model on weighted samples
5. Use simple model to explain $x$
Interpretable representations

\[ x \text{ (embeddings)} \]

\[ 3.1 \ 1.7 \ 8.4 \ 0.1 \ \ldots \]

\[ x' \text{ (words)} \]

This is a horrible movie.

We use \( x' \) to perturb and explain
Interpretable representation: images

$x$ (3 color channels / pixel)

$x'$ (contiguous superpixels)
Sampling example - images

Original Image
P(labrador) = 0.21

Perturbed Instances  P(Labrador)

- 0.92
- 0.001
- 0.34

Explanation
Locally weighted regression
Explaining Global behavior

LIME explains a single prediction

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

Chosen set must be representative...

...and diverse

How: Submodular optimization
LIME

1. Purpose
   • What are explanations for?
     - Augment humans
     - Evaluate the AI
     - Improve the AI
   • What questions do you have?
     - general
     - specific

2. Explanation Type
   - Linear model
   - Partial Dependency Plot
   - Anchor
   - Counterfactual
   - Sliced Statistics
   - Whole model

3. Technique
   - Interpretable models
   - Black box
     - LIME (Ribeiro et al. 2016)
Experiments

Purpose

- [ ] Augment humans
- [X] Evaluate the AI
- [ ] Improve the AI
Experiment: Wolf or a Husky?

Only 1 mistake!
Neural Network Explanations

We built a deep snow detector
Do ML people get this insight?

1. Would you trust this model?
2. What is the model doing?
Explanations help experts get insights, avoid mistakes

| % of subjects (out of 27) | Didn't trust the model | "Snow insight"
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Before explanations</td>
<td>62,3</td>
<td>44,4</td>
</tr>
<tr>
<td>After explanations</td>
<td>88,9</td>
<td>92,6</td>
</tr>
</tbody>
</table>
Experiments

Purpose

- Augment humans
- Evaluate the AI
- Improve the AI
Experiment: Model selection

Turkers asked to pick model that generalizes better
Choosing between competing models

Turkers can do model selection

<table>
<thead>
<tr>
<th>Model Type</th>
<th>% Picked Better Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guessing</td>
<td>50</td>
</tr>
<tr>
<td>LIME - Random</td>
<td>75</td>
</tr>
<tr>
<td>LIME - Submodular</td>
<td>89</td>
</tr>
</tbody>
</table>
Fixing bad classifiers

Atheism Christianity

Repeat

New Dataset

Turkers don’t know about this dataset

Atheism Christianity

Posting

Host

Keith

Atheism Christianity

Posting

Host

Keith
Fixing bad classifiers

No user input
1 round
2 rounds
3 rounds

Accuracy on hidden set

Train on 20 newsgroups
Train on hand-cleaned 20 newsgroups
Train on turkers clean data

Turkers can do feature engineering
Experiments

Purpose

- Augment humans
- Evaluate the AI
- Improve the AI
Cheat sheet

1. Purpose
   - What are explanations for?
     - Augment humans
     - Evaluate the AI
     - Improve the AI
   - What questions do you have?
     - general
     - specific

2. Explanation Type
   - Black box
   - Interpretable models
     - Linear model
     - Counterfactual
     - Partial Dependency Plot
     - Sliced Statistics
     - Anchor
     - Whole model

3. Technique
   - Interpretable models
     - LIME [Ribeiro et al 2016]
     - SHAP [Lundberg et al 2017]
LIME vs SHAP

What’s the difference?
Different meanings for weights

LIME: weight is local approximation
If you increase age, \( f(x) \) goes down, if you decrease it, \( f(x) \) goes up.
Ergo, Age has negative weight

SHAP: weight is contribution w.r.t baseline
John’s age contributes positively towards \( f(x) \) w.r.t. the average.
Ergo, Age has positive weight

Which one is right?
Another example

LIME: Age doesn’t matter

SHAP: Age has positive weight

$\phi(x)$

Age

John
SHAP experiments

Purpose

- Augment humans
- Evaluate the AI
- Improve the AI

Lundberg et al., Explainable machine-learning predictions for the prevention of hypoxemia during surgery, Nature Biomedical Engineering 2018 (cover article)
Cheat sheet

① Purpose
- What are explanations for?
  - Augment humans
  - Evaluate the AI
  - Improve the AI

- What questions do you have?
  - general
  - specific

② Explanation Type
- Linear model
- Partial Dependency Plot
- Anchor
- Whole model
- Counterfactual
- Sliced Statistics

③ Technique
- Interpretable models
- Black box
  - LIME [Ribeiro et al 2016]
  - SHAP [Lundberg et al 2017]
Explaining Explainable AI

Marco Tulio Ribeiro (Microsoft Research)