DATA MINING 2

Explainability

Riccardo Guidotti

a.a. 2022/2023
Definitions

• To *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts.

• In AI, and in data mining and machine learning, interpretability is the *ability to explain* or to provide the meaning *in understandable terms to a human*.

- [https://www.merriam-webster.com/](https://www.merriam-webster.com/)
A black box is a model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.

Example:
- DNN
- SVM
- Ensemble

Interpretable Models

Decision Tree

if condition$_1$ $\land$ condition$_2$ $\land$ condition$_3$ then outcome

Linear Model

PREDICTION: $p(\text{survived} = \text{yes} | X) = 0.671$

OUTCOME: YES

<table>
<thead>
<tr>
<th>Feature</th>
<th>contribution</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClass</td>
<td>-0.344</td>
<td>3rd</td>
</tr>
<tr>
<td>Age</td>
<td>-0.034</td>
<td>52</td>
</tr>
<tr>
<td>Sex</td>
<td>1.194</td>
<td>female</td>
</tr>
</tbody>
</table>
Motivations For Explanation Methods
COMPAS Recidivism

DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3

BERNARD PARKER

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.
The Wolf and the Husky

(a) Husky classified as wolf   (b) Explanation
Since 25 May 2018, GDPR establishes a right for all individuals to obtain “meaningful explanations of the logic involved” when “automated (algorithmic) individual decision-making”, including profiling, takes place.
Explanation in different AI fields

• Machine Learning

Feature Importance, Partial Dependence Plot, Individual Conditional Expectation

Auto-encoder


Surrogate Model

Explanation in different AI fields

- Machine Learning
- Computer Vision

**Uncertainty Map**

Alex Kendall, Yarin Gal: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NIPS 2017: 5580-5590

**Saliency Map**

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning

Abduction Reasoning (in Bayesian Network)


Diagnosis Inference

Explanation in different AI fields

• Machine Learning
• Computer Vision
• Knowledge Representation and Reasoning
• Multi-agent Systems
Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling

Explanation in different AI fields

- Machine Learning
- Computer Vision
- Knowledge Representation and Reasoning
- Multi-agent Systems
- NLP
- Planning and Scheduling
- Robotics

Robots: I have decided to turn left.
Human: Why did you do that?
Robots: I believe that the correct action is to turn left
BECAUSE:
I’m being asked to go forward
AND This area in front of me was 20 cm higher than me
*highlights area*
AND the area to the left has maximum protrusions of less
than 5 cm *highlights area*
AND I’m tilted to the right by more than 5 degrees.
Here is a display of the path through the tree that lead to
this decision. *displays tree*
Human: How confident are you in this decision?
Robots: The distribution of actions that reached this leaf
node is shown in this histogram. *displays histogram*
This action is predicted to be correct 67% of the time.
Human: Where did the threshold for the area in front come
from?
Robots: Here is the histogram of all training examples that
reached this leaf. 80% of examples where this area was
above 20 cm predicted the appropriate action to be “drive
forward”.

From Decision Tree to human-friendly information
Raymond Ka-Man Sheh: "Why Did You Do That?" Explainable Intelligent
Robots. AAAI Workshops 2017
Explanation as *Machine-Human Conversation*

- Humans may have follow-up questions
- Explanations cannot answer all users’ concerns

[Weld and Bansal 2018]
Explanation as Machine-Human Conversation

- Humans may have follow-up questions
- Explanations cannot answer all users’ concerns
Explanation as *Machine-Human Conversation*

- Humans may have follow-up questions
- Explanations cannot answer all users’ concerns
Explanation as Machine-Human Conversation

- Humans may have follow-up questions
- Explanations cannot answer all users’ concerns

[Weld and Bansal 2018]
Role-based Interpretability

“Is the explanation interpretable?” \(\rightarrow\) “To whom is the explanation interpretable?”

No Universally Interpretable Explanations!

- **End users** “Am I being treated fairly?”
  “Can I contest the decision?”
  “What could I do differently to get a positive outcome?”

- **Engineers, data scientists**: “Is my system working as designed?”

- **Regulators** “Is it compliant?”

An ideal explainer should model the *user background*.

XAI is Interdisciplinary

• For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure

• [Tim Miller 2018]
How to Open the Black Box
XAI Taxonomy of Explanation Methods
XAI Taxonomy of Explanation Methods

- Explainable By Design Methods (Intrinsic Explainability)
- Black Box Explanation Methods (Post-hoc Explainability)
XAI Taxonomy of Explanation Methods

Explanation Methods

Explainable By Design Methods (Intrinsic Explainability)

Black Box Explanation Methods (Post-hoc Explainability)

Input Data

Interpretability

Transparent System

Black-box System

\( \hat{y} \)

\( y \)

\( \hat{y} \)

Explanation Sub-system
XAI Taxonomy of Explanation Methods

Explanation Methods

- Explainable By Design Methods (Intrinsic Explainability)
- Black Box Explanation Methods (Post-hoc Explainability)

are

Global and Model Specific
Explainable by Design Method

Dataset $X$ → Interpretable Model $c$ → Prediction and Explanation $y, e = c(X)$ → User

Outlook

- Sunny
- Rain

Humidity
- High
- Normal

Wind
- Strong
- Weak

- No
- Yes

Overcast

- Yes
XAI Taxonomy of Explanation Methods

Explanation Methods

- Explainable By Design Methods (Intrinsic Explainability)
  - Global and Model Specific
    - Global
    - Local

- Black Box Explanation Methods (Post-hoc Explainability)

Black Box Explanations: Global vs Local

- **Global Explanation**
  - Dataset $X$
  - Black Box Model $b$
  - Prediction $y = b(X)$
  - Explanation Method $f$
  - Explanation $e = f(b, x)$
  - User

**Example:**
- If Outlook = Sunny and Humidity = Normal
  - then Play Tennis = Yes

- **Local Explanations**
  - Outlook: 0.7
  - Humidity: -0.4
  - Wind: 0.0

**Global Explanation Diagram**

- Outlook
  - Sunny
  - Rain

- Humidity
  - High
  - Normal
  - Strong
  - Weak
  - Yes
  - No

- Wind
  - Yes
  - No

- Overcast
  - Yes

**Decision Tree**

- If Outlook = Sunny and Humidity = Normal
  - then Play Tennis = Yes
Black Box Explanations: Specific vs Agnostic

Model Specific

- Dataset $X$
- Black Box Model $b$
- Explanation Method $f$
- Prediction $y = b(X)$
- Explanation $e = f(b, x)$
- User

Model Agnostic

- Dataset $X$
- Black Box Model $b$
- Explanation Method $f$
- Prediction $y = b(X)$
- Explanation $e = f(b, x)$
- User

Note: $f$ is black box dependent in model specific and black box independent in model agnostic.
Types of Data

- **Text (TXT)**
- **Tabular (TAB)**
- **Images (IMG)**

Table of baby-name data (baby-2010.csv)

<table>
<thead>
<tr>
<th>name</th>
<th>rank</th>
<th>gender</th>
<th>year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jacob</td>
<td>1</td>
<td>boy</td>
<td>2010</td>
</tr>
<tr>
<td>Isabella</td>
<td>1</td>
<td>girl</td>
<td>2010</td>
</tr>
<tr>
<td>Ethan</td>
<td>2</td>
<td>boy</td>
<td>2010</td>
</tr>
<tr>
<td>Sophia</td>
<td>2</td>
<td>girl</td>
<td>2010</td>
</tr>
<tr>
<td>Michael</td>
<td>3</td>
<td>boy</td>
<td>2010</td>
</tr>
</tbody>
</table>

2000 rows all told
Types of Explanations

- Tabular Data
  - Rule-based
  - Decision Tree
  - Features Importance
  - Prototypes
  - Counter-exemplars

- Images
  - Saliency Maps
  - Concept Attributions
  - Prototypes
  - Counter-exemplars

- Text
  - Sentence Highlighting
  - Attention-based
  - Prototypes
  - Counter-exemplars

If Outlook = Sunny and Humidity = Normal then Play Tennis = Yes

- Outlook: 0.7
- Humidity: -0.4
- Wind: 0.0
Explanations and Explanation Methods
TREPAN
• Global explainer designed to explain NN but usable for any type of black box.

• It aims at approximating a NN with a DT classifier using best-m-of-n rules.

• At each node split the feature to split is selected on the original data extended with random samples respecting the current path.

• It learns to predict the label returned by the black box, not the original one.
Trepan

```
01 T = root_of_the_tree()
02 Q = <T, X, {}>
03 while Q not empty & size(T) < limit
04 N, X_N, C_N = pop(Q)
05 Z_N = random(X_N, C_N)
06 Y_Z = b(Z), y = b(X_N)
07 if same_class(y U Y_Z)
     continue
08 S = best_split(X_N U Z_N, y U Y_Z)
09 S' = best_m-of-n_split(S)
10 N = update_with_split(N, S')
11 for each condition c in S'
     C = new_child_of(N)
     C_C = C N U {c}
12     X_C = select_with_constraints(X_N, C_N)
13     put(Q, <C, X_C, C_C>)
```
LIME
Local Explanation

• The overall decision boundary is complex
• In the neighborhood of a single decision, the boundary is simple
• A single decision can be explained by auditing the black box around the given instance and learning a *local* decision.
Local Interpretable Model-agnostic Explanations

- Local model-agnostic explainer that reveals the black box decisions through features importance/saliency maps.
- It locally approximates the behavior of a black box with a local surrogate expressed as a logistic regressor (with Lasso or Ridge penalization).
- Synthetic neighbors are weighted w.r.t. the distance with the instance to explain.
<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>b(setosa)</th>
<th>b(versic)</th>
<th>b(virgi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Sepal length</td>
<td>Sepal width</td>
<td>Petal length</td>
<td>Petal width</td>
<td>b(setosa)</td>
<td>b(versic)</td>
<td>b(virgi)</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>Sepal length</td>
<td>Sepal width</td>
<td>Petal length</td>
<td>Petal width</td>
<td>b(setosa)</td>
<td>b(versic)</td>
<td>b(virgi)</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Sepal length</td>
<td>Sepal width</td>
<td>Petal length</td>
<td>Petal width</td>
<td>b(setosa)</td>
<td>b(vers)</td>
<td>b(virgi)</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>0.0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>
## LIME

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>b(setosa)</th>
<th>b(versic)</th>
<th>b(virgi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>0.0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Train a Linear Regressor

Returns the coefficients as Explanation
## LIME

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>b(setosa)</th>
<th>b(versic)</th>
<th>b(virgi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>0.0</td>
<td>0.3</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>0.0</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Train a Linear Regressor
LIME

01 Z = {}
02 x instance to explain
03 x' = real2interpretable(x)
04 for i in {1, 2, ..., N}
05      z_i = sample_around(x')
06      z = interpretable2real(z')
07      Z = Z ∪ {<z_i, b(z_i), d(x, z)>}
08 w = solve_Lasso(Z, k)
09 return w

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. KDD.
LIME

- LIME **turns** an image $x$ to a vector $x'$ of interpretable superpixels expressing presence/absence.
- It **generates** a synthetic neighborhood $Z$ by randomly perturbing $x'$ and labels them with the black box.
- It **trains** a linear regression model (interpretable and locally faithful) and assigns a weight to each superpixel.
LORE
LOcal Rule-based Explainer

• LORE extends LIME adopting as local surrogate a decision tree classifier and by generating synthetic instances through a genetic procedure that accounts for both instances with the same labels and different ones.

• It can be generalized to work on images and text using the same data representation adopted by LIME.
01 x instance to explain
02 \( Z_\ = \ \text{geneticNeighborhood}(x, \ \text{fitness}_\, =, \ N/2) \)
03 \( Z_\# = \ \text{geneticNeighborhood}(x, \ \text{fitness}_\, \neq, \ N/2) \)
04 \( Z = Z_\ = \cup Z_\# \)
05 \( c = \text{buildTree}(Z, b(Z)) \)
06 \( r = (p \to y) = \text{extractRule}(c, x) \)
07 \( \phi = \text{extractCounterfactual}(c, r, x) \)
08 \text{return } e = <r, \phi>

\[
\begin{array}{|c|c|c|c|}
\hline
\text{parent} & \text{25} & \text{clerk} & \text{10k} & \text{yes} \\
\hline
\text{children 1} & \text{25} & \text{other} & \text{5k} & \text{no} \\
\hline
\text{children 2} & \text{30} & \text{clerk} & \text{10k} & \text{yes} \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|c|}
\hline
\text{parent} & \text{25} & \text{clerk} & \text{10k} & \text{yes} \\
\hline
\text{children 1} & \text{25} & \text{other} & \text{5k} & \text{no} \\
\hline
\text{children 2} & \text{27} & \text{clerk} & \text{7k} & \text{yes} \\
\hline
\end{array}
\]

r = \{\text{age} \leq 25, \text{job} = \text{clerk}, \text{income} \leq 900\} \to \text{deny}

\( \Phi = \{\{\text{income} > 900\} \to \text{grant}\}, \\
\{17 \leq \text{age} < 25, \text{job} = \text{other}\} \to \text{grant}\} \)
\( x_1 = \{ \text{Education} = \text{Bachelors}, \text{Occupation} = \text{Prof-specialty}, \text{Sex} = \text{Male}, \text{NativeCountry} = \text{Vietnam}, \text{Age} = 35, \text{Workclass} = 3, \text{HoursWeek} = 40, \text{Race} = \text{Asian-Pac-Islander}, \text{MaritalStatus} = \text{Married-civ}, \text{Relationship} = \text{Husband}, \text{CapitalGain} = 0, \text{CapitalLoss} = 0 \}, \geq 50k \)

\( x_2 = \{ \text{Education} = \text{College}, \text{Occupation} = \text{Sales}, \text{Sex} = \text{Male}, \text{NativeCountry} = \text{US}, \text{Age} = 19, \text{Workclass} = 2, \text{HoursWeek} = 15, \text{Race} = \text{White}, \text{MaritalStatus} = \text{Married-civ}, \text{Relationship} = \text{Husband}, \text{CapitalGain} = 2880, \text{CapitalLoss} = 0 \}, \leq 50k \)

\( r_{\text{lore}} = \{ \text{Education} > 5-6\text{th}, \text{Race} > 0.86, \text{WorkClass} \leq 3.41, \text{CapitalGain} \leq 20000, \text{CapitalLoss} \leq 1306 \} \rightarrow \geq 50k \)

\( r_{\text{lore}} = \{ \text{Education} \leq \text{Masters}, \text{Occupation} > -0.34, \text{HoursWeek} \leq 40, \text{WorkClass} \leq 3.50, \text{CapitalGain} \leq 10000, \text{Age} \leq 34 \} \rightarrow \leq 50k \)

\( c_{\text{lore}} = \{ \text{CapitalLoss} \geq 436 \} \rightarrow \leq 50k \)

\( c_{\text{lore}} = \{ \text{Education} > \text{Masters} \} \rightarrow > 50k \)
\( \{ \text{CapitalGain} > 20000 \} \rightarrow > 50k \)
\( \{ \text{Occupation} \leq -0.34 \} \rightarrow > 50k \)
LORE on Medical Images

• The goal is to classify dermoscopic images among categories such as: Melanoma (MEL), Melanocytic Nevus (NV); Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), etc.

• The original is classified as AK

• The counterfactual as BCC.
SHAP
Shapely Values

• A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values -- a method from coalitional game theory -- tells us how to fairly distribute the "payout" among the features.

• Example: A black box predicts apartment prices. For a certain apartment it predicts €300,000 and you need to explain this prediction. The apartment has an area of 50 m², is located on the 2nd floor, has a park nearby and cats are banned.
Shapely Values and Game Theory

• The average prediction is €310,000. How much has each feature value contributed to the prediction compared to the average prediction?
• The "game" is the prediction task for a single instance of the dataset.
• The "gain" is the actual prediction for this instance minus the average prediction for all instances.
• The "players" are the feature values of the instance that collaborate to receive the gain (= predict a certain value).
• The explanation could be: The park-nearby contributed €30,000; area-50 contributed €10,000; floor-2nd contributed €0; cat-banned contributed -€50,000. The contributions add up to -€10,000, the final prediction minus the average predicted apartment price.
The Shapley value is the average marginal contribution of a feature value across all possible coalitions (combination of fixed feature values).

We evaluate the contribution of cat-banned when it is added to a coalition of park-nearby and area-50.

We simulate that only park-nearby, cat-banned and area-50 are in a coalition by randomly drawing another apartment from the data and using its value for the floor feature.

The floor-2nd is replaced by the randomly drawn floor-1st.

Then we predict the price of the apartment with this combination (€310,000).
Shapely Values Example

• In a second step, we remove cat-banned from the coalition by replacing it with a random value of the cat allowed/banned from the randomly drawn apartment. In the example it was cat-allowed, but it could have been cat-banned again.

• We predict the apartment price for the coalition of park-nearby and area-50 (€320,000).

• The contribution of cat-banned was €310,000 - €320,000 = -€10,000. This estimate depends on the values of the randomly drawn apartment that served as a "donor" for the cat and floor feature values.

• We get better estimates if we repeat this sampling step and average the contributions.
Shapely Values Example

• We repeat this computation for all possible coalitions.

• The Shapley value is the average of all the marginal contributions to all possible coalitions.

• The computation time increases exponentially with the number of features.

• For each of these coalitions we compute the predicted apartment price with and without the feature value cat-banned and take the difference to get the marginal contribution.

• We replace the feature values of features that are not in a coalition with random feature values from the apartment dataset to get a prediction from the black box.

• If we estimate the Shapley values for all feature values, we get the complete distribution of the prediction (minus the average) among the feature values.
SHAP

• SHAP (SHapley Additive exPlanations) assigns each feature an importance value for a particular prediction by means of an additive feature attribution method.

• It assigns an importance value to each feature that represents the effect on the model prediction of including that feature.

SHAP on Tabular Data

Coalitions \[ h_x(z^i) \] \rightarrow \text{Feature values}

Instance \( x \)

\[ x = \begin{array}{ccc} \text{Age} & \text{Weight} & \text{Color} \\ 1 & 1 & 1 \end{array} \]

Instance with "absent" features

\[ z = \begin{array}{ccc} \text{Age} & \text{Weight} & \text{Color} \\ 1 & 0 & 0 \end{array} \]

\[ x = \begin{array}{ccc} \text{Age} & \text{Weight} & \text{Color} \\ 0.5 & 20 & \text{Blue} \end{array} \]

\[ z = \begin{array}{ccc} \text{Age} & \text{Weight} & \text{Color} \\ 0.5 & 17 & \text{Pink} \end{array} \]
SHAP on Images

Coalitions of super pixels $h(x^i)$ Image

Instance $x$

Instance $x$ with absent features
Saliency Maps
Saliency Maps

A saliency map is an image in which a pixel's brightness represents how salient the pixel is. A positive value (red) means that the pixel has contributed positively to the classification, while a negative one (blue) means that has contributed negatively.

There are two methods for creating SMs.

1. Assign to every pixel a saliency value.
2. Segment the image into different pixel groups (superpixels or segments) and then assign a saliency value for each group.
## Saliency Maps

<table>
<thead>
<tr>
<th>Model Prediction</th>
<th>5</th>
<th>6</th>
<th>3</th>
<th>dog</th>
<th>deer</th>
<th>deer</th>
<th>puck</th>
<th>shower cap</th>
<th>seashore</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Original</strong></td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>LIME</strong></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>&amp;-LRP</strong></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
<td><img src="image22.png" alt="Image" /></td>
<td><img src="image23.png" alt="Image" /></td>
<td><img src="image24.png" alt="Image" /></td>
<td><img src="image25.png" alt="Image" /></td>
<td><img src="image26.png" alt="Image" /></td>
<td><img src="image27.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>IntGrad</strong></td>
<td><img src="image28.png" alt="Image" /></td>
<td><img src="image29.png" alt="Image" /></td>
<td><img src="image30.png" alt="Image" /></td>
<td><img src="image31.png" alt="Image" /></td>
<td><img src="image32.png" alt="Image" /></td>
<td><img src="image33.png" alt="Image" /></td>
<td><img src="image34.png" alt="Image" /></td>
<td><img src="image35.png" alt="Image" /></td>
<td><img src="image36.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>DeepLift</strong></td>
<td><img src="image37.png" alt="Image" /></td>
<td><img src="image38.png" alt="Image" /></td>
<td><img src="image39.png" alt="Image" /></td>
<td><img src="image40.png" alt="Image" /></td>
<td><img src="image41.png" alt="Image" /></td>
<td><img src="image42.png" alt="Image" /></td>
<td><img src="image43.png" alt="Image" /></td>
<td><img src="image44.png" alt="Image" /></td>
<td><img src="image45.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Saliency Maps

Model Prediction

| 5 | 6 | 3 | dog | deer | deer | puck | shower cap | seashore |

Original

SmoothGrad

XRAI

GradCam

GradCam++

RISE
Integrated Gradient

- INTGRAD can only be applied to differentiable models.
- INTGRAD constructs a path from the baseline image $x'$ to the input $x$ and computes the gradients of points along the path.
- The points are taken by overlapping $x$ with $x'$, and gradually modifying the opacity of $x$. Saliency maps are obtained by cumulating the gradients of these points.

01 x instance to explain
02 **varying** x into x’ maximizing b(x)~b(x’)
03 the variation runs replacing a region R of x with:
   - constant value, noise, blurred image
04 reformulation: find smallest R such that b(x_R) ≪ b(x)

**Sentence Highlighting**

**INTGRAD**

the movie is not that bad, ringo lam sucks, i hate when van dam ##me has love in his movies, van dam ##me is good only when he doesn't have love in his movies.

**LIME**

the movie is not that bad, ringo lam sucks, i hate when van dam ##me has love in his movies, van dam ##me is good only when he doesn't have love in his movies.

**DeepLift**

the movie is not that bad, ringo lam sucks, i hate when van dam ##me has love in his movies, van dam ##me is good only when he doesn't have love in his movies.

**Gradient x Input**

the movie is not that bad, ringo lam sucks, i hate when van dam ##me has love in his movies, van dam ##me is good only when he doesn't have love in his movies.
Example-based Explanations
Example-based Explanations

• Example-based explanation methods select particular instances of the dataset or generate synthetic instances to explain black box behaviors.

• Example-based explainers are mainly local explainers.

• Example-based explanations only make sense if we can represent an instance of the data in a humanly understandable way.

• This works well for:
  • images
  • tabular data with not many features
  • short texts
Example-based Explanations

• We mainly recognize the following example-based explanations:

  • **Prototypes**: a selection of representative instances having the same class of the instance under analysis. Among prototypes we also recognize:
    • **Criticisms**: instances that are not well represented by prototypes.
    • **Influential Instances**: training points that were the most influential for the training of the black-box or for the prediction itself.

  • **Counterfactuals**: a selection of representative instances having a different class w.r.t. the instance under analysis.
Counterfactual Explanations

• A counterfactual explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred".

• Thinking in counterfactual terms requires imagining a hypothetical reality that contradicts the observed facts.

• Even if the relationship between the inputs and the outcome to be predicted might not be causal, we can see the inputs of a model as the cause of the prediction.

• A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output.
Counterfactual Explanations

• Counterfactuals answer why a decision has been made by highlighting what changes in the input would lead to a different outcome.

• CF are not generalizations!!!
A simple and naive approach to generating counterfactual explanations is *searching by trial and error*: randomly changing feature values of the instance of interest and stopping when the desired output is predicted.

As an alternative we can define a *loss function* that consider the instance of interest, a counterfactual and the desired (counterfactual) outcome. Then, we can find the *counterfactual explanation that minimizes this loss using an optimization algorithm*.

Many methods proceed in this way but differ in their definition of the loss function and optimization method.
### Counterfactuals with a Brute Force Procedure

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1200$</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>500$</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1200$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>500$</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>10000$</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1200$</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>income</th>
<th>other debts</th>
<th>car owner</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>500$</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Counterfactuals by Optimization Problems

• Most of the counterfactual explainers return counterfactuals by solving an optimization problem.

• The problem is typically designed through the definition of a loss function aimed at guaranteeing a set of desired properties.

• The objective is to find a counterfactual instance that minimizes this loss using an optimization (OPT) algorithm.
Optimized CF Search

Wachter et al. suggest minimizing the following loss:

\[
L(x, x', y', \lambda) = \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')
\]

\[
d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}
\]

\[
MAD_j = \text{median}_{i \in \{1, \ldots, n\}}(|x_{i,j} - \text{median}_{l \in \{1, \ldots, n\}}(x_{l,j})|)
\]

1. Sample a random CF \( x' \)
2. Optimize the loss \( L \)
3. If not \( |\hat{f}(x') - y'| \leq \epsilon \)
4. Increase Lambda. Go to 2.
5. Return the CF \( x' \) that minimizes the loss.

---

Optimized CF Search

• The loss function minimized by Wachter et al. is

\[ \lambda (b(x') - y')^2 + d(x, x') \]

• where the first term is the quadratic distance between the desired outcome \( y' \) and the classifier prediction on \( x' \), and the second term is the distance between \( x \) and \( x' \).

• Lambda balances the contribution of the first term against the second term.

Wachter S, Mittelstadt BD, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. HarvJL & Tech 31:841
Distance Functions

- Manhattan distance weighed with the inverse median absolute deviation MAD (used by Wachter)

\[ d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \quad MAD_j = \text{median}_{i\in\{1,\ldots,n\}}(|x_{i,j} - \text{median}_{l\in\{1,\ldots,n\}}(x_{l,j})|) \]

- Mixed Distance (used by Mothilal)

\[ d(a, b) = \frac{m_{\text{con}}}{m} \sum_{i\in\text{con}} \frac{|a_i - b_i|}{MAD_i} + \frac{m_{\text{cat}}}{m} \sum_{i\in\text{cat}} 1_{a_i \neq b_i} \]
DICE - Diverse Counterfactual Explanations

• DICE solves an optimization problem with penalization terms to ensure plausibility by similarity and diversity.
• It returns a set of $k$ plausible and different counterfactuals for $x$.

$$C(x) = \arg\min_{c_1, \ldots, c_k} \frac{1}{k} \sum_{i=1}^{k} y \text{loss}(f(c_i), y) + \frac{\lambda_1}{k} \sum_{i=1}^{k} \text{dist}(c_i, x) - \lambda_2 \text{dpp\_diversity}(c_1, \ldots, c_k)$$

Counterfactuals through Heuristic Strategies

• Heuristic strategies are typically much more efficient than optimization algorithms.

• Efficiency is paid with solutions that are not necessarily optimal.

• The search strategy is typically designed such that at each iteration, $x'$ is updated with the objective of minimizing a cost function.

• The cost function is based on a local and heuristic choice aiming for a valid counterfactual similar to $x$. 
The search is guided by local improvements via best-first search with pruning.

- $b(\text{"the quick brown fox jumps over the lazy dog"}) = y(0.8)$
- $b(\text{"the quick brown fox jumps over the lazy dog"}) = y'(0.3)$
- $b(\text{"the quick brown fox jumps over the lazy dog"}) = y(0.7)$
- ...
- $b(\text{"the quick brown fox jumps over the lazy dog"}) = y'(0.4)$

GSG - Growing Spheres Generation

• GSG relies on a generative approach growing a sphere of synthetic instances around $x$ to find the closest counterfactual $x'$.  
• GSG ignores in which direction the closest classification boundary might be.

Counterfactuals with Instance-Based Strategies

• The very simple but effective idea of instance-based (or case-based) approaches for counterfactual explanation is to search into a reference population instances to be used as counterfactuals.
NNCE - Nearest-Neighbor Counterfactual Explainer

- NNCE is an endogenous counterfactual explainer inspired by kNN classifiers that select as counterfactual(s) the instances in $x' \in X$ most similar to $x$ and with a different label, i.e., $b(x') \neq b(x)$.
- Candidate counterfactuals are sorted with respect to the distance between $x$, and the $k$ most similar ones are selected.

CBCE - Case-Based Counterfactual Explainer

• CBCE refines NNCE.
• It adopts the notion of explanation case ($xc$).
• Given $X$, an $xc$ is a couple of instances $(x,x')$ such that $(x,x')$ are the two most similar instances in $X$ and $b(x') \neq b(x)$.

Open The Black Box!

• *To empower* individual against undesired effects of automated decision making
• *To reveal* and protect new vulnerabilities
• *To implement* the “right of explanation”
• *To improve* industrial standards for developing AI-powered products, increasing the trust of companies and consumers
• *To help* people make better decisions
• *To align* algorithms with human values
• *To preserve* (and expand) human autonomy
Open Research Questions

• There is no agreement on what an explanation is
• There is not a formalism for explanations
• How to evaluate the goodness of explanations?
• There is no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans
• What if there is a cost for querying a black box?


References


Explanation Toolboxes and Repositories

- https://github.com/jphall663/awesome-machine-learning-interpretablility
- https://github.com/pbiecek/xai_resources
- https://github.com/ModelOriented/DrWhy
- https://fat-forensics.org/
- https://github.com/Trusted-AI/AIX360
- https://captum.ai/
- https://github.com/interpretml/interpret
- https://github.com/SeldonIO/alibi
- https://github.com/pair-code/what-if-tool