Andreas Holzinger
185.A83 Machine Learning for Health Informatics
2019S, VU, 2.0 h, 3.0 ECTS
Lecture 07 – Dienstag, 07.05.2019

Causality, Explainability, Ethical, Legal, and Social Issues of AI/ML in health

andreas.holzinger AT tuwien.ac.at
https://human-centered.ai/machine-learning-for-health-informatics-class-2019
Agenda

- 00 Reflection – follow-up from last lecture
- 01 Causality
- 02 Explainability and Causability
- 03 AI Ethics
- 04 Social implications of AI
00 Reflection
Five Mainstreams in Machine Learning

- **Symbolic ML**
  - First order logic, inverse deduction
  - Tom Mitchell, Steve Muggleton, Ross Quinlan, ...

- **Bayesian ML**
  - Statistical learning
  - Judea Pearl, Michael Jordan, David Heckermann, ...

- **Cognitive ML**
  - Analogisms from Psychology, Kernel machines
  - Vladimir Vapnik, Peter Hart, Douglas Hofstadter, ...

- **Connectionist ML**
  - Neuroscience, Backpropagation
  - Geoffrey Hinton, Yoshua Bengio, Yann LeCun, ...

- **Evolutionary ML**
  - Nature-inspired concepts, genetic programming
  - John Holland (1929-2015), John Koza, Hod Lipson, ...
Reflection from last lectures

1. Graph showing sensitivity vs. 100-specificity.
2. Diagram illustrating the relationship between uncertainty, diagnosis, and choice.
3. Figure depicting the signal-to-noise ratio with probability distribution and criterion.
4. Decision tree for cancer treatment options.
Key Challenges

- Remember: Medicine is an complex application domain – dealing most of the time with **probable information**!

- Some challenges include:

  - (a) defining hospital system architectures in terms of generic tasks such as diagnosis, therapy planning and monitoring to be executed for (b) medical reasoning in (a);
  - (c) patient information management with (d) minimum uncertainty.

- Other challenges include: (e) knowledge acquisition and encoding, (f) human-ai interface and ai-interaction; and (g) system integration into existing clinical legacy and proprietary environments, e.g. the enterprise hospital information system; to mention only a few.
01 Causality
- David Hume (1711-1776): Causation is a matter of perception: observing fire > result feeling heat
- Karl Pearson (1857-1936): Forget Causation, you should be able to calculate correlation
- Judea Pearl (1936-): Be careful with purely empirical observations, instead define causality based on known causal relationships, and beware of counterfactuals ...


What is a counterfactual? (and see Slides 21-23)

- Hume again: “... if the first object had not been, the second never had existed ...”
- Causal inference as a missing data problem
- \( x_i : = f_i(\text{ParentsOf}_i, \text{Noise}_i) \)
- Interventions can only take place on the right side

Remember: Correlation is NOT Causality

Dependence vs. Causation

Hans Reichenbach (1891-1953): **Common Cause Principle**

Links causality with probability:

- If X and Y are statistically dependent, there is a Z influencing both
- Whereas:
  - A, B, ... events
  - X, Y, Z random variables
  - P ... probability measure
  - Px ... probability distribution of X
  - p ... probability density
  - p(X) .. Density of Px
  - p(x) probability density of Px evaluated at the point x

Correlation does not tell anything about causality!


https://plato.stanford.edu/entries/physics-Rpcc/
- $X_1, \ldots, X_n$ ... set of observables
- Draw a directed acyclic graph $G$ with nodes $X_1, \ldots, X_n$

Parents = direct causes

$x_i: = f_i(\text{ParentsOf}_i, \text{Noise}_i)$

Remember: Noise means unexplained (exogenous) and denote it as $U_i$

Question: Can we recover $G$ from $p$?
Answer: under certain assumptions, we can recover an equivalence class containing the correct $G$ using conditional independence testing
But there are problems!

Remember: the mapping is important

<table>
<thead>
<tr>
<th>Explainability</th>
<th>in a technical sense highlights decision-relevant parts of the used representations of the algorithms and active parts in the algorithmic model, that either contribute to the model accuracy on the training set, or to a specific prediction for one particular observation. It does not refer to an explicit human model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causability</td>
<td>as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use.</td>
</tr>
</tbody>
</table>

- Causability := a property of a person, while
- Explainability := a property of a system
Probabilistic vs. causal inference problems

Compare this with usability

“How do humans generalize from few examples?”

- Learning relevant representations
- Disentangling the explanatory factors
- Finding the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$


Decide if $X \rightarrow Y$, or $Y \rightarrow X$ using only observed data

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf
- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C

- **Inductive reasoning** = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions

- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.
Important Definition: Ground truth

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
    - Causal inference is an example of causal reasoning.
Empirical Inference Example

Gottfried W. Leibniz (1646-1716)
Hermann Weyl (1885-1955)
Vladimir Vapnik (1936-)
Alexey Chervonenkis (1938-2014)
Gregory Chaitin (1947-)
Remember: hard inference problems

- High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent

What makes it hard ... ?

Example 3.4 (Eye disease) There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured ($B = 0$); if untreated, these patients turn blind within a day ($B = 1$). For the remaining 1%, the treatment has the opposite effect and they turn blind ($B = 1$) within a day. If untreated, they regain normal vision ($B = 0$).

Which category a patient belongs to is controlled by a rare condition ($N_B = 1$) that is unknown to the doctor, whose decision whether to administer the treatment ($T = 1$) is thus independent of $N_B$. We write it as a noise variable $N_T$.

Assume the underlying SCM

\[
\begin{align*}
T & := N_T \\
B & := T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{align*}
\]

with Bernoulli distributed $N_B \sim \text{Ber}(0.01)$; note that the corresponding causal graph is $T \rightarrow B$.

Now imagine a specific patient with poor eyesight comes to the hospital and goes blind ($B = 1$) after the doctor administers the treatment ($T = 1$). We can now ask the counterfactual question “What would have happened had the doctor administered treatment $T = 0$?” Surprisingly, this can be answered. The observation $B = T = 1$ implies with (3.5) that for the given patient, we had $N_B = 1$. This, in turn, lets us calculate the effect of $do(T := 0)$.

To this end, we first condition on our observation to update the distribution over the noise variables. As we have seen, conditioned on $B = T = 1$, the distribution for $N_B$ and the one for $N_T$ collapses to a point mass on 1, that is, $\delta_1$. This leads to a modified SCM:
\[ \mathcal{C}|B = 1, T = 1 : \begin{align*}
T & := 1 \\
B & := T \cdot 1 + (1 - T) \cdot (1 - 1) = T
\end{align*} \tag{3.6} \]

Note that we only update the noise distributions; conditioning does not change the structure of the assignments themselves. The idea is that the physical mechanisms are unchanged (in our case, what leads to a cure and what leads to blindness), but we have gleaned knowledge about the previously unknown noise variables for the given patient.

Next, we calculate the effect of \( do(T = 0) \) for this patient:

\[ \mathcal{C}|B = 1, T = 1; do(T := 0) : \begin{align*}
T & := 0 \\
B & := T
\end{align*} \tag{3.7} \]

Clearly, the entailed distribution puts all mass on \((0, 0)\), and hence

\[ p^{\mathcal{C}|B=1,T=1;do(T:=0)}(B = 0) = 1. \]

This means that the patient would thus have been cured \((B = 0)\) if the doctor had not given him treatment, in other words, \( do(T := 0) \). Because of

\[ p^{\mathcal{C};do(T:=1)}(B = 0) = 0.99 \quad \text{and} \quad p^{\mathcal{C};do(T:=0)}(B = 0) = 0.01, \]

however, we can still argue that the doctor acted optimally (according to the available knowledge). \( \square \)
Interestingly, Example 3.4 shows that we can use counterfactual statements to falsify the underlying causal model (see Section 6.8). Imagine that the rare condition $N_B$ can be tested, but the test results take longer than a day. In this case, it is possible that we observe a counterfactual statement that contradicts the measurement result for $N_B$. The same argument is given by Pearl [2009, p.220, point (2)]. Since the scientific content of counterfactuals has been debated extensively, it should be emphasized that the counterfactual statement here is falsifiable because the noise variable is not unobservable in principle but only at the moment when the decision of the doctor has to be made.

Remember: Medical Action = Decision Making
Search Task in $\mathcal{H}$
Problem: Time ($t$)
02 Explainability & Causability
What can we learn out of this?

Explaining individual classification decisions

**Linear classification**
- \( S_1 \): sepal width
- \( S_2 \): sepal width
- \( V_1 \): sepal width

**Non-linear classification**
- \( S_1 \): sepal width & length
- \( S_2 \): sepal width
- \( V_1 \): sepal length

How can we map these to effectively?

- Causability := a property of a person (Human)
- Explainability := a property of a system (Computer)
**We need effective tools for Human-AI Interaction**

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

**A possible solution**

- **Explanation Interface**
- **Explainable Model**

*The domain expert can understand why ...*
*The domain expert can learn and correct errors ...*
*The domain expert can re-enact on demand ...*
Example for an Explanation Interface


\[
f(x) = \sum \text{Feature Relevances} = \sum \text{Pixel Relevances}
\]
The simplest possibility ...

Fig 3. An exemplary real-valued prediction function for classification with the dashed black line being the decision boundary which separates the blue from the green dots. The blue dots are labeled negatively, the green dots are labeled positively. Left: Local gradient of the classification function at the prediction point. Right: Taylor approximation relative to a root point on the decision boundary. This figure depicts the intuition that a gradient at a prediction point $x$—here indicated by a square—does not necessarily point to a close point on the decision boundary. Instead it may point to a local optimum or to a far away point on the decision boundary. In this example the explanation vector from the local gradient at the prediction point $x$ has a too large contribution in an irrelevant direction. The closest neighbors of the other class can be found at a very different angle. Thus, the local gradient at the prediction point $x$ may not be a good explanation for the contributions of single dimensions to the function value $f(x)$. Local gradients at the prediction point in the left image and the Taylor root point in the right image are indicated by black arrows. The nearest root point $x_0$ is shown as a triangle on the decision boundary. The red arrow in the right image visualizes the approximation of $f(x)$ by Taylor expansion around the nearest root point $x_0$. The approximation is given as a vector representing the dimension-wise product between $Df(x_0)$ (the black arrow in the right panel) and $x - x_0$ (the dashed red line in the right panel) which is equivalent to the diagonal of the outer product between $Df(x_0)$ and $x - x_0$.

doi:10.1371/journal.pone.0130140.g003
How to generate a heatmap

**Fig 4.** Local and global predictions for input images are obtained by following a series of steps through the classification- and pixel-wise decomposition pipelines. Each step taken towards the final pixel-wise decomposition has a complementing analogue within the Bag of Words classification pipeline. The calculations used during the pixel-wise decomposition process make use of information extracted by those corresponding analogues. Airplane image in the graphic by Pixabay user tpsdave.

doi:10.1371/journal.pone.0130140.g004
Relevance propagation

Montavon et al. (2017)
03 AI Ethics
What is Ethics?

- Ethics = moral philosophy
- Recommending and defending concepts of right and wrong conduct.
- Three areas:
  - 1) Meta-ethics, concerning the theoretical meaning and reference of moral propositions, and how their truth values (if any) can be determined
  - 2) Normative ethics, concerning the practical means of determining a moral course of action
  - 3) Applied ethics, concerning what a person is obligated (or permitted) to do in a specific situation or a particular domain of action -> AI ethics

https://www.iep.utm.edu/ethics/
What is Ethics for us as Engineers?

- Ethics is a **practical discipline**
- It is the good things – It is the right things
- How do we define what is good?
Biomedical Ethics is well-established

**UNESCO’s 15 Bioethical principles**

<table>
<thead>
<tr>
<th>Human dignity &amp; human rights</th>
<th>Benefit &amp; harm</th>
<th>Autonomy-individual responsibility</th>
<th>Consent</th>
<th>Persons without the capacity to consent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human vulnerability &amp; personal integrity</td>
<td>Privacy / Confidentiality</td>
<td>Equality, Justice, Equity</td>
<td>Non-discrimination</td>
<td>Respect for cultural diversity</td>
</tr>
<tr>
<td>Solidarity &amp; cooperation</td>
<td>Social responsibility &amp; health</td>
<td>Sharing of benefits</td>
<td>Protecting future generations</td>
<td>Protecting biodiversity, biosphere &amp; environment</td>
</tr>
</tbody>
</table>

Independent review and approval by ethics board:

1) Informed consent
2) Risk-Benefit ratio and minimization of risk
3) Fair selection of study population (inclusion-, exclusion-criteria)
4) Scientific validity (‘scholarly review’) 
5) Social value 
6) Respect for participants and study communities
7) Confidentiality and privacy, data security 
8) No Conflict of interest
Accountability ... we have to take responsibility for our developments, governments have to take responsibility for decisions and laws affecting all citizens
Trust ... confidence in the reliability, truth, ability (a trustee holds the property as its nominal owner for the good of beneficiaries
Transparency ... implies openness, communication, accountability, trust, ...
Understandability ... property of a system according to the principles of usability, we can say it is a kind of domain usability, and can be perceived as the relation and good fit between the “language of the human” and the “language of the machine”
Isaac Asimov three laws of robotics

The three laws of (fictional) robotics:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.
(Some) Grand Questions of AI Ethics

- Is it morally justified to create super-intelligent systems?
- Should our AI have any free will? And if it is possible: Can we prevent them from having free will?
- Will AI have consciousness? (Strong AI)
  - If so, will it accept to be constrained by artificial AI-ethics placed on them by humans?
- If AI develop their own ethics and morality, will we like what they do with us?

https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/
What about existing AI?
If the robot looks like a human, do we have different expectations?

Would you “kill” a robot car?

Would you “kill” a robot insect that would react by squeaky noises and escape in panic?

Would you “kill” a robot biped that would react by begging you to save his life?
04 Social Issues of AI
Watch the Obama Interview on how artificial intelligence will affect our jobs:

https://human-centered.ai/2016/10/14/obama-on-humans-in-the-loop
For sure explainability and ethical issues belong together ...
“Does your car have any idea why my car pulled it over?”
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,knatesan,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 [1-3]. Selbst and Powles interpret the concept of “meaningful information” as information that should be understandable to the audience (potentially individuals
Alexa, what about legal aspects of AI?
Legal aspects of AI

- mid-2018 active a/cs (bn):
  - Facebook: 2.2
  - YouTube: 1.9
  - WhatsApp: 1.5
  - WeChat: 1

- explosive growth of data volumes:
  - fuels big data analytics & AI
  - growing 10x every 5 yrs

- machine learning:
  - deep learning
  - unsupervised
  - supervised

- cheaper sensors/cameras:
  - speech (to/from text)
  - image recognition
  - machine vision

- mid-2018:
  - 7.6bn population
  - 20bn+ connected things
  - 5bn+ mobile users
  - 4bn+ Internet users

- year on year Cloud growth >50%

Conclusion: Human-in-control
Engineers create a set of logical rules to represent knowledge (Rule based Expert Systems)

- Advantage: works well in narrowly defined problems of well-defined domains
- Disadvantage: No adaptive learning behaviour and poor handling of $p(x)$

Image credit to John Launchbury
Engineers create learning models for specific tasks and train them with “big data” (e.g. Deep Learning).

- Advantage: works well for standard classification tasks and has prediction capabilities.
- Disadvantage: No contextual capabilities and minimal reasoning abilities.
- A contextual model can perceive, learn and understand and abstract and reason
- Advantage: can use transfer learning for adaptation on unknown unknowns
- Disadvantage: Superintelligence ...
Three (selected) dangers and myths about AI generally

- Myth 1a: Superintelligence by 2100 is inevitable!
- Myth 1b: Superintelligence by 2100 is impossible!
- Fact: We simply don’t know it!
- Myth 2: Robots are our main concern
  Fact: Cyberthreats are the main concern: it needs no body – only an Internet connection
- Myth 3: AI can never control us humans
  Fact: Intelligence is an enabler for control: We control tigers by being smarter ...
The future: human-in-control

High-performance medicine: the convergence of human and artificial intelligence

Eric J. Topol

Nature Medicine 25, 44–56 (2019) | Download Citation

Towards a Code of Ethics for Artificial Intelligence

Paula Boddington

Machine Learning Health 07
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 ...

Thank you!