Andreas Holzinger
706.315 Selected Topics on Knowledge Discovery:
Interactive Machine Learning
2015W, SE, 2.0 h, 3.0 ECTS
Week 42 - 16.10.2015  10:00-11:30

Introduction to Machine Learning (ML):
automatic ML - interactive ML

a.holzinger@hci-kdd.org
01 The HCI-KDD approach
02 What is Machine Learning?
03 Application Area: Health Informatics
04 Probabilistic Information
05 Gaussian Processes
06 Automatic Machine Learning (aML)
07 Interactive Machine Learning (iML)
08 Conclusion and Future Outlook
09 Questions
10 Appendix
Quiz: Which of the following tasks include ML

1. All Mail

2. Statistical analysis

3. Financial market analysis

4. Online shopping

5. Social media monitoring

6. Fraud detection

7. DNA sequence analysis

8. Protein structure analysis

9. Image classification

10. Weather forecasting

11. Text classification

12. Network security analysis

Solutions in the Appendix
What is the HCI-KDD approach

01 HCI-KDD
In the Knowledge Discovery pipeline ML is the heart ...

Interactive

Data Mining

Knowledge Discovery

1. Data Mapping
2. Learning Algorithms
3. Graph-based Data Mining
4. Topological Data Mining
5. Entropy-based Data Mining
6. Data Visualization
7. Privacy, Data Protection, Safety and Security

Successful ML needs a concerted effort...

concerted effort
international
without boundaries...
Grand challenge: Transfer results from $\mathbb{R}^n$ to $\mathbb{R}^2$

Human intelligence (Cognitive Science)

Machine intelligence (Computer Science)

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)
Science is to test crazy ideas – Engineering is to put these ideas into Business

“solve intelligence – then solve everything else”

Demis Hassabis, 22 May 2015

The Royal Society, Future Directions of Machine Learning Part 2

https://youtu.be/XAbLn66iHcQ?t=1h28m54s
Cognitive Science vs. Computer Science

- **Cognitive Science → human intelligence**
  - Study the principles of *human learning* to understand biological intelligence

- **Human-Computer Interaction → the bridge**
  - Interacting with algorithms that learn shall enhance user friendliness and let concentrate on problem solving - Opening the “black-box” to a “glass-box”

- **Computer Science → computational intelligence**
  - Study the principles of *machine learning* to understand artificial intelligence
CS aims to reverse engineer **human intelligence**;
ML provides powerful sources of insight into *how machine intelligence* is possible.
CS therefore raises challenges for, and draws inspiration from ML;
Insights about the human mind may help inspire **new directions for ML**...
Is the human brain a inference engine?

- Learning concepts from examples (babies!)
- Causal inference and reasoning
- Predicting everyday events
- Even little children solve complex problems unconsciously, effortlessly, and ... successfully
- Compare your best Machine Learning algorithm with a seven year old child!


02 What is ML?
What is (machine) learning?

- No hand-crafted pre-programmed solutions – adapting/modifying behavior on experience
- Arthur Samuel (1959) [1]: "Field of study that gives computers the ability to learn without being explicitly programmed"

People are awesome ...

See Youtube: “people are awesome” ... hundreds of examples
Machine Learning and Statistics are closely related

- Machine Learning is the development of algorithms which can **learn from data**
- Machine Learning has a pre-history in **statistical learning**, which is the application of statistical models and the assessment of **uncertainty**

The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
How can ML help to solve a problem task?

Tasks are addressed by models, whereas learning problems are solved by learning algorithms that produce models. Important: Using the right features to build the right models that achieve the right tasks!

Flach, P. 2012. Machine learning: the art and science of algorithms that make sense of data, Cambridge University Press, p.11, Figure 3
When do we need machine learning?

- ML is necessary ...
  - when human expertise does not yet exist or is absent (bioinformatics, biomarkers, ...)
  - when solutions change over time and/or solutions need to adapt to new situations (personalization, ...)
  - when humans are unable (or unwilling) to explain their knowledge
  - When the problem space is so enormous large and data are of very high-dimensions which cannot processed manually
- ML is useful ... for many, many applications!
How do human babies learn?

- “... The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion; and to the very end of life, our location of all things in one space is due to the fact that the original extents or bignesses of all the sensations which came to our notice at once, coalesced together into one and the same space ...”

“An ultra-intelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion*" and the intelligence of man would be left far behind ... It is curious that this point is made so seldom outside of science fiction.”

Irving John Good, Trinity College, Oxford, 1965
Colleague of Alan Turing in Bletchley Park


*) https://intelligence.org/ie-faq/

Deep Learning is a very old concept ...

**Psychological Review**
Vol. 65, No. 6, 1958

**THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN**

F. ROSENBLATT  
*Cornell Aeronautical Laboratory*

If we are eventually to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking, we must first have answers to three fundamental questions:

1. How is information about the physical world sensed, or detected, by the biological system?
2. In what form is information stored, or remembered?
3. How does information contained in storage, or in memory, influence recognition and behavior?

**Fig. 1.** Organization of a perceptron.

2015 everything is machine learning ...


1) Which hypothesis space $\mathcal{H}$ to choose?
2) How to measure the degree of fit?
3) How to trade-off degree of fit vs. complexity?
4) How to find a good hypothesis $h$?
5) How to know that a good $h$ will predict well?

ad 3) Occam's Razor: the most simple shall be chosen
1) Which hypothesis space $\mathcal{H}$ to choose?
   - Deterministic $h’s$, mathematical expressions and/or logical sentences; implicit relevance determination

2) How to measure the degree of fit?
   - $h$ shall be consistent with the underlying data

3) How to trade-off degree of fit vs. complexity?
   - The underlying theory must be totally correct

4) How to find a good hypothesis $h$?
   - Intuition, imagination, inspiration, inventiveness, ...

5) How to know that a good $h$ will predict well?
   - David Hume’s problem of induction: most scientists give up

ad 2) Ernest Rutherford: If your experiment needs statistics, you ought to have done a better experiment
ad 3) excluding the experimental error
ad 4) Thomas Alva Edison: Ingenious is 1 % inspiration and 99% perspiration
1) Which hypothesis space \( \mathcal{H} \) to choose?
   - All Turing Machines or programs of a Universal TM

2) How to measure the degree of fit?
   - Fit must be perfect – program shall output data exactly

3) How to trade-off degree of fit vs. complexity?
   - Minimize the size of the program

4) How to find a good hypothesis \( h \) ?
   - Undecidable ... unless we bound time complexity of \( h \)

5) How to know that a good \( h \) will predict well?
   - Theory from PAC learning

Ad 5) In computational learning theory, Probably Approximately Correct learning (PAC learning) is a framework for mathematical analysis of machine learning, proposed by Leslie Valiant, 1984.
1) Which hypothesis space \( \mathcal{H} \) to choose?
   - e.g., linear combinations of features: \( h_w(x) = w^T x \)

2) How to measure the degree of fit?
   - Loss function, e.g., squared error \( \Sigma_j (y_j - w^T x)^2 \)

3) How to trade-off degree of fit vs. complexity?
   - Regularization: complexity penalty, e.g., \( ||w||^2 \)

4) How to find a good hypothesis \( h \)?
   - Optimization (closed-form, numerical); discrete search

5) How to know that a good \( h \) will predict well?
   - Try it and see (cross-validation, bootstrap, etc.)

1) Which hypothesis space $\mathcal{H}$ to choose?
   - Probability model $P(y | x, h)$, e.g., $Y \sim N(w^T x, \sigma^2)$

2) How to measure the degree of fit?
   - Data likelihood $\prod_j P(y_j | x_j, h)$

3) How to trade-off degree of fit vs. complexity?
   - Regul. or prior: $\arg\max_h P(h) \prod_j P(y_j | x_j, h)$ (MAP)

4) How to find a good hypothesis $h$?
   - Optimization (closed-form, numerical); discrete search

5) How to know that a good $h$ will predict well?
   - Empirical process theory (generalizes Chebyshev, CLT, PAC...); Key assumption is (i)id
1) Which hypothesis space $\mathcal{H}$ to choose?
   - All hypotheses with nonzero a priori probability

2) How to measure the degree of fit?
   - Data probability, as for MLE/MAP

3) How to trade-off degree of fit vs. complexity?
   - Use prior, as for MAP

4) How to find a good hypothesis $h$?
   - Don’t! Bayes predictor ...
     \[ P(y|x,D) = \sum h P(y|x,h) P(D|h) P(h) \]

5) How to know that a good $h$ will predict well?
   - Bayesian prediction is already optimal 😊
• Progress in ML is driven by the explosion in the availability of big data and low-cost computation.

• Health is amongst the biggest challenges

03 Application Area: Health Informatics
Is this our ML lab?
Does all these work here as well?
Why is this application area complex?
In medicine we have two different worlds ...

Our central hypothesis:
Information may bridge this gap

Where is the problem in building this bridge?
What are the (main) problems?

Heterogeneity of Data

Curse of Dimensionality

Complexity

Uncertainty

04 Probabilistic Information $p(x)$
Who is this?
The foundation for machine learning was laid in 1763 ...

Probabilistic Information p(x)


\[ p(x_i) = \sum P(x_i, y_j) \]

**Thomas Bayes 1701 - 1761**

\[ p(x_i, y_j) = p(y_j|x_i)P(x_i) \]

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

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

Learning and Inference

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

- **Likelihood**
- **Prior Probability**
- **Posterior Probability**

\( d \) ... data  \( h \) ... hypothesis  \( H \) ... \( \{H_1, H_2, ..., H_n\} \)  \( \forall h, d \) ...

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

Feature parameter \( \theta \)
Biomedical Example
To find structural anomalies in such data...

Learning from previous examples ...

The more examples we have the better ...

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, ...
Big Data is good for automatic Machine Learning
\[ \mathcal{D} = x_{1:n} = \{ x_1, x_2, \ldots, x_n \} \]

\[ p(\mathcal{D}|\theta) \]

\[ p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \ast p(\theta)}{p(\mathcal{D})} \]

**posterior = likelihood \ast prior \over evidence**

The inverse probability allows to learn from data, infer unknowns, and make predictions.
Let the data do the work!

\[
\max_{x \in \mathcal{A} \subset \mathbb{R}^d} f(x)
\]

\[
p(h|d) \propto p(\mathcal{D}|\theta) \ast p(h)
\]

\[
p(f(x)|\mathcal{D}) \propto p(\mathcal{D}|f(x)) \ast p(f(x))
\]

- Machine Learning is the development of algorithms which can **learn from data**
- assessment of **uncertainty**, making **predictions**
- Automating automation - getting computers to program themselves – let the data do the work!
- 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
Who is this?
05 Gaussian Processes
From Bayesian Optimization to Gaussian Process (GP) approximation

$p(f(x)|\mathcal{D}) \propto p(\mathcal{D}|f(x)) p(f(x))$
Demo on how Bayesian Optimization works ...

Why is this relevant for health informatics?
- 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

\[
p(x_i | y_j) = \frac{p(y_j | x_i) p(x_i)}{\sum p(x_i, y_j)p(x_i)}
\]
Grand Goal of automatic Machine Learning
06 aML
Today most ML-applications are using automatic Machine Learning (aML) approaches.

Automatic Machine Learning (aML) := algorithms which interact with agents and can optimize their learning behaviour through this interaction.
Best practice examples of aML
Lenovo Einsteiger Notebooks Arbeitsspeicher und \nArtikel von Lenovo
von Lenovo
70 Kunden
Bestseller Nr. 1 in Notebooks

Unverb. Preisempf.: EUR 349,99
Preis: EUR 273
Sie sparen: EUR 76,01
Alle Preise für Details.

Auf Lager.
Verkauf und Versand durch Amazon
20 neu ab EUR 273,99 4 gebraucht

Größe: 500GB
1TB 800GB

Stil Intel Pentium
Intel Core i3 Intel Pentium

- Prozessor: Intel Pentium N35
- Besonderheiten: HD Glare Display
- Akku: bis zu 4 Stunden Akkulaufzeit
- Herstellergarantie: 12 Monate
- Lieferumfang: Lenovo ideapad

Für größere Ansicht Maus über das Bild ziehen

Wenger 600638 IBEX 17" Laptop Backpack with Tablet / eReader Pocket (Black / Blue)
n von Wenger

EUR 66,99 Prime

Andere Angebote
EUR 60,00 neu (22 Angebote)
EUR 57,70 gebraucht (1 Angebot)

Nur Artikel von Wenger anzeigen

🌟🌟🌟🌟🌟 295

Holzinger Group
Fully automatic autonomous vehicles ("Google car")

Cyber-Physical Systems (CPS):
Tight integration of networked computation with physical systems

- Automotive
  - E-Corner, Siemens
- Building Systems
- Telecommunications
- Transportation (Air traffic control at SFO)
- Avionics
- Instrumentation (Soleil Synchrotron)
- Factory automation
  - Courtesy of Kuka Robotics Corp.
  - Courtesy of General Electric
  - Courtesy of Doug Schmidt

Big Data is necessary for aML!

Again: Does this all work here as well?
Medical Decision Making as a Search Task in $\mathcal{H}$

Problem: Time (t)
Search in an arbitrarily high-dimensional space < 5 min.
What is a good decision?

- Maximizing Expected Utility Theory
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,
  - Protein-Folding,
  - k-Anonymization,
  - Graph Coloring, Category Discovery, etc. etc....
Example: Discovery of causal relationships from data ...


https://www.youtube.com/watch?v=9KiVNIUMmCc
Sometimes we (still) need a human-in-the-loop
interactive Machine Learning (iML) := algorithms which interact with agents*) and can optimize their learning behaviour through this interaction

*) where the agents can be human

A group of experts-in-the-loop
A crowd of people-in-the-loop
Nature-inspired-agents-in-the-loop
A) Unsupervised ML: Algorithm is applied on the raw data and learns fully automatic – Human can check results at the end of the ML-pipeline

B) Supervised ML: Humans are providing the labels for the training data and/or select features to feed the algorithm to learn – the more samples the better – Human can check results at the end of the ML-pipeline

C) Semi-Supervised Machine Learning: A mixture of A and B – mixing labeled and unlabeled data so that the algorithm can find labels according to a similarity measure to one of the given groups
D) 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 ...

Constraints of humans: Robustness, subjectivity, transfer?
Open Questions: Evaluation, replicability, ...

Three examples for the usefulness of the iML approach

- Example 1: k-Anonymity
- Example 2: Protein Folding
- Example 3: Subspace Clustering
Example 1: k-Anonymization of Medical Data

87% of the population in the USA can be uniquely re-identified by Zip-Code, Gender and date of birth

Proteins are the building blocks of life ...

Example 2 Protein Folding: ὀμολογέω (homologeo)

The sequence of a protein can NOT (yet) be used to predict its 3D structure ...

Humans can help here: Crowdsourcing

Example 3: Subspace Clustering

- Patterns may be found in subspaces (dimension combinations)
- Clustering and subset selection: Non-convex & NP-hard
- Real data are often noisy and corrupted
- Little prior knowledge about low-dim structures
- Data points in different subgroups can be very close

What did the “doctor-in-the-loop” do?

- **Positive** subspace clusters
  - One homogeneous cluster (healthy patients)
    - hyper, CVD, neoplasm, psy.disorder, drug allergy
    - No medications: statins, anticoagulants, analgesics and clear (preserved renal function)

- **Negative** subspace clusters
  - Cluster with obvious reasons for neg. outcome
    - Impairment of certain pathophysiologic mechanism increased MCV, decreased VITB12, FOLNA, CORTIS despite no: DM, drug allergy, Fglu, E/HB (anaemia)
The usefulness of a Human kernel...

Judgment 1 out of 33

This is the first function from the system. Please try to predict the new points as well as you can based on the points you can see.

Please click along the blue line to say what you think the height of the point is for that location.

Once you have selected a position along the line, **hit the 's' key to submit the point.**


\[
\hat{p}^a_{bc} = \frac{\mu + \delta_{ac}}{2\mu + \delta_{ab} + \delta_{ac}} \quad \text{and} \quad K_{ii} = 1,
\]
Experimental Proof for the Human in the Loop

Algorithm 2: Ant Colony Algorithm iML

Input: ProblemSize, PopulationSize, m, ρ, β, σ, q0
Output: P_{best}

\begin{align*}
P_{best} & \leftarrow \text{CreateHeuristicSolution(ProblemSize)}; \\
P_{best}\text{cost} & \leftarrow \text{Cost}(S_{n}); \\
\text{Pheromone}_{\text{init}} & \leftarrow \frac{1}{\text{ProblemSize} \times P_{best}\text{cost}}; \\
\text{Pheromone} & \leftarrow \text{InitializePheromone(\text{Pheromone}_{\text{init}})}; \\
\text{while } -\text{StopCondition()} & \text{ do} \\
\quad \text{for } i = 1 \text{ to } m & \text{ do} \\
\quad \quad S_i & \leftarrow \text{ConstructSolution(\text{Pheromone}, \text{ProblemSize}, \beta, \sigma)}; \\
\quad \quad S_{i}\text{cost} & \leftarrow \text{Cost}(S_i); \\
\quad \quad \text{if } S_{i}\text{cost} \leq P_{best}\text{cost} & \text{ then} \\
\quad \quad \quad P_{best}\text{cost} & \leftarrow S_{i}\text{cost}; \\
\quad \quad \quad P_{best} & \leftarrow S_i; \\
\quad \quad \text{end} \\
\quad \text{end} \\
\quad \text{LocalUpdateAndDecayPheromone(\text{Pheromone}, S_i, S_{i}\text{cost}, \sigma)}; \\
\text{end} \\
\text{GlobalUpdateAndDecayPheromone(\text{Pheromone}, P_{best}, P_{best}\text{cost}, \rho)}; \\
\text{while isUserInteraction()} & \text{ do} \\
\quad \text{GlobalAddAndRemovePheromone(\text{Pheromone}, P_{best}, P_{best}\text{cost}, \rho)}; \\
\text{end} \\
\text{end} \\
\text{return } P_{best}; \\
\end{align*}
Please take part in this online Experiment:
http://hci-kdd.org/projects/iml-experiment
08 Conclusion
Coming to the conclusion ...

1. Heterogeneous data sources
   - need for data integration and data fusion
2. Complexity – reduction of search space
   - combining the best of Human & Computer
3. What is interesting? – and relevant!
   - need of effective mapping $\mathbb{R}^N \rightarrow \mathbb{R}^2$
4. Clinical time limits “5 Minutes”
   - need of efficient solutions

Three Main future challenges

**Multi-Task Learning (MTL)**

for improving prediction performance, help to reduce catastrophic forgetting

**Transfer learning (TL)**

is not easy: learning to perform a task by exploiting knowledge acquired when solving previous tasks:

a solution to this problem would have major impact to AI research generally and ML specifically.

**Multi-Agent-Hybrid Systems (MAHS)**

To include swarm-intelligence and crowdsourcing and making use of discrete models – avoiding to seek perfect solutions – better have a good solution < 5 min.
concerted effort
international
without boundaries ...
ML-Algorithms are key but needs also concerted effort

Thank you!
09 Questions
Sample Questions

- What is Active Learning?
- Where are the advantages of AL?
- Describe a few scenarios for AL?
- How does the robot scientist by King et al (2004) work?
- What does “Probable Approximate Correct” mean?
- What is the basic assumption of PL?
- What is the core essence of the “programming by feedback” approach?
- What could be huge disadvantages with the “human-in-the-loop”?
- What is a utility function?
- Why is multi-task learning of extreme importance for future research?
- When are humans better in TL?
- Explain the 3 types of TL and the 4 TL approaches!
- What is the main idea of inductive TL?
10 Appendix
Recommended Books 1 (General Introductions)

http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html

https://www.cs.bris.ac.uk/~flach/mlbook/


https://www.microsoft.com/en-us/research/people/cmbishop/

http://www.cmpe.boun.edu.tr/~ethem/i2ml3e/
Solutions to the Quiz Questions

• ad 1) yes; e.g. Naïve Bayes classifier for spam learning; K-nearest neighbor with training data; collaborative filtering, clustering, etc. etc.
• ad 2) yes; e.g. SVD/PCA, ensemble methods, regression, clustering etc.;
• ad 3) yes; e.g. SVM, reinforcement learning, regression etc.;
• ad 4) yes; e.g. Google-Page-Rank Algorithm;
• ad 5) yes; reverse image search, e.g. feature detection algorithms for matching deformations (SIFT, PCA-SIFT and SURF) etc. Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images
• ad 6) yes, recommender systems, collaborative filtering.
• ad 7) yes, fraud detection, e.g. profiling methods
• ad 8) yes, sequence motif, pattern of nucleotides in a DNA sequence or amino acids in a protein; structural motif, a pattern in a protein structure formed by the spatial arrangement of amino acids;
• ad 9) yes, structural alignment, establishing homology between polymer structures based on their shape and 3D-conformation. (protein tertiary structures);
• ad 10) yes, learning from DNA data
• Ad 11) yes, feature extraction from cancerous pathological samples
• Ad 12) yes, machine learning approach to predict protein–ligand binding affinity with applications to molecular docking in Bioinformatics
Popular ML research data sets

- University California at Irvine Machine Learning Repository: [http://archive.ics.uci.edu/ml/](http://archive.ics.uci.edu/ml/) - some examples include:
  - Wisconsin Breast Cancer data set: Given the results of a diagnostic test on breast tissue, predict whether the mass is a tumor or not
  - Heart Disease data set: Given the results of various diagnostic tests on a patient, predict the risk of heart disease of the patient.
  - Human activity recognition data set: From smart phone movement data predict the type of activity performed by the person holding the smart phone.
- More information see:
The doctor of the future will give no medicine, but will instruct his patient in the care of
the human frame, in diet and in the cause and prevention of disease.

Thomas Alva Edison (1847 – 1931)