From Machine Learning to Explainable AI

Assoc.Prof. Dr. Andreas Holzinger

Holzinger-Group, Human-Centered AI
Institute for Medical Informatics/Statistics, Medical University Graz, Austria
&
Institute of interactive Systems & Data Science
Graz University of Technology, Austria

andreas.holzinger@human-centered.ai
http://human-centered.ai
Andreas Holzinger: Background

- PhD in Cognitive Science 1998
- Habilitation Computer Science 2003
- Lead Holzinger Group HC-AI
  [www.human-centered.ai](http://www.human-centered.ai)
- Visiting Professor for Machine Learning in Health Informatics: TU Vienna, Univ. Verona, UCL London, RWTH Aachen
- Research statements:
AI = Artificial Intelligence
ML = Machine Learning
DL = Deep Learning
aML = automatic (autonomous) ML
iML = interactive ML
HCI = Human-Computer Interaction
KDD = Knowledge Discovery from Data
HC-AI = Human-Centered AI
Ex-AI = explainable AI

01 What is HC-AI?
02 Probabilistic Machine Learning
03 aML
04 iML
05 Why Explainability?
06 Methods
01 What is the HC-AI approach?

Privacy 4 – Transparency, Accountability, Ethics

1 - Data
2 - Learning
3 - Visualization

Space and Time 5 - Graphs, 6-Topology, 7-Entropy


Our goal is that human values are aligned to ensure responsible machine learning.

Holzinger, A. (2013). Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Lecture Notes in Computer Science 8127 (pp. 319-328)
- ML is a very practical field – algorithm development is at the core – however, successful ML needs a concerted effort of various topics...
... successful ML needs ...

concerted effort

international

without boundaries ...

Image Source: http://www.bach-cantatas.com
CD-MAKE 2019
Cross Domain Conference for Machine Learning and Knowledge Extraction

Image with friendly permission of Michael D. Beckwith
02 Probabilistic Machine Learning
Probability theory is nothing but common sense reduced to calculation ...

Pierre Simon de Laplace (1749-1827)
Learning from data

The “inverse probability” allows to learn from previous data and to make predictions.

\[ D = x_{1:n} = \{ x_1, x_2, \ldots, x_n \} \]

\[ p(\theta | D) = \frac{p(D | \theta) \ast p(\theta)}{p(D)} \]
Learning and Inference

$d \ldots \text{data}$

$h \ldots \text{hypotheses}$

$\mathcal{H} \ldots \{H_1, H_2, \ldots, H_n\} \quad \forall \ h, d \ldots$

$$p(h|d) = \frac{p(d|h) \cdot p(h)}{\sum_{h' \in H} p(d|h') \cdot p(h')}$$

Prior Probability

Likelihood

Posterior Probability

Problem in $\mathbb{R}^n \rightarrow \text{complex}$

Feature parameter $\theta$
Fully automatic → Goal: Taking the human out of the loop

### Algorithm 1 Bayesian optimization

1. for $n = 1, 2, \ldots$ do
2. select new $x_{n+1}$ by optimizing acquisition function $\alpha$
   \[
   x_{n+1} = \arg \max_x \alpha(x; D_n)
   \]
3. query objective function to obtain $y_{n+1}$
4. augment data $D_{n+1} = \{D_n, (x_{n+1}, y_{n+1})\}$
5. update statistical model
6. end for

- **PI**: Probability of Improvement
- **EI**: Expected Improvement
- **UCB**: Upper Confidence Bound
- **TS**: Thompson Sampling
- **PES**: Predictive Entropy Search

03 aML
Fully automatic autonomous vehicles

Autonomous aerial vehicle (AAV): passenger drone


http://www.ehang.com/ehang184/
... and thousands of industrial aML applications ...

Big Data is necessary for aML!

Sometimes we do not have “big data”, where aML-algorithms benefit.

Sometimes we have

- Small amount of data sets
- Rare Events – no training samples
- NP-hard problems, e.g.
  - Subspace Clustering,
  - k-Anonymization,
  - Protein-Folding, ...
Not our Goal: Humanoid AI

Humanoid AI ≠ Human-level AI
To reach a level of human-level AI

- 1) ... learn from prior data
- 2) ... extract knowledge
- 2) ... generalize, i.e. guessing where a probability mass function concentrates
- 4) ... fight the curse of dimensionality
- 5) ... disentangle underlying explanatory factors of data, i.e.
- 6) ... understand the data in the context of an application domain
Motivation for the human-in-the-loop

- Humans can generalize even from few examples ...
  - They learn relevant representations
  - Can disentangle the explanatory factors
  - Find the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

10 million $200 \times 200$ px images from the Web

\[ x^* = \arg \min_x f(x; W, H), \text{ subject to } \|x\|_2 = 1. \]


Even Children can make inferences from little, noisy, incomplete data ...

Consequently ...

Sometimes we (still) need a human-in-the-loop
04 iML
iML := algorithms which interact with agents*) in a multi-agent system and can optimize their learning behaviour through this interaction

*) where the agents can be human

Interactive Machine Learning: Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions ...

iML: bringing in human intuition

Three examples for the usefulness of iML ...

- **Example 1: Subspace Clustering**
- **Example 2: k-Anonymization**
- **Example 3: Protein Design**


Input: ProblemSize, \( m, \beta, \rho, \sigma, q_0 \)

Output: \( P_{\text{best}} \)

\[
P_{\text{best}} \leftarrow \text{CreateHeuristicSolution}(\text{ProblemSize});
\]

\[
P_{\text{bestcost}} \leftarrow \text{Cost}(P_{\text{best}});
\]

\[
P_{\text{Pheromoneinit}} \leftarrow \frac{1.0}{\text{ProblemSize} \times P_{\text{bestcost}}};
\]

\[
P_{\text{Pheromone}} \leftarrow \text{InitializePheromone}(P_{\text{Pheromoneinit}});
\]

\[\textbf{while} \neg \text{StopCondition()} \textbf{do}\]

\[\quad \textbf{for} \ i = 1 \text{ to } m \textbf{ do}\]

\[\quad \quad S_i \leftarrow \text{ConstructSolution}(P_{\text{Pheromone}}, \text{ProblemSize}, \beta, q_0);\]

\[\quad \quad S_{i_{\text{cost}}} \leftarrow \text{Cost}(S_i);\]

\[\quad \quad \text{if } S_{i_{\text{cost}}} \leq P_{\text{bestcost}} \text{ then}\]

\[\quad \quad \quad P_{\text{bestcost}} \leftarrow S_{i_{\text{cost}}};\]

\[\quad \quad \quad P_{\text{best}} \leftarrow S_i;\]

\[\quad \quad \text{end}\]

\[\quad \text{LocalUpdateAndDecayPheromone}(P_{\text{Pheromone}}, S_i, S_{i_{\text{cost}}}, \rho);\]

\[\text{end}\]

\[\text{GlobalUpdateAndDecayPheromone}(P_{\text{Pheromone}}, P_{\text{best}}, P_{\text{bestcost}}, \rho);\]

\[\textbf{while} \ \text{isUserInteraction()} \textbf{ do}\]

\[\quad \text{GlobalAddAndRemovePheromone}(P_{\text{Pheromone}}, P_{\text{best}}, P_{\text{bestcost}}, \rho);\]

\[\text{end}\]

\[\textbf{end}\]

\[\textbf{return} \ P_{\text{best}};\]

05 Why Explainability?
Deep Convolutional Neural Network Pipeline


Limitations of Deep Learning approaches

- **Non-convex** difficult to set up, to train, to optimize, needs a lot of expertise, error prone, ...

- **Data intensive**, needs often millions of training samples ...

- “Black-Box” approaches – lack **transparency**, do not foster trust and acceptance among end-user, **legal** aspects make them difficult ...

Houston, we have a problem ...

a.holzinger@human-centered.ai

ACRAI, Vienna, 03.05.2019
Example: Classifier Errors

- Result of the classifier: **This is a horse**
- Why is this a horse?

Source: Image is in the public domain
Example: Context recognition understanding

Image Captions by deep learning: github.com/karpathy/neuraltalk2
State-of-the-Art of the Stanford Machine Learning Group

Benefits ...

Verify that algorithms/classifiers work as expected
Wrong decisions can be costly and dangerous ...

Understanding the weaknesses and errors
Detection of bias – bring in human intuition to know the error ...

Scientific replicability and causality
The “why” is often more important than the prediction ...

06 Methods of Explainability
State-of-the-art Methods

- 1) Gradients
- 2) Sensitivity Analysis
- 4) Optimization (LIME - Local interpretable model agnostic Explanations, BETA - Black Box Explanation through Transparent Approximation, ...)
- 5) Deconvolution and Guided Backpropagation
- 6) Concept Activation Vectors (CAV)
Explanation by Decomposition (general idea)

- Given: a prediction $f(x)$ over an input set $x = (x_1, \ldots, x_d)$
- Goal: Computing a relevance score $r_d(x)$ for each input $x_d$ in dimension $d$

\[
f(x) = \sum_{d=1}^{D} r_d(x)
\]

- Decompose the prediction depending on the test data
- $r_d(x) = ?$
- Looking for a linear mapping which can be a meaningful explanation for a human expert
Classification (extremely simplified)

\[ x_j = \sigma \left( \sum_i x_i w_{ij} + b_j \right) \]
Explanation (extremely simplified)

\[
\sum_p r_p = \ldots = \sum_i r_i = \sum_j r_j = \ldots = f(x)
\]
LRP Layer-Wise Relevance Propagation

\[ a_j^{(l+1)} = \sigma \left( \sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)} \right) \]

\[ R_i^{(l)} = \sum_j \frac{z_{ij}}{z_{i1}} R_j^{(l+1)} \]

Forward propagation

Layer-wise relevance propagation

\[ R_i = \left| \frac{\partial}{\partial x_i} f(x) \right| \]

\[ \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(x) \]
Example: Concept Activation Vector (CAV)

Humans work in another vector space which is spanned by implicit knowledge vectors corresponding to an unknown set of human interpretable concepts.

\[
\frac{\partial h_k(x)}{\partial x_{a,b}}
\]

\[
S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon} = \nabla h_{l,k}(f_l(x)) \cdot v_C^l
\]


- Causability := a property of a person (Human)
- Explainability := a property of a system (Computer)

We work on mapping Explainability with Causability for effective Human-AI interaction.
We need effective Human-AI mapping

Why did the algorithm do that?
Can I trust these results?
How can I correct an error?

A possible solution

The domain expert can understand why ...
The domain expert can learn and correct errors ...
The domain expert can re-enact on demand ...
Conclusion
1. Automatic approaches can find in \( \mathbb{R}^N \) what no human would be able to see.

2. Black box approaches cannot explain **why** a decision has been made ...

3. Enabling human experts to understand context, need **effective** mapping explanations to causability.
Thank you!