Machine Learning & Knowledge Extraction (MAKE) for Health Informatics: Towards educating a new kind of graduates

Andreas Holzinger
Holzinger Group HCI-KDD, Institute for Medical Informatics/Statistics
Medical University Graz, Austria
andreas.holzinger@medunigraz.at
Graz, April, 2018

Abstract

Machine learning (ML) is the fastest growing field in computer science, and health informatics is among the greatest application challenges, providing huge benefits in intelligent decision support systems and biomedical data analysis towards personalized medicine. The enormous scientific, industrial and economic challenges involving privacy, data protection, safety and security require a new kind of graduates. Machine learning studies algorithms which can learn from data to extract knowledge from experience to make decisions and predictions. Health informatics studies the effective use of probabilistic information for decision making under uncertainty; to augment human intelligence with artificial intelligence, students need a cross-disciplinary skill set. My teaching in the last years revolved around a combination of machine learning and knowledge extraction (MAKE) with health informatics towards explainable AI. This is important, because new European privacy regulations make it necessary in the future to explain *why* a decision has been made. To support broad applicability of methods, algorithms and tools in industry, I foster within my teaching probabilistic programming with Python.

1 Introduction and Motivation for Teaching

A great privilege of my academic position is to work with students. Teaching and explaining is an essential part of the research process. Consequently, in my teaching I follow a Research-Based Teaching (RBT) approach [1], providing inspiration and excitement to my students. This is relatively easy to do in AI/ML, due to the fact that it is currently the most exciting area of computer science with many future aspects and opportunities for science, engineering and business, whilst health informatics is generally accepted as the greatest application challenge. In ML jargon, I regard teaching as a practical application of probabilistic multi-objective optimization, with central but competitive principles aiming at optimizing each objective adaptively to reach a trade-off in each direction. This is important in ML, as worldwide our data driven health industries need a new kind of education to provide future professionals with the necessary cross-domain skill set to solve challenging future problems. The application domain health is of importance, as health systems worldwide are challenged by increasingly heterogeneous, high-dimensional data and increasing amounts of unstructured information. Cognitive complexity and high-level visualizations challenge the appropriate understanding of information in the application *context*. The tailoring of information representations to the specificity of human information processing is crucial, as in many domains we are facing an enormous diversity of end users, e.g. medical doctors have to understand complex information for decision making.

2 Successful ML for Health Informatics needs a concerted effort

ML is a field at the intersection of cognitive science and computer science [2], and progressed enormously in the last two decades with huge application challenges and business potential [3], [4]. To see health informatics among the greatest challenges is not surprising, because here we are confronted with uncertainty, with probabilistic, unknown, incomplete, heterogenous, noisy, dirty, erroneous, inaccurate, missing, yet even contradictory data sets in arbitrarily high dimensional spaces [5], [6]. ML is an extremely broad field and successful application of ML requires a concerted cross-domain effort - following the HCI-KDD approach [7] [8] (see figure 1). The central goal is in bringing ML-pipelines directly into the work-flows of the end-users [9] and to support explainable AI [10].

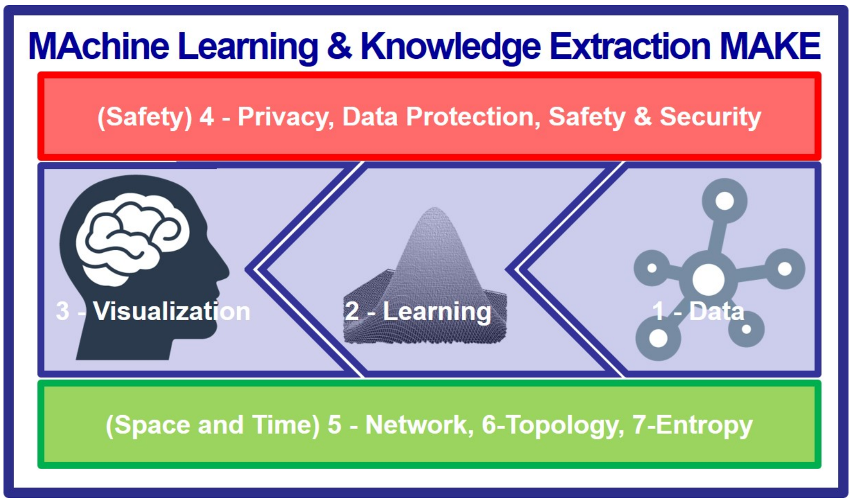

Figure 1: The big picture of the HCI-KDD approach: The horizontal process chain (blue box) encompasses the whole machine learning pipeline from physical aspects of raw data, to human aspects of data visualization; while the vertical topics (green box) include important aspects of structure (graphs/networks), space (computational topology) and time (entropy); privacy, data protection, safety and security are mandatory topics within the health domain and provide kind of a base compartment

3 My Teaching experience in MAKE

My teaching experience in MAKE with the application in the health informatics domain[1] includes courses at undergraduate, graduate and postgraduate levels at various institutions and schools, and tutorials at international conferences and workshops. Some recent courses include:

MAKE-Health (2 ECTS) at the University of Verona, where I helped to establish health informatics among the computer science faculty.

LV 185.A83 Machine Learning for Health Informatics (3 ECTS) This is a graduate course at Vienna University of Technology since 2016 - see it as sample in the next section

LV 706.046 AK HCI - Intelligent User Interfaces - HCI meets AI (5 ECTS) This is a practical graduate course at Graz University of Technology since 2003

LV 706.315 Selected Topics on interactive Knowledge Discovery (3 ECTS)

LV 709.049 Biomedical Informatics: discovering knowledge in data (3 ECTS)

Moreover I regularly offer the following seminars: LV 706.996 and 706.998 Seminar for Master's Students; LV 706.997 and 706.999 Seminar for PhD Students; LV 706.119 Project Information Systems; LV 706.116 Master's Project Software Development; LV 709.036 Biomedical Seminar; LV 706.502 Master's Project Web and Data Science to mention only the most relevant ones.

[1] Watch a sample video: https://youtu.be/lc2hvuh0FwQ

4 Example Curriculum: MAKE for Health Informatics

The Master's course 183.A83 [2] [11] is a modular system consisting of 1 primer, 1 introduction, 12 modules and 6 practicals, adaptable to the prior knowledge of the students (figure 2).

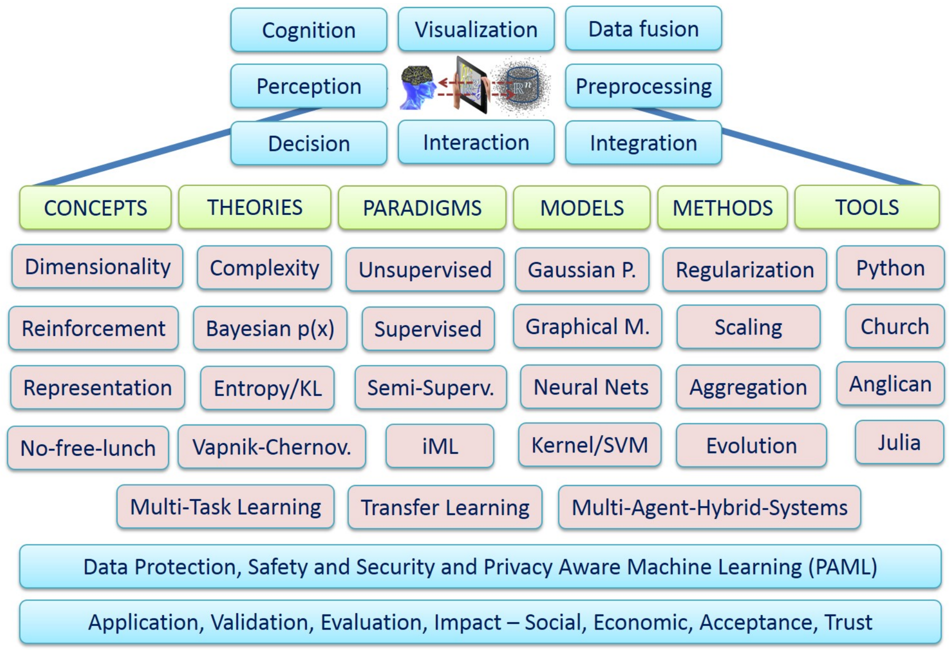

Figure 2: Topical orientation of the course 183.A83 Machine Learning for Health Informatics

Throughout this course I foster probabilistic programming with Python [12], [13].

Primer on Probability, Information and Learning from Data includes the necessary mathematical skill set to understand the course and brings the students to an equal level; students which have attended probability classes may skip this module.

Module 00: Introduction and Overview of ML and HI explains the HCI-KDD approach, shows the complexity of the application area health informatics, demonstrates what automatic ML can do, demonstrates the limitations of aML, and the usefulness of interactive ML with a human-in-the-loop on practical examples and outlines some future challenges in health informatics.

Module 01: Fundamentals of Data and Information discusses specifics of biomedical data, data integration in the life sciences, introduction to probabilistic information with a focus on the problem of estimating the parameters of a Gaussian distribution (maximum likelihood problem) and shows the importance of the Kullback–Leibler divergence which is very important, particularly for sparse variational methods between stochastic processes [14].

Tutorial 01: **Data augmentation** discusses the artificial generation of new data through the expansion of an existing data set by introducing new samples created by perturbation of original samples. This is, for example, required for training neural networks that have many millions of learning parameters and thousands of categories. In deep learning it is often used in scenarios where only low numbers of samples are available or where severe class imbalance is present [15].

Module 02: Probabilistic Graphical Models Part I is a primer for the second tutorial on probabilistic programming, with Monte Carlo sampling from probability distributions based on MCMC, which is very important and awesome, as it is similar as our brain may work and allows for computing hierarchical models having a large number of unknown parameters and also works well for rare event sampling which is often the case in the health informatics domain.

[2] http://hci-kdd.org/machine-learning-for-health-informatics-course

Tutorial 02 Probabilistic Programming with Python is playing with the Python framework PyMC3[16], which allows automatic Bayesian inference on user-defined probabilistic models. MCMC sampling allows inference on increasingly complex models. This class of MCMC, known as Hamiltonian Monte Carlo, requires gradient information which is often not readily available.

Module 03: Probabilistic Graphical Models Part II continues with graphical model structure learning for knowledge discovery, learning tree structures and directed acyclic graphs (DAG), learning causal DAGs, and undirected Gaussian graphical models and gives an outline of graph bandits.

Module 04: Human Learning vs. Machine Learning: Decision Making starts with reinforcement learning and discusses the differences of humans and machines on the example of decision making under uncertainty, shows then multi-armed bandits and applications in health informatics and finally gives an outlook on the importance of transfer learning.

Module 05: Dimensionality Reduction and Subspace Clustering provides an introduction into classification vs. clustering, feature spaces, feature engineering, discusses the curse of dimensionality and methods of dimensionality reduction, and demonstrates the usefulness of subspace clustering with the expert-in-the-loop; discusses the question "what is relevant?" on projection pursuit.

Module 06: Machine Learning from Text focuses on natural language understanding and the problems involved, and highlights word vectors for sentiment analysis (continuous bag-of-words model, skip-gram model, global vectors for word embedding) with giving an outline on neural probabilistic language models and alternative models.

Tutorial 03: Machine Learning from Text focuses on natural language understanding and the problems involved, and highlights word vectors for sentiment analysis (continuous bag-of-words model, skip-gram model, global vectors for word embedding) with giving an outline on neural probabilistic language models and alternative models.

Module 07: Evolutionary Computing for HI I poses medical decision making as search problem and shows evolutionary principles (Lamarck, Darwin, Baldwin, Mendel) and applications of evolutionary computing with the special case of genetic algorithms and k-armed bandits and genetic algorithms (global optimization problem).

Module 08: Evolutionary Computing for HI II continues with examples from medical applications for EA, discusses natural computing concepts and their usefulness for solving health problems, focuses then on Ant Colony Optimization and the traveling salesman problem with motivation on protein folding, simulated annealing, and the human-in-the-loop, and finalizes with multi-agents and neuro-evolution.

Module 09: Towards Open Data Sets: Privacy Aware Machine Learning motivates privacy, data protection safety and security and discusses anonymization methods (k-Anonymization, l-diversity, t-closeness, delta-presence, pertubative approaches, differentially private kernel learning, etc.), and how iML can help anonymization.

Tutorial 04: Privacy-Aware Machine Learning and Secure Federated Learning asks questions including 1) how do multi-class classification, prediction, etc., behave under perturbation, 2) is ML on graph structures more robust under the effects of perturbation, and 3) can iML with a Human-in-the loop yield more robust heuristics for cost functions so that information loss in anonymization can be minimized.

Module 10: Active Learning and Active Preference Learning discusses some principles of active learning, preference learning, active preference learning with an excursion on PAC-learning, and programming by feedback, highlights some problems of the human-in-the-loop and shows some examples where humans are better than machines.

Module 11: Multi-Task Learning and Transfer Learning discusses the grand challenges of artificial intelligence of the future which are in answering the question: "How can we perform a task by exploiting knowledge, extracted during solving previous tasks?" and to help to overcome the problem of catastrophic forgetting.

Module 12: Discrete Multi-Agent Systems on the topic of stochastic simulation of tumor kinetics and key problems for cancer research, tumor growth modeling, cellular Potts model, tumor growth visualization and towards using open tumor growth data for machine learning in the international context.

Tutorial 05: Discrete Multi-Agent Systems based on module 12 aspects of stochastic simulation of tumor kinetics and key problems for cancer research, tumor growth modeling, cellular potts model and tumor growth visualization will be practiced .

Tutorial 06: Experimenting, Evaluating and Benchmarking of learning methods, models, algorithms and tools is key for success in health informatics and includes ROC Analysis and AUC, probabilistic and qualitative measures, and metrics to determine the accuracy of error rates.

5 Future Teaching Plans

My goal is to consolidate this curriculum and expanding it towards explainable AI. I like to work with students, to teach and to mentor, and actually this is my main motivation for staying in academia. I am convinced that teaching is inherently connected with the development of critical thought and reasoning thus the cornerstone of successful research. I am looking forward to continue my role as internationally orientated catalytic research based teacher. One fact is, that women are underrepresented in computer science and mathematics generally, and machine learning specifically, so one goal for the near future is to help to motivate female students to get into this exciting field.

References

[1] Andreas Holzinger. *Successful Management of Research and Development*. BoD, Norderstedt, 2011.

[2] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

[3] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349 (6245):255–260, 2015.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[5] Andreas Holzinger, Matthias Dehmer, and Igor Jurisica. Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. *BMC Bioinformatics*, 15(S6): I1, 2014.

[6] Sangkyun Lee and Andreas Holzinger. Knowledge discovery from complex high dimensional data. In *Lecture Notes in Artificial Intelligence LNAI 9580*, pages 148–167. Springer, Cham, 2016.

[7] Andreas Holzinger. Human–computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together? In *Lecture Notes in Computer Science LNCS 8127*, pages 319–328. Springer, Heidelberg, Berlin, New York, 2013.

[8] Andreas Holzinger. Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. *IEEE Intelligent Informatics Bulletin*, 15(1):6–14, 2014.

[9] Andreas Holzinger and Igor Jurisica. Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Lecture Notes in Computer Science LNCS 8401*, pages 1–18. Springer, Heidelberg, Berlin, 2014.

[10] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable ai systems for the medical domain? *arXiv:1712.09923*, 2017.

[11] Andreas Holzinger. *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges, Lecture Notes in Artificial Intelligence LNAI 9605*. Springer International, Cham, 2016.

[12] Marcus D. Bloice and Andreas Holzinger. A tutorial on machine learning and data science tools with python. In Andreas Holzinger, editor, *Machine Learning for Health Informatics, Lecture Notes in Artificial Intelligence LNAI 9605*, pages 437–483. Springer, Heidelberg, 2016.

[13] Andrew D Gordon, Thomas A Henzinger, Aditya V Nori, and Sriram K Rajamani. Probabilistic programming. In *Proceedings of the on Future of Software Engineering*, pages 167–181. ACM, 2014.

[14] Alexander Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 41, pages 231–239. JMLR, 2016.

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In Zhoubin Ghahramani, Max Welling, C. Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 766–774. Curran, 2014.

[16] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.