Seminar Explainable AI
Module 4
Overview on Explainable AI Methods
Birds-Eye View
Global-Local-Ante-hoc-Post-hoc

Andreas Holzinger
Human-Centered AI Lab (Holzinger Group)
Institute for Medical Informatics/Statistics, Medical University Graz, Austria
and
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada
00 Reflection – follow-up from last lecture
01 Basics, Definitions, ...
02 Please note: xAI is not new!
03 Examples for Ante-hoc models (explainable models, interpretable machine learning)
04 Examples for Post-hoc models (making the “black-box” model interpretable)
05 Explanation Interfaces: Future human-AI interaction
06 A few words on metrics of xAI (measuring causability)
01 Basics, Definitions, …
Inconsistent Definitions: What is the difference between explainable, interpretable, verifiable, intelligible, transparent, understandable ... ?


The expectations of xAI are extremely high

- **Trust** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?

- **Causality** - inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge

- **Transferability** – humans have a much higher capacity to generalize, and can transfer learned skills to completely new situations; compare this with e.g. susceptibility of CNNs to adversarial data (please remember that we rarely have iid data in real world)

- **Informativeness** - for example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision

- **Fairness and Ethical decision making** – interpretations for the purpose of assessing whether decisions produced automatically by algorithms conform to ethical standards

Glossar (1/3)

- **Ante-hoc Explainability (AHE)** = such models are interpretable by design, e.g. glass-box approaches; typical examples include linear regression, decision trees/lists, random forests, Naive Bayes and fuzzy inference systems; or GAMs, Stochastic AOGs, and deep symbolic networks; they have a long tradition and can be designed from expert knowledge or from data and are useful as framework for the interaction between human knowledge and hidden knowledge in the data.

- **BETA** = Black Box Explanation through Transparent Approximation, developed by Lakkarju, Bach & Leskovec (2016) it learns two-level decision sets, where each rule explains the model behaviour; this is an increasing problem in daily use of AI/ML, see e.g. [http://news.mit.edu/2019/better-fact-checking-fake-news-1017](http://news.mit.edu/2019/better-fact-checking-fake-news-1017)

- **Bias** = inability for a ML method to represent the true relationship; High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting);

- **Causability** = is a property of a human (natural intelligence) and a measurement for the degree of human understanding; we have developed a causability measurement scale (SCS).

- **Decomposition** = process of resolving relationships into the constituent components (hopefully representing the relevant interest). Highly theoretical, because in real-world this is hard due to the complexity (e.g. noise) and untraceable imponderabilities on our observations.

- **Deduction** = deriving of a conclusion by reasoning

- **Explainability** = motivated by the opaqueness of so called “black-box” approaches it is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized). Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact. (German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz)
Explanation = set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questions (“Why is this a cat?”). (German: Erklärung, Begründung)

Explanatory power = is the ability of a set hypothesis to effectively explain the subject matter it pertains to (opposite: explanatory impotence).

Explicit Knowledge = you can easily explain it by articulating it via natural language etc. and share it with others.

European General Data Protection Regulation (EU GDPR) = Regulation EU 2016/679 – see the EUR-Lex 32016R0679, will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made (see explainable AI).

Gaussian Process (GP) = collection of stochastic variables indexed by time or space so that each of them constitute a multidimensional Gaussian distribution; provides a probabilistic approach to learning in kernel machines (See: Carl Edward Rasmussen & Christopher K.I. Williams 2006. Gaussian processes for machine learning, Cambridge (MA), MIT Press); this can be used for explanations. (see also: Visual Exploration Gaussian)

Gradient = a vector providing the direction of maximum rate of change.

Ground truth = generally information provided by direct observation (i.e. empirical evidence) instead of provided by inference. For us it is the gold standard, i.e. the ideal expected result (100 % true);

Inverse Probability = an older term for the probability distribution of an unobserved variable, and was described by De Morgan 1837, in reference to Laplace’s (1774) method of probability.

Implicit Knowledge = very hard to articulate, we do it but cannot explain it (also tacit knowledge).

Kernel = class of algorithms for pattern analysis e.g. support vector machine (SVM); very useful for explainable AI

Kernel trick = transforming data into another dimension that has a clear dividing margin between the classes


Post-hoc Explainability (PHE) = such models are designed for interpreting black-box models and provide local explanations for a specific decision and re-enact on request, typical examples include LIME, BETA, LRP, or Local Gradient Explanation Vectors, prediction decomposition or simply feature selection.


Saliency map = image showing in a different representation (usually easier for human perception) each pixel’s quality.

Tacit Knowledge = Knowledge gained from personal experience that is even more difficult to express than implicit knowledge.

Transfer Learning (TL) = The ability of an algorithm to recognize and apply knowledge and skills learned in previous tasks to novel tasks or new domains, which share some commonality. Central question: Given a target task, how do we identify the commonality between the task and previous tasks, and transfer the knowledge from the previous tasks to the target one? Pan, S. J. & Yang, Q. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/TKDE.2009.191.
Expected trends in medical AI

Interpretable Models vs. Interpreting Models

- **Interpretable Models**, the model itself is already interpretable, e.g.
  - Regression
  - Naïve Bayes
  - Random Forests
  - Decision Trees/Graphs
  - ...

- **Interpreting Black-Box Models** (the model is not interpretable and needs a post-hoc interpretability method, e.g.:
  - Decomposition
  - LIME/BETA
  - LRP
  - ...

Rule-Based Models:
- Easy to interpret, the rules provide clear explanations
- Can learn even from little data sets
- Problems with high-dimensional data, with noise, and with images (ambiguity)

Neuro-Symbolic Models:
- Not easy to interpret ("black box")
- Needs a lot of top-quality training data
- Can well generalize even from high-dimensional data, with noise and good for images
- Needs previous knowledge
This is far too naïve: Explainability (better: interpretability !) does not correlate with performance !!

DARPA Causal Model Induction (CRA)


(a) | $c^+$ | $c^-$ |
---|---|---|
$e^+$ | $N(e^+, c^+)$ | $N(e^+, c^-)$ |
$e^-$ | $N(e^-, c^+)$ | $N(e^-, c^-)$ |

(b) Graph 0: $B \rightarrow C$
Graph 1: $B \rightarrow C \rightarrow E$

02 Please note: xAI is not new!
Explainability was the most requested feature of early medical decision support systems!

THE QUESTION-ANSWERING MODULE

** Is either chloramphenicol or kanamycin okay for treatment of salmonella infections?
YES.

RULE083

IF:
THE IDENTITY OF THE ORGANISM IS SALMONELLA

THEN:
I RECOMMEND (IN ORDER OF PREFERENCE):
1-CHLORAMPHENICOL
2-AMPICILLIN
3-CEPHALOTHIN

** How do you know if an organism is a bacteroides?

RULE030

IF:
THE GRAMSTAIN OF THE ORGANISM IS GRAMNEG, AND THE
MORPHOLOGY OF THE ORGANISM IS ROD, AND THE
AEROBICITY OF THE ORGANISM IS ANAEROBIC

THEN:
CONCLUDE THAT THE IDENTITY OF THE ORGANISM IS
BACTEROIDES (MODIFIER: THE CERTAINTY TALLY FOR
THE PREmise TIMES .6)

INTERNIST was lacking to explain its “thinking”

The evaluation demonstrated that the present form of the program is not sufficiently reliable for clinical applications. Specific deficiencies that must be overcome include the program's inability to reason anatomically or temporally, its inability to construct differential diagnoses spanning multiple areas, its occasional attribution of findings to improper causes, and its inability to explain its "thinking".

The success of deep learning brought new problems!

- AI is actually the oldest field of computer science, aiming at solving task where humans are good (e.g. speech, vision, problem solving, ...)
- Note that while the first goal was to mimic human intelligence across tasks, the success today is only on very narrow AI e.g. solving one specific task, playing a game, driving a car, classifying objects, ... due to the advancements in “deep learning” this works well,
- But the best performing methods remain opaque, i.e. are considered as so-called “black-box” models
Success in deep learning *) resulted in “deep problems” (e.g. complex and exploding gradients)

*) Note: “DL” methods are representation learning methods with multiple layers of representations (see LeCun, Bengio & Hinton (2015), Nature 521, 7553)

Problem in our society: “Secret algorithms” make important decisions about individuals (discussion of “bias, fairness, see Module 09)

Black box Type 1 = too complicated for a human to understand

Black box Type 2 = proprietary = “secret algorithm”

A black box model could be either

- (1) a function that is too complicated for any human to comprehend or
- (2) a function that is proprietary

Post-Hoc (latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ... (see module 5)

Ante-hoc (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

Explainable AI trees: From local explanations to global understanding

Explainable AI trees: From local explanations to global understanding

03 Examples for Ante Hoc Models (interpretable Machine Learning)
Differences: Post-hoc versus Ante-hoc

- **Post-Hoc** (latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ... (see module 5)

- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems (see module 3); typical methods decision trees, linear regression, and Random Forests.

Example for an Action Influence Graph of a Starcraft agent

State variables:
W - Worker number
S - Supply depot number
B - barracks number
E - enemay location
An - Ally unit number
Ah - Ally unit health
Al - Ally unit location
Du - Destroyed units
Db - Destroyed buildings

Actions:
As - build supply depot
Ab - build barracks
Am - train offensive unit
Aa - attack

Algorithm 1 Task Prediction: Action Influence Model

Input: trained regression models $\mathcal{L}$, current state $S_t$
Output: predicted action $a$

1: $\vec{F}_p \leftarrow []$ ; vector of predicted difference
2: for every $\hat{L} \in \mathcal{L}$ do
3: $P_y \leftarrow \hat{L} \cdot \text{predict}(S_{x,t})$; predict variable $S_y$ at $S_{t+1}$
4: $\vec{F}_p \leftarrow |S_y - P_y|$; difference with actual $S_y$ value
5: end for
6: return $\max(\vec{F}_p) \cdot \text{getAction()}$

<table>
<thead>
<tr>
<th>Env - RL</th>
<th>Size</th>
<th>Accuracy (%)</th>
<th>Performance (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>DT</td>
</tr>
<tr>
<td>Cartpole-PG</td>
<td>4/2</td>
<td>83.8</td>
<td>81.6</td>
</tr>
<tr>
<td>MountainCar-DQN</td>
<td>3/3</td>
<td>69.7</td>
<td>57.8</td>
</tr>
<tr>
<td>Taxi-SARSA</td>
<td>4/6</td>
<td>68.2</td>
<td>74.2</td>
</tr>
<tr>
<td>LunarLander-DDQN</td>
<td>8/4</td>
<td>68.4</td>
<td>63.7</td>
</tr>
<tr>
<td>BipedalWalker-PPO</td>
<td>14/4</td>
<td>56.9</td>
<td>56.4</td>
</tr>
<tr>
<td>Starcraft-A2C</td>
<td>9/4</td>
<td>94.7</td>
<td>91.8</td>
</tr>
</tbody>
</table>

Stochastic AND-OR Templates for visual objects

Zhangzhang Si & Song-Chun Zhu 2013. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35, (9), 2189-2205, doi:10.1109/TPAMI.2013.35.
Framework for vision: AND-OR Graphs

- Algorithm for this framework
  - Top-down/bottom-up computation
- Generalization of small sample
  - Use Monte Carlos simulation to synthesis more configurations
- Fill semantic gap

Images credit to Zhaoyin Jia (2009)
Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- **Terminal (leaf) node:** $T(pg)$
- **And-Or node:** $V^{or}(pg), V^{and}(pg)$
- **Set of links:** $E(pg)$
- **Switch variable at Or-node:** $w(t)$
- **Attributes of primitives:** $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_v(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$
Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: \( T(pg) \)
- And-Or node: \( V^{or}(pg), V^{and}(pg) \)
- Set of links: \( E(pg) \)
- Switch variable at Or-node: \( w(t) \)
- Attributes of primitives: \( \alpha(t) \)

\[
p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\zeta(pg))
\]

\[
\zeta(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_t(\alpha(t))
\]

\[
+ \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})
\]

SCFG: weigh the frequency at the children of or-nodes
Terminal (leaf) node: \( T(pg) \)

And-Or node: \( V^{or}(pg), V^{and}(pg) \)

Set of links: \( E(pg) \)

Switch variable at Or-node: \( w(t) \)

Attributes of primitives: \( \alpha(t) \)

\[
p(pg; \Theta, R, \Lambda) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))
\]

\[
\xi(pg) = \sum_{v \in V^{or}(pg)} \lambda_v(w(v)) + \sum_{v \in V^{and}(pg) \cup T(pg)} \lambda_v(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})
\]

Weigh the local compatibility of primitives (geometric and appearance)
Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^\text{or}(pg), V^\text{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V^\text{or}(pg)} \lambda_i(w(v)) + \sum_{v \in V^\text{and}(pg) \cup T(pg)} \lambda_i(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$

Spatial and appearance between primitives (parts or objects)
Terminal (leaf) node: $T(pg)$

And-Or node: $V_{or}^{\text{pg}}, V_{\text{and}}^{\text{pg}}$

Set of links: $E(pg)$

Switch variable at Or-node: $w(t)$

Attributes of primitives: $\alpha(t)$

$$p(pg; \Theta, R, \Delta) = \frac{1}{Z(\Theta)} \exp(-\xi(pg))$$

$$\xi(pg) = \sum_{v \in V_{or}^{pg}} \lambda_v(w(v)) + \sum_{v \in V_{\text{and}}^{pg} \cup T(pg)} \lambda_q(\alpha(t)) + \sum_{(i,j) \in E(pg)} \lambda_{ij}(v_i, v_j, \gamma_{ij}, \rho_{ij})$$
Input: an input image $I$, and a set of constructed And-Or graphs of compositional object categories.
Output: a parsing graph $pg$ of the scene that consists of the parsing graphs of detected objects.

- Repeat the following steps

1. Schedule the next node $A$ to visit from the candidate parts.
2. Call Bottom-up($A$) to update $A$’s open list.
   - Detect terminal instances of $A$ from the image.
   - Bind non-terminal instances of $A$ from its children’s open or closed lists
3. Call Top-down($A$) to update $A$’s open or closed lists.
   - Accept hypotheses from $A$’s open list to its closed list.
   - Remove (or disassemble) hypotheses from $A$’s closed list.
   - Update the open lists for particles that overlap with node $A$.
- Until the particles in open list with weights higher than the empirical threshold are exhausted. Output all parsing graphs whose root nodes are reached.

Example: Bayesian Rule Lists

Example: Interpretable Decision Sets

04 Examples for Post Hoc Models
(e.g. LIME, BETA, LRP)
Differences: Post-hoc versus Ante-hoc

- **Post-Hoc** (latin) = after-this (event), i.e. such approaches provide an explanation for a specific solution of a “black-box” approach, e.g. LIME, BETA, LRP, ... (see module 5)

- **Ante-hoc** (latin) = before-this (event), i.e. such methods can be (human) interpreted immanently in the system, i.e. they are transparent by nature (glass box), similar to the "interactive machine Learning" (iML) model.

- Note: Many ante-hoc approaches appear to the new student particularly novel, but these have a long tradition and were used since the early beginning of AI and applied in expert systems (see module 3); typical methods decision trees, linear regression, and Random Forests.

Caveat – Post hoc explanation can be misleading!

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.
Example LRP Layer-Wise Relevance Propagation


\[ f(x) = \sum \text{Feature Relevances} = \sum \text{Pixel Relevances} \]
A NN-classifier during prediction time

\[ f(x) = \cdots = \sum_{d \in \mathbb{R}^{l+1}} R_d^{(l+1)} = \sum_{d \in \mathbb{R}^l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)} \]

Example Taylor Decomposition

Heatmap Computation


a.holzinger@human-centered.ai
Pixel-wise decomposition for bag-of-words features

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek

a.holzinger@human-centered.ai

Last update: 11-10-2019
**Definition 1.** A heatmapping \( R(x) \) is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

\[
\forall \ x: \ f(x) = \sum_p R_p(x).
\]

**Definition 2.** A heatmapping \( R(x) \) is *positive* if all values forming the heatmap are greater or equal to zero, that is:

\[
\forall \ x, \ p: \ R_p(x) \geq 0
\]

**Definition 3.** A heatmapping \( R(x) \) is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with Definitions 1 and 2.

Sensitivity Analysis vs. Decomposition

function to analyze:
\[ f(x) = \max(0, x_1) + \max(0, x_2) \]

sensitivity analysis:
\[ (\partial f/\partial x_1)^2 = 1_{x_1 > 0} \]
\[ (\partial f/\partial x_2)^2 = 1_{x_2 > 0} \]

decomposition:
\[ R_1(x) = \max(0, x_1) \]
\[ R_2(x) = \max(0, x_2) \]
Relevance propagation
Relevance Redistribution
Example 1
LIME – Local Interpretable Model Agnostic Explanations

Example LIME – Model Agnostic Explanation

In [12]:
```python
explainer = lime.lime_tabular.LimeTabularExplainer(X_train, feature_names=breast.feature_names, class_names=breast.target_names)
```

Here we will take a sample from the test set (in this case the sample at index 76) and create an explainer instance for this sample. This will let us see why the algorithm made its prediction visually.

In [18]:
```python
# For this demonstration, let’s take the same sample each time, in this case sample index 86
i = 76
# For a random sample uncomment out the following line
# i = np.random.randint(0, X_test.shape[0])

exp = explainer.explain_instance(X_test[i], random_forest.predict_proba, num_features=4)
exp.show_in_notebook(show_table=True, show_all=False)
```

As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is malignant.

When using the explainer, we set the `num_features` parameter to 4, meaning the explainer shows the top 4 features that contributed to the prediction probabilities.

We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature’s contribution to the probability):
Remember: there are myriads of classifiers ...


a.holzinger@human-centered.ai
Black Box Explanations through Transparent Approximations

If Age < 50 and Male = Yes:
    If Past-Depression = Yes and Insomnia = No and Melancholy = No, then Healthy
    If Past-Depression = Yes and Insomnia = Yes and Melancholy = Yes and Tiredness = Yes, then Depression

If Age ≥ 50 and Male = No:
    If Family-Depression = Yes and Insomnia = No and Melancholy = Yes and Tiredness = Yes, then Depression
    If Family-Depression = No and Insomnia = No and Melancholy = No and Tiredness = No, then Healthy

Default:
    If Past-Depression = Yes and Tiredness = No and Exercise = No and Insomnia = Yes, then Depression
    If Past-Depression = No and Weight-Gain = Yes and Tiredness = Yes and Melancholy = Yes, then Depression
    If Family-Depression = Yes and Insomnia = Yes and Melancholy = Yes and Tiredness = Yes, then Depression

05 Principles of Making Neural Networks transparent
Example: Interpretable Deep Learning Model

Visualizing a Conv Net with a De-Conv Net

The world is compositional (Yann LeCun)

The world is compositional (Yann LeCun)


a.holzinger@human-centered.ai 65
06 Explanation
Interfaces: Future
Human-AI Interaction
Explanation is a reasoning process

Open questions:

- What is a good explanation?
- When is it enough (degree of saturation)?
- Context dependent (Emergency vs. research)
- How can we measure the degree of comprehensibility of a given explanation
- (obviously the explanation was good when it has been understood by the human)
- What can the system learn from the human?
- What can the human learn from the system?
- Measuring explanation effectiveness!

Explainability := a property of a system ("the AI explanation")
Causability := a property of a person ("the Human explanation")
Our goal is to provide effective mapping

- **Causability**: a property of a person (Human)
- **Explainability**: a property of a system (Computer)
Combination of Deep Learning with Ontologies

(1) Explaining the reasons for judgment

Deep Tensor

Output both inference result and reasons (inference factors)

Inference result

Inference factors

Knowledge Graph

Knowledge graph generates a logical path from input to the inference result

(2) Explaining the basis (evidence) for judgment

Basis formation

Explainable AI with Deep Tensor and Knowledge Graph

What is a good explanation?
(obviously if the other did understand it)
Experiments needed!
What is explainable/understandable/intelligible?
When is it enough (Sättigungsgrad – you don’t need more explanations – enough is enough)
But how much is it ...
Explanations in Artificial Intelligence will be necessary

- Justification, Explanation and Causality
- Trust > scaffolded by justification of actions (why)
- Please always take into account the inherent uncertainty and incompleteness of medical data!

IBM is doing it now: teaching meaningful explanations

Teaching Meaningful Explanations

Noel C. F. Codella,* Michael Hind,† Karthikeyan Natesan Ramamurthy,* Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei, Aleksandra Mojsilović

* These authors contributed equally.

IBM Research
Yorktown Heights, NY 10598
{nccodell,hindm,knatesa,mcam,adhuran,krvarshn,dwei,aleksand}@us.ibm.com

Abstract

The adoption of machine learning in high-stakes applications such as healthcare and law has lagged in part because predictions are not accompanied by explanations comprehensible to the domain user, who often holds ultimate responsibility for decisions and outcomes. In this paper, we propose an approach to generate such explanations in which training data is augmented to include, in addition to features and labels, explanations elicited from domain users. A joint model is then learned to produce both labels and explanations from the input features. This simple idea ensures that explanations are tailored to the complexity expectations and domain knowledge of the consumer. Evaluation spans multiple modeling techniques on a simple game dataset, an image dataset, and a chemical odor dataset, showing that our approach is generalizable across domains and algorithms. Results demonstrate that meaningful explanations can be reliably taught to machine learning algorithms, and in some cases, improve modeling accuracy.

1 Introduction

New regulations call for automated decision making systems to provide “meaningful information” on the logic used to reach conclusions [11–13]. Selbst and Powles interpret the concept of “meaningful information” as information that should be understandable to the audience (potentially individuals
07 Metrics of xAI
Central question: How can we measure whether and to what extent an “explanation” given by a machine has been understood by a human?

Therefore we need to know:

(1) the “goodness” of explanations,
(2) the “satisfaction” of the user
(3) the “understandability”
(4) the “trustability”
(5) the “human-AI interaction”

Please note that the terms are in “quotation marks” because it is extremely difficult to measure!

Bias and Fairness: Is AI more objective than Humans?


a.holzinger@human-centered.ai
Bias is the real AI Danger!

https://www.youtube.com/watch?v=EuBBz3bl-aA

https://www.youtube.com/watch?v=fMym_BKWQzk

For the love of us all...

"There are errors in these systems which propagate very quickly. Because of their scale of their action space – they can be hitting a billion or two billion users per day – that means the costs of getting it wrong are very very high."

-Mustafa Suleyman
cofounder DeepMind
How to measure “trust”? 

<table>
<thead>
<tr>
<th>Items in the Cognitive Effort construct</th>
<th>Mean</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Organized view</td>
<td>List view with “why”</td>
<td></td>
</tr>
<tr>
<td>I easily found the information I was looking for (reverse scale);</td>
<td>2.47</td>
<td>3.07</td>
</tr>
<tr>
<td>Selecting a product using this interface required too much effort.</td>
<td>2.61</td>
<td>3.14</td>
</tr>
</tbody>
</table>

Cronbach’s alpha = 0.73


### Trust Measurement

<table>
<thead>
<tr>
<th></th>
<th>Perceived Competence</th>
<th>Intention to Return</th>
<th>Cognitive Effort</th>
<th>Completion Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived Competence</td>
<td>1</td>
<td>.778** (.000)</td>
<td>-.826** (.000)</td>
<td>-.018 (.830)</td>
</tr>
<tr>
<td>Intention to Return</td>
<td>.778** (.000)</td>
<td>1</td>
<td>-.675** (.000)</td>
<td>-.042 (.619)</td>
</tr>
<tr>
<td>Cognitive Effort</td>
<td>-.826** (.000)</td>
<td>-.675** (.000)</td>
<td>1</td>
<td>.069 (.414)</td>
</tr>
<tr>
<td>Completion Time</td>
<td>-.018 (.830)</td>
<td>-.042 (.619)</td>
<td>.069 (.414)</td>
<td>1</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed).
Measuring Causability:
Mapping machine explanations with human understanding
The process of explanations (System Causability Scale)

Computational approaches can find in $R^n$ what no human is able to see

However, still there are many hard problems where a human expert in $R^2$ can understand the context and bring in experience, expertise, knowledge, intuition, ...

Black box approaches can not explain WHY a decision has been made ...
Thank you!