185.A83 Machine Learning for Health Informatics  
2020S, VU, 2.0 h, 3.0 ECTS  
Andreas Holzinger, Rudolf Freund  
Marcus Bloice, Florian Endel, Anna Saranti

Lecture 01 – Introduction

From health informatics to ethical responsible medical AI

Contact: andreas.holzinger AT tuwien.ac.at  
@aholzinz #KandinskyPatterns


Learning Goals

At the end of this lecture you will ...  
- ... see why an integrative approach is important  
- ... understand that machine learning can help medicine  
- ... be fascinated by the possibilities of machine learning  
- ... disillusioned at the limits of machine learning  
- ... be aware of the complexity of the health domain  
- ... realize why a human-in-the-loop is sometimes crucial  
- ... identify the main challenges in this area  
- ... recognize some ideas of future work

Advance Organizer (1/2)

- **Bioinformatics** – discipline, as part of biomedical informatics, at the interface between biology and information science and mathematics; processing of biological data;  
- **Biomarker** – a characteristic (e.g., body temperature (fever) as a biomarker for an infection, or proteins measured in the urine) as an indicator for normal or pathogenic biological processes, or pharmacologic responses to therapeutic intervention;  
- **Biomedical data** – composed of general data, it is characterized by large volumes, complex structures, high dimensionality, evolving biological concepts, and insufficient data modeling practices;  
- **Biomedical Informatics** – 2011 definition: similar to medical informatics but including the optimal use of biomedical data, e.g. from genomics, proteomics, metabolomics;  
- **Classical Medicine** – is both the science and the art of healing and encompasses a variety of practices to maintain and restore health;  
- **Genomics** – branch of molecular biology which is concerned with the structure, function, mapping & evolution of genomes;  
- **Interactive Machine Learning** – as defined by algorithms that can interact with both computational agents and human agents and can optimize their learning behavior through these interactions, by bringing in a human into the loop;  
- **Machine Learning** – addresses the question of how to design algorithms that improve automatically through experience from big data – doing it automatically (AM) without a human-in-the-loop;  
- **Medical Informatics** – 1970 definition: “... scientific field that deals with the storage, retrieval, and optimal use of medical information, data, and knowledge for problem solving and decision making”; – see the better 2011 definition by the AMIA  
- **Metabolomics** – study of chemical processes involving metabolites (e.g. enzymes). A challenge is to integrate proteomic, transcriptomic, and metabolomic information to provide a more complete understanding of living organisms;  
- **Molecular Medicine** – emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs;
(0) A few definitions first

- **Omics data**: data from e.g. genomics, proteomics, metabolomics, etc.
- **Pervasive Computing**: similar to ubiquitous computing (Ubicomp), a post-desktop model of Human-Computer Interaction (HCI) in which information processing is integrated into everyday, miniaturized and embedded objects and activities, having some degree of "intelligence".
- **Pervasive Health**: all noninvasive, analytical, diagnostic, support etc. information functions to improve health care, e.g. remote, automated patient monitoring, diagnosis, home care, self-care, independent living, etc.;
- **Proteome**: the entire complement of proteins that is expressed by a cell, tissue, or organism;
- **Proteomics**: field of molecular biology concerned with determining the proteome;
- **P-Health Model**: Preventive, Participatory, Pre-emptive, Personalized, Predictive, Pervasive (= available to anybody, anytime, anywhere);
- **Space**: a set with some added structure;
- **Technological Performance**: machine "capabilities", e.g. short response time, high throughput, high availability, etc.
- **Time**: a dimension in which events can be ordered along a time line from the past through the present into the future;
- **Translational Medicine**: based on interventionist epidemiology; progress of Evidence-Based Medicine (EBM), integrates research from basic science for patient care and prevention;
- **von-Neumann-Computer**: a 1945 architecture, which still is the predominant machine architecture of today (e.g. Non-Vons, incl. analogue, optical, quantum computers, cell processors, DNA arid neural nets (in silico));

(0) A few definitions – for mutual understanding

- (1) **Machine Learning Health examples**
- (2) **A brief look at the application area health**
- (3) **Statistical Machine Learning**
- (4) **Automatic Machine Learning (aML)**
- (5) **Interactive Machine Learning (iML)** and why we need the human-in-the-loop
- (6) **Explainable AI and Methods of Explainability**
- (7) **Causability – Measuring the Quality of (6)**

---

**Health, Medicine, Informatics, Biomedical Informatics, ... ?**

- **Health**: state of physical, mental and social well-being
- **Medicine**: art, science, practice of patient diagnosis, prognosis, prevention, treatment of injury or disease.
- **Informatics**: study of information processing (Schrodinger [1]: Life is Information Processing)
- **Biomedical informatics (BMI)**: interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific problem solving, and decision making, motivated by efforts to improve human health [2]


Artificial Intelligence, Machine Learning, Deep Learning, ...?

- AI := not clearly defined “umbrella term” of “making intelligent machines”
- ML := the workhorse of AI - algorithms that improve by learning from data
- DL := a particular family of neural networks - currently very successful

What was the first definition of Machine Learning?

- Some studies in machine learning using the game of checkers

What is Standardized Medicine versus Personalized Medicine?

- Standardized Medicine
- Personalized Medicine
- Pervasive Health

Preventive, Participatory, Pre-emptive, Pervasive, Predictive, Personalized?

- Preventive
- Participatory
- Pre-emptive
- Pervasive
- Predictive
- Personalized


(1) Machine learning health examples
Why is Machine Learning in health enormously progressing?

- Progress is driven by the explosion in the availability of big data and low-cost computation.
- Health concerns everyone ...


What are open general questions?

(1) Why can AI solve some tasks better than humans?

(2) How did AI get these results in the first place?

(3) What happens if I change, replace, disturb, remove, ... input data?

What do we need to make AI even more successful?

Robustness & Explainability

\[ x + 0.04 \times sign(\nabla_x f(\theta, x, y)) = x + \epsilon \times sign(\nabla_x f(\theta, x, y)) \]


https://github.com/slgfin/adversarial-medicine

What do we need to reach robust intelligence?

- 1) learning from few data
- 2) extracting knowledge
- 3) generalize
- 4) fight the curse of dimensionality
- 5) disentangle the independent explanatory factors of data, i.e.
- 6) causal understanding of the data in the context of an application domain
(2) Application Area
Health

Why is health a complex application domain?

Our central hypothesis:
Information may bridge this gap

Why is probabilistic information so important in medicine?

- 400 BC Hippocrates (460-370 BC):
  - A medical record should reflect the course of a disease ...
  - ... and should indicate the probable cause of a disease

- 1890 William Osler (1849-1919):
  - Medicine is a science of uncertainty and an art of probabilistic decision making

- Today
  - Prediction models are based on data features, the patient health status is modelled as high-dimensional feature vectors ...

What are the main problems in “building a bridge”?

- Heterogeneity
- Dimensionality
- Complexity
- Uncertainty

(3) Probable Information and Statistical Machine Learning

Who laid the foundations for modern statistical machine learning?

"Il est remarquable qu'une science qui a commencé avec l'ère la prise en compte des jeux de hasard ... aurait dû devenir l'objet le plus important de la connaissance humaine."

Pierre Simon de Laplace, 1812

Who inspired Laplace for his work?

Bayes' Rule is a corollary of the Sum Rule and Product Rule:

\[ p(x|y) = \frac{p(y|x)p(x)}{\sum p(x_i,y_j)p(x_i)} \]

Probability Theory is nothing, but common sense reduced to calculation...

Pierre-Simon Laplace (1749-1827)


Probability theory is the formal study of laws of chance and managing uncertainty; allows to measure (many) events

- Frequentist* view: coin toss
- Bayesian* view: probability as a measure of belief (this is what made machine learning successful)
- \( p(x) = 1 \) means that all events occur for certain
- Information is a measure for the reduction of uncertainty
- If something is 100% certain its uncertainty = 0
- Uncertainty is max. if all choices are equally probable (i.i.d. = independent and identically distributed)
- Uncertainty (as information) sums up for independent sources: \( \sum x p(x = X) = 1 \)

*) Bayesian vs. Frequentist: please watch the excellent video of Kristin Lennox (2016): https://www.youtube.com/watch?v=Q05DhyO6QY
Why is the work of Laplace so important for us?

What is the simplest mathematical operation for us?

\[ p(x) = \sum_{x} p(x, y) \]  
\[ p(x, y) = p(y|x) * p(x) \]  
Laplace (1773) showed that we can write:

\[ p(x, y) * p(y) = p(y|x) * p(x) \]  
Now we introduce a third, more complicated operation:

\[ \frac{p(x, y) * p(y)}{p(y)} \]  
We can reduce this fraction by \( p(y) \) and we receive what is called Bayes rule:

\[ p(x, y) = \frac{p(x|y) * p(y)}{p(y)} \]  
\[ p(h|d) \propto p(d|h) * p(h) \]  
\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

What are the fundamentals of the work of Bayes-Price-Laplace?

- 1763: Richard Price publishes posthum the work of Thomas Bayes
- 1781: Pierre-Simon Laplace: Probability theory is nothing, but common sense reduced to calculation ...
- 1812: Théorie Analytique des Probabilités, now known as Bayes’ Theorem, should be correctly named as Bayes-Price-Laplace T.

Who are our giants in the foundations of statistical machine learning?

- Newton, Leibniz, ... developed calculus – mathematical language for describing and dealing with rates of change
- Bayes, Laplace, ... developed probability theory - the mathematical language for describing and dealing with uncertainty
- Gauss generalized those ideas

How does probabilistic (“Bayesian”) machine learning work?

Bayes’ Rule is a corollary of the Sum Rule and Product Rule:

\[ P(x_i | y_j) = \sum_{y_i} P(y_i | x_j) P(x_j) \]  
\[ p(x_i | y_j) = \sum_{y_i} p(y_i | x_j) p(x_i) \]  
\[ P(h|d) \propto p(d|h) * p(h) \]  
\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]
How does inference work in practice?

\[ d \text{ ... data } \quad H \leftarrow \{H_1, H_2, \ldots, H_n\} \quad \forall h, d \rightarrow \]

\[ p(h \mid d) = \frac{p(d \mid h) \cdot p(h)}{\sum_{h' \in H} p(d \mid h') \cdot p(h')} \]

Posterior Probability

Prior Probability

Problem in \( \mathbb{R}^n \rightarrow \) complex

How can we use this in medicine to make predictions?

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

\[ p(D \mid \theta) \]

\[ p(\theta \mid D) = \frac{p(D \mid \theta) \cdot p(\theta)}{p(D)} \]

\[ \text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \]

The inverse probability allows to learn from data, infer unknowns, and make predictions

What is the Probability Density Function vs. Probability Distribution?

How do we reason under uncertainty for decision making in medicine?

- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and predict into the future, e.g. how likely will the patient be ...
- Prior = belief before making a particular observation
- Posterior = belief after making the observation and is the prior for the next observation – intrinsically incremental

\[ p(x_j \mid y_j) = \frac{p(y_j \mid x_j) \cdot p(x_j)}{\sum_{p(x_j \mid y_j) \cdot p(x_j)} p(x_j)} \]
Expectation and Expected Utility Theory

What is the result of the Expected Utility Theory $E(U|d)$?

For a single decision variable an agent can select $D = d$ for any $d \in \text{dom}(D)$. The expected utility of decision $D = d$ is

$$E(U|d) = \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n | d) U(x_1, \ldots, x_n, d)$$

An optimal single decision is the decision $D = d_{\text{max}}$ whose expected utility is maximal:

$$d_{\text{max}} = \arg \max_{d \in \text{dom}(D)} E(U|d)$$

Why does uncertainty matter?

Example: Learning to complement humans

What was the origin of probabilistic decision making?


How does this work in medical decision making?

Two doctors, with equally good training, looking at the same CT scan, will have the same information ... but they may have a different bias/criteria!

What does a correct rejection mean?

Remember: Two doctors, with equally good training, looking at the same CT scan data, will have the same information ... but they may gain different knowledge due to bias/criteria.

Positive = identified and negative = rejected
True positive = correctly identified (hit)
False positive = incorrectly identified, false alarm, type I error
True negative = correctly rejected (correct rejection)
False negative = incorrectly rejected, miss, type II error


And please look up the Wikipedia page:

Why do we need specificity and sensitivity?

Andrew P. Bradley 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30, (6), 1145-1159. doi:10.1016/S0031-3203(96)00145-2
What is the advantage of the concordant pAUC?

Optimal performance is critical for decision-making.

- common performance measures may be too general or too specific.
- AUC too general because including unrealistic decision thresholds.
- Accuracy, sensitivity or the F1 score are measures at a single threshold that reflect an individual single probability or predicted risk, rather than a range of individuals or risk.
- Deep ROC examines groups of probabilities or predicted risks for more insightful analysis.
- that can improve model selection in some cases and provide interpretation and assurance for patients in each risk group.


What is the principle of Bayesian diagnostic in medicine?


How does Bayesian inference work?

- Your MD has bad news and good news for you.
- Bad news first: You are tested positive for a serious disease, and the test is 99% accurate if you are infected (T)
- Good news: It is a rare disease, striking 1 in 10,000 (D)

**How worried would you now be?**

\[
p(x) = \frac{\text{likelihood} \times \text{prior } p(x)}{\text{evidence}}
\]

\[
p(h|d) = \frac{p(d|h)p(h)}{p(d)}
\]

\[
p(T = 1|D = 1) = p(d|h) = 0.99 \quad \text{and}
\]

\[
p(D = 1) = p(h) = 0.0001
\]

\[
p(D = 1 | T = 1) = \frac{(0.99) 	imes (0.0001)}{(1-0.99) \times (1-0.0001) + 0.99 \times 0.0001} = 0.0098
\]

---

**What is the Connection to Cognitive Science/Decision Making?**


---

**Can we scale into the high dimensions?**

What is a Gaussian process?


How can we reduce uncertainty?


Why do we take the human-out-of-the-loop?


(4) aML
Big Data is necessary for aML


When does deep learning fail?


When does aML fail ...

- 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, ...

https://human-center-ed.ai/project/aML
What is now our big problem with deep learning approaches?

- High dimensionality (curse of dim., many factors contribute)
- Complexity of medical problems (medical world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Sensitive to small disturbances (noise, bias, one-pixel attacks, ...)
- 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
- However, most of all ...

05 iML

- iML := algorithms which interact with agents*) and can optimize their learning behaviour through this interaction
  * where the agents can be human
Humans can understand abstract concepts!

What is Deductive vs Inductive vs Abductive Reasoning?

- **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, conclusion: 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 ...
  
  - 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.

What is 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.

How can humans learn so much from so little?

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


What can we do with rare examples in typical classification tasks?

What is the difference between Correlation and Causation?


What is adversarial machine learning?


Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Abstract

Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes, such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.
Correlation ≠ Causality
Why we need the Human-in-the-loop

- Hans Reichenbach (1891-1953):
  Common Cause Principle
  Links causality with probability:
  - If A and B are statistically dependent, there is a C influencing both
  - Whereas:
    A, B, C ... events
    p ... probability density

Hans Reichenbach 1956. The direction of time (edited by Maria Reichenbach), Mineola, New York, Dover.
(6) explainable AI and Methods of Explainability

- **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 by algorithms conform to ethical standards, avoiding bias and misconceptions ..


**What are interpretable models vs. interpreting models?**

- **Interpretable Models, = ante-hoc** - the "glass-box" model itself is ante-hoc interpretable, e.g. Regression, Naïve Bayes, Decision Trees, Graphs, ...

- **Interpreting Black-Box Models, = post-hoc** - the model is not interpretable and needs a post-hoc interpretability method.

---

**What are typical examples of post-hoc methods of explainable AI?**

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Simple Taylor expansions
- 5) Excitation Backpropagation
- 6) Optimization (LIME, BETA, Smooth Grad, ...) BETA transparent approximation, ...
- 7) Deconvolution (Occlusion-based, meaningful perturbations, ...)
- 8) Qualitative Testing with Concept Activation Vectors TCAV

---

**Why is explainable AI only a first step in human-centered AI?**


(7) Causability measures the quality of explanations obtained from (6).

Example: How do human pathologists make diagnoses?

What is ground truth? Where is the ground truth?

- := information truth 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 conclusions about a causal connection based on the conditions of the occurrence of an effect
- Causal machine learning is key to ethical AI in health to model explainability for bias avoidance and algorithmic fairness for decision making
Definition 1: A Kandinsky Figure is ...

- a square image containing 1 to n geometric objects.
- Each object is characterized by its shape, color, size and position within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.

Hubel & Wiesel (1962): Our world is compositional!


Source: https://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/models-of-information-processing-in-the-visual-cortex

Related Work (1): Bongard Problems

Bongard, M. Mikhail, 1967. The problem of recognition (in Russian), Moscow, Nauka (1970 in English)

Related Work (2): CLEVR

https://cs.stanford.edu/people/jjohns/clevr/

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick & Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 Hawaii. IEEE.

Related Work (3): CLEVRER


Related Work (4): Bongard-LOGO


Related Work (5): CURI

Definition 2 A statement $s(k)$

- about a Kandinsky Figure $k$ is ...
- either a mathematical function $s(k) \rightarrow B$; with $B (0,1)$
- or a natural language statement which is true or false
  - The evaluation of a natural language statement is always done in a specific context.
  - we follow well known concepts from human perception and linguistic theory.
  - If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.


Intelligence Test for Machines


Difficult Concepts

- A) True (the cells are smaller and closer together – it is a tumor ...)
- B) False
- C) Counterfactual (What if the cells are slightly bigger ?)
Measuring the quality of Explanations: The Systems Causability Scale


Definitionen: Explainability vs. Causability

- Causability is neither a typo nor a synonym for Causality
- Causa-bil-ity ... in reference to ... Us-a-bil-ity.
- While xAI is about implementing transparency and traceability, Causability is about the measurement of the quality of explanations.
- **Explainability**: technically highlights decision relevant parts of machine representations and machine models i.e., parts which contributed to model accuracy in training, or to a specific prediction.
  - Explainability does not refer to a human model!
- **Causability**: the measurable extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency, satisfaction in a specified context of use.
  - Causability does refer to a human model!


Conclusio

Thank you very much!

Explainability needs a framework to ensure common understanding and adaptive Question/Answering Interfaces

Appendix

Thank you very much!
“The most interesting facts are those which can be used several times, those which have a chance of recurring...
which, then, are the facts that have a chance of recurring?
In the first place, simple facts.”