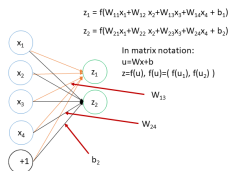


Chapter 1: Introduction to Machine Learning – Overview



Christian Heumann

(largely based on slides from Nina Poerner, Benjamin Roth, Marina Speranskaya)

CIS LMU München, Department of Statistics LMU München

November 2020

Course mode

- 60 - 70 minutes of video material per chapter (split into smaller videos of approx. 15 - 20 minutes)
- 30 minutes discussion of the videos during the regular lecture time
- 45 minutes (live) discussion of the exercise sheet's solution afterwards
- Examination: 15 minutes Oral Exam (Date: 17.02.2021)

References:

- Deep learning: ▶ Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep learning*
- Statistical learning:
▶ Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*
- Introduction to Machine Learning (online course):
▶ Bernd Bischl, Fabian Scheipl, Heidi Seibold, Christoph Molnar and Daniel Schalk. *I2ML*

Why Machine Learning is Useful Today

- Humans get lost in the lake of ever-growing volumes and varieties of available data **► Big Data**, thus automatic methods for analyzing the data or making decisions based on the data are necessary.
- Unstructured data (e.g. text-heavy data) **► Unstructured Data** often needs a lot of preprocessing before it can be further used. For humans this is usually an awkward task and is better delegated to a "machine".
- Scientific laws (e.g. laws of classical mechanics) can most often not be applied to the data at hand, we can only learn something (usually called a *model*) from the observed *example data*.
- *Accurate model building* can be a time-consuming and frustrating process, especially for complex data and has to be done again and again for different data. Machine learning promises to reduce the use of human resources, also leading to more accurate models.



Disadvantages of Machine Learning

- The resulting models are often black boxes. That makes it often difficult to evaluate the reliability and robustness of a model. Since there is a zoo of available ML algorithms, *model agnostic* methods are needed.
- It may be difficult to explain and interpret a model. For example, a credit scoring model should be transparent why a certain customer is getting a credit but another not.
- The model can only learn from the data which is available to *train* it. If the *training data* does not reflect or cover the *target population* well or contains *stereotypes* (e.g. gender stereotypes in text data), its application may lead to so-called *biased* results.
- The computational resources needed are often very high (e.g. GPUs are still expensive).
- Depending on the complexity of the problem, a high number of examples (statistically: a high sample size) is needed to build accurate models.

A Formal Definition

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .”

(Mitchell 1997)

- Learning: Attaining the ability to perform a task T .
- A set of examples (“*experience*”) represents a more general task.
- Examples are (usually) described by *features*. The features are often represented by numeric vectors $\mathbf{x} \in \mathbb{R}^m$ (sometimes also by matrices, tensors).

Data (Experience)

Typical rectangular data sets are defined as follows.

- Dataset: collection of examples
- Data is stored in a *design matrix*:

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

- m : number of examples
- n : number of features
- Examples: $X_{i,j}$ is the count of feature j (e.g. a stem form) in document i or the intensity of the j 'th pixel in image i

Characterization of the Learning Process

An often used characterization is *supervised* versus *unsupervised* learning.

- Unsupervised learning:
 - ▶ Model the data in \mathbf{X} , or find interesting properties of \mathbf{X} .
 - ▶ Example: Clustering (find groups of similar images/documents)
 - ▶ Training data: only \mathbf{X} .
- Supervised learning:
 - ▶ *Predict specific* additional properties from \mathbf{X}
 - ▶ E.g., sentiment classification: Predict sentiment (1–5) of amazon reviews
 - ▶ In the training data, additional *labels* (the *outputs*) are available for each example, i.e. an additional label vector $\mathbf{y} \in \mathbb{R}^m$ together with the *input* \mathbf{X} can be used for training the model.
 - ▶ The task is then to predict the labels of new data, where only the features \mathbf{X} are known.
 - ▶ The *performance* of this prediction is measured by a *loss function* and the task is usually solved by an optimization which minimizes the loss.

Supervised and Unsupervised Learning: data generation process (DGP)

- Unsupervised learning: Learn interesting properties, such as the probability distribution $p(\mathbf{x})$, i.e. estimate the distribution
- Supervised learning: learn mapping from \mathbf{x} to y , typically by estimating the conditional distribution $p(y|\mathbf{x})$. The supervised learning task is often formulated as the problem of estimating (or approximating) a function f , such that

$$y = f(\mathbf{x}) + \varepsilon ,$$

where ε is a random component (*error term*) describing the probabilistic nature of the task.

- Supervised learning in an unsupervised way (Bayes theorem):

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'} p(\mathbf{x}, y')} = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})}$$

Note: if y is continuous, the sum (in the denominator) must be replaced by an integral.

Machine Learning Tasks

Types of Tasks:

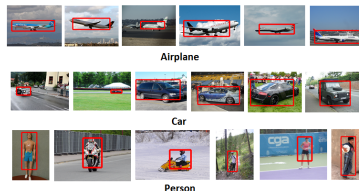
- Classification
- Regression
- Structured Prediction
- Anomaly Detection
- Synthesis and sampling
- Imputation of missing values
- Denoising
- Clustering
- Reinforcement learning
- ...

Task: Classification

- Which of k classes does an example belong to?

$$f : \mathbb{R}^n \rightarrow \{1 \dots k\}$$

- Typical example: Categorize image patches
 - ▶ Feature vector: color intensities for each pixel; derived features.
 - ▶ Output categories: Predefined set of labels



- Typical example: Spam Classification
 - ▶ Feature vector: High-dimensional, sparse vector.
Each dimension indicates occurrence of a particular word, or other email-specific information.
 - ▶ Output categories: “spam” vs. “ham”

Task: Classification

Identifying civilians killed by police with distantly supervised entity-event extraction

**Katherine A. Keith, Abram Handler, Michael Pinkham,
Cara Magliozzi, Joshua McDuffie, and Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst

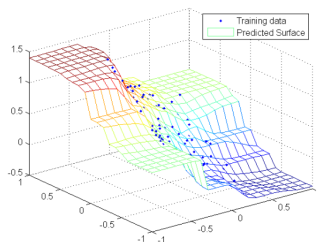
- EMNLP 2017: Given a **person name** in a sentence that contains **keywords** related to **police** (“officer”, “police” ...) and to **killing** (“killed”, “shot”), was the person a civilian killed by police?

| Text | Person killed by police? |
|--|-----------------------------|
| Alton Sterling was killed by police. | True |
| Officers shot and killed Philando Castile . | True |
| Officer Andrew Hanson was shot. | False |
| Police report Megan Short was fatally shot in apparent murder-suicide. | False |

Task: Regression

- Predict a numerical value given some input.

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$



- Typical examples:
 - ▶ Predict the risk of an insurance customer.
 - ▶ Predict the value of a stock.

Task: Regression

Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses

Ryan Lowe♡*

Michael Noseworthy♡*

Iulian V. Serban[◇]Nicolas A.-Gontier[♡]Yoshua Bengio Joelle Pineau

► [arXiv link](#)

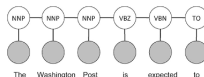
- ACL 2017: Given a response in a multi-turn dialogue, predict the value (on a scale from 1 to 5) how natural a response is.

| Context | Reference response | Model responses | Human score | ADEM score |
|---|---|--|-------------|------------|
| photo to see my television debut go to - some. some on <url> - hehe <url> → it really was you? i thought ppl were recognizing someone who looked like you! were the oysters worth the wait? | yeah it was me . haha i 'd kinda forgotten about it it was filmed a while ago | 1) i'm not sure. i just don't know what to do with it. | 3 | 1.602 |
| | | 2) you heard the horsepower productions remix of lee scratch perry's 'exercising' off his 'mighty upsetter' album? | 1 | 1.513 |
| | | 3) you wont chug a fuzzy peach navel | 1 | 1.744 |
| | | 4) they were! | 5 | 3.274 |
| just beat call of duty!! → want a cookie? → yes!! → come get it | im in kenmore at the moment | 1) i'm gonna get a new phone some moro | 1 | 1.848 |
| | | 2) no way man. | 5 | 4.265 |
| | | 3) wow i just got a free pizza coupon! get yours before theres no more! <url> | 1 | 0.921 |
| | | 4) i'm going to go to the mall. | 1 | 2.634 |
| am i out of twitter jail yet? testing → yeah. i posted bail → thanks. i am a right chatter tweetbox on sundays. same happened last sunday lol | any news on meeting our user ? i go to the us on friday and i don 't want to miss anything arranged | 1) i'm not sure if i'm going to be able to get it. | 3 | 1.912 |
| | | 2) good to see another mac user in the leadership ranks | 4 | 1.417 |
| | | 3) awww poor baby hope u get to feeling better soon. maybe some many work days at piedmont | 2 | 1.123 |
| | | 4) did you tweet too much? | 5 | 2.533 |

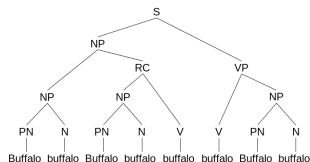
Task: Structured Prediction

- Predict a multi-valued output with special inter-dependencies and constraints.
- Typical examples:

- ▶ Part-of-speech tagging



- ▶ Syntactic parsing



▶ Wikipedia

- ▶ Machine Translation

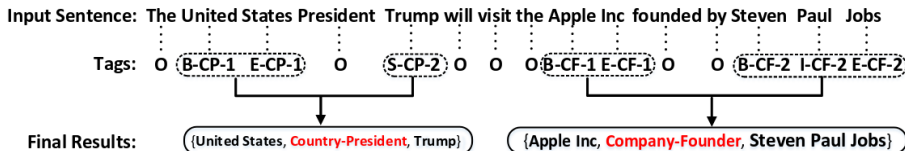
- Often involves search and problem-specific algorithms.

Task: Structured Prediction

Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, Bo Xu
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, P.R. China

- ACL 2017: jointly find all relations of interest in a sentence by tagging arguments and combining them. [▶ Paper](#)



Task: Reinforcement Learning

- In **reinforcement learning**, the model (also called **agent**) needs to select a series of actions, but only observes the outcome (**reward**) at the end.
- The goal is to predict actions that will maximize the outcome.

Deal or No Deal? End-to-End Learning for Negotiation Dialogues

Mike Lewis¹, Denis Yarats¹, Yann N. Dauphin¹, Devi Parikh^{2,1} and Dhruv Batra^{2,1}




¹Facebook AI Research

²Georgia Institute of Technology

- EMNLP 2017: The computer negotiates with humans in natural language in order to maximize its points in a game. [▶ Paper](#)

Divide these objects between you and another Turker. Try hard to get as many points as you can!

Send a message now, or enter the agreed deal!

| Items | Value | Number You Get |
|---|-------|--------------------------------|
|  | 8 | <input type="text" value="1"/> |
|  | 1 | <input type="text" value="1"/> |
|  | 0 | <input type="text" value="0"/> |

Mark Deal Agreed ✓

Fellow Turker: I'd like all the balls

You: Ok, if I get everything else

Fellow Turker: If I get the book then you have a deal

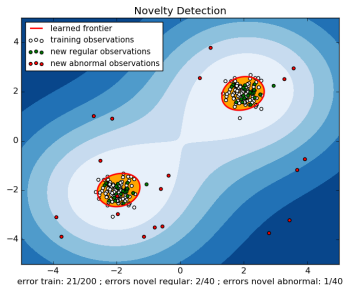
You: No way - you can have one hat and all the balls

Fellow Turker: Ok deal

Type Message Here:

Task: Anomaly Detection

- Detect atypical items or events.
- Common approach: Estimate density and identify items that have low probability.



- Examples:
 - ▶ Quality assurance
 - ▶ Detection of criminal activity
- Often items categorized as outliers are sent to humans for further scrutiny.


Task: Anomaly Detection

Using Automated Metaphor Identification to Aid in Detection and Prediction of First-Episode Schizophrenia

E. Darío Gutiérrez¹ Philip R. Corlett² Cheryl M. Corcoran³ Guillermo A. Cecchi¹

- ACL 2017: Schizophrenia patients can be detected by their non-standard use of metaphors, and more extreme sentiment expressions. [▶ Paper](#)

Frequently Used Models and Algorithms in supervised ML

| Supervised | Label | Task |
|--|----------------|----------------------------|
| linear regression | numeric | regression |
| logistic regression | binary | regression, classification |
| decision trees (e.g. C 4.5) ¹ | class | classification |
| CART ² | class, numeric | classification, regression |
| random forest ³ | class, numeric | classification, regression |
| Boosting, e.g.  | class, numeric | classification, regression |
| (artificial) neural networks | (see later) | (see later) |

Note: Class variables are often called categorical variables in statistics.

¹Quinlan, J.R. (1992). C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann.

²L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone: CART: Classification and Regression Trees. Wadsworth: Belmont, CA, 1984.

³Breiman L., Random forests. In: Machine Learning, 2001, 45(1), Seiten 5–32, doi:10.1023/A:1010933404324

Some Algorithms for unsupervised ML

| Unsupervised | Task |
|---|-------------------------------------|
| Clustering (k -means) | group similar objects |
| Clustering (partition around medoids (PAM)) | group similar objects |
| PCA (principal components analysis) | dimension reduction |
| k -nearest neighbor | outlier detection |
| Isolation forest ⁴ | anomaly detection |
| Autoencoders | representation, data compression |

⁴Liu, F. T., Ting, K. M., and Zhou, Z.-H. 2008a. Isolation Forest. In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, 413–422.

Performance Measures

“A computer program is said to learn [...] with respect to some [...] **performance measure P** , if its performance [...] **as measured by P** , improves [...]”

- Quantitative measure of algorithm performance.
- Task-specific.

Discrete vs. Continuous Loss Functions

• Discrete Loss Functions

- ▶ Accuracy (how many samples were correctly labeled?)
- ▶ Error Rate (1 - accuracy)
- ▶ Precision / Recall
- ▶ Accuracy may be inappropriate for skewed label distributions, where relevant category is rare. Often used is the F1-Score (harmonic mean of precision and recall):

$$\text{F1-score} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$$

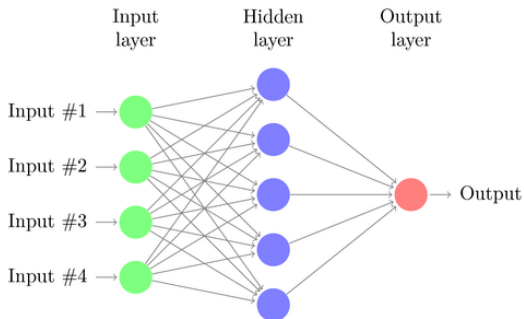
- ▶ F1-Score does not take into account the true negatives. Proposed alternative: Matthews correlation coefficient (MCC), $-1 \leq \text{MCC} \leq 1$.
- Discrete loss functions cannot indicate **how wrong** a wrong decision is.
- They are not differentiable (hard to optimize)
- Often algorithms are optimized using a continuous loss (e.g. hinge loss) and evaluated using another loss (e.g. F1-Score).

Examples for Continuous Loss Functions

- Squared error (regression): $(y - f(\mathbf{x}))^2$
- Hinge loss (classification):
 - ▶ $\max(0, 1 - f(\mathbf{x}) \cdot y)$
 - ▶ (assume that $y \in \{-1, 1\}$)
- ...
- These loss functions are differentiable. So we can use them for gradient descent (more on that later).

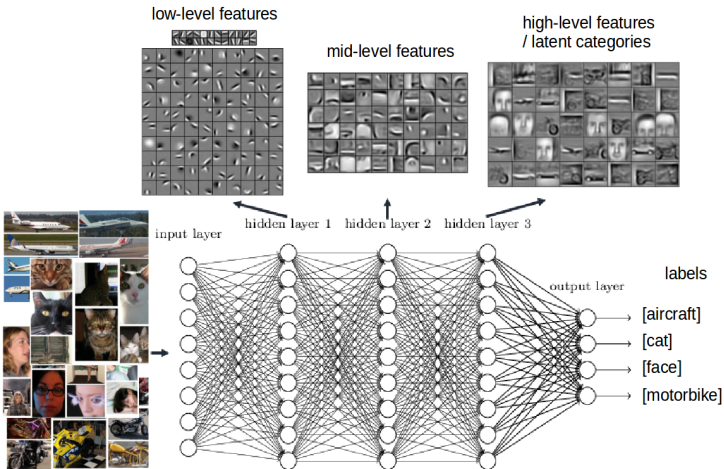
Deep Learning

- Learn complex functions, that are (recursively) composed of simpler functions.
- Many parameters have to be estimated.



Deep Learning

- Main Advantage: Feature learning
 - ▶ Models learn to capture *most essential* properties of data (according to some performance measure) as intermediate representations.
 - ▶ No need to hand-craft feature extraction algorithms



Neural Networks

- First training methods for deep nonlinear NNs appeared in the 1960s (Ivakhnenko and others).
- Increasing interest in NN technology (again) since around 10 years ago (*“Neural Network Renaissance”*):
Orders of magnitude more data and faster computers now.
- Many successes:
 - ▶ Image recognition and captioning
 - ▶ Speech recognition
 - ▶ NLP and Machine translation
 - ▶ Game playing (AlphaGO)
 - ▶ ...

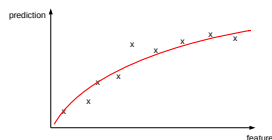
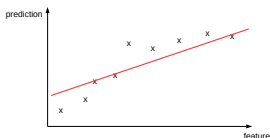
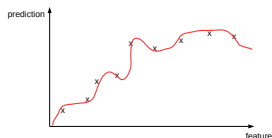
Machine Learning

- Deep Learning builds on general Machine Learning concepts

$$\operatorname{argmin}_{\theta \in \mathcal{H}} \sum_{i=1}^m \mathcal{L}(f(\mathbf{x}_i; \theta), y_i)$$

\mathcal{L} is called *loss function* or *cost function* (we will discuss it further in the next chapter).

- Fitting data vs. generalizing from data



Summary

- Machine learning definition
 - ▶ Data
 - ▶ Task
 - ▶ Cost function
- Machine learning tasks
 - ▶ Classification
 - ▶ Regression
 - ▶ ...
- Deep Learning
 - ▶ many successes in recent years
 - ▶ feature learning instead of feature engineering
 - ▶ builds on general machine learning concepts