

Ringvorlesung Artificial Intelligence Introduction, Statistical Learning, PAC, and VC-Theory

Prof. Dr. Jörg Schäfer
Frankfurt University of Applied Sciences

Fachbereich 2 Informatik und Ingenieurwissenschaften
Computer Science and Engineering

<https://joerg-schaefer.biz/academics>

 2025–2025 Jörg Schäfer

10.03.2025

Abstract

This short note “Introduction, Statistical Learning, PAC, and VC-Theory” introduces into the field of AI and provides a broad overview. It also sketches the main framework and results for machine learning, in particular statistical learning theory including the theory of probably approximately correct PAC results, generalisation bound by Vapnik and the core definitions of VC-theory. It does not provide any mathematical details and proofs.

Contents

1	Introduction	5
1.1	AI Definition	5
1.2	Examples of Machine Learning	12
1.3	Definition of Machine Learning	16
1.4	Machine Learning Types	17
1.5	Machine Learning Model	19
1.6	Summary AI Landscape	22
2	Statistical Learning, PAC and VC Theory	23
2.1	Statistical Learning Model	23
2.2	Problem Statement	23
2.2.1	Features	24
2.2.2	Labels	24
2.2.3	Model	24
2.2.4	Hypothesis	24
2.2.5	Error Measures (KPIs)	25
2.2.6	Confusion Matrix	25
2.2.7	Learning or Training	25
2.3	Examples	25
2.3.1	Credit Approval	25
2.3.2	Cats and Dogs	26
2.3.3	Regression	26
2.4	Probably Approximately Correct (PAC) Theory	26
2.4.1	Problem Statement	26
2.4.2	Generalization Bound	27
2.4.3	Feasibility of Learning	29
2.4.4	KPIs and Validation	29

List of Figures

1.1	MNIST data set	13
1.2	MNIST Fashion data set	14
1.3	Cats or Dogs	14
1.4	Grace Hopper	14
1.5	Multidimensional scaling plot of patients with prostate disease. Black dots indicate individual patients with prostate cancer; gray dots, patients with noncancerous disease.	15
1.6	Performance of Several ML-Algorithm using the Breast Cancer Wisconsin (Diagnostic) Database	16
1.7	Autonomous Driving	16
1.8	Learning Model, source: [1]	20
2.1	Learning Model, source: [1]	24
2.2	Regression Housing Prices	26
2.3	Counting Hypothesis	28

Chapter 1

Introduction

1.1 AI Definition

AI

“Artificial intelligence is the science of making machines do things that would require intelligence if done by men.” – Marvin Minsky

“Artificial intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs.”
– John McCarthy

“Artificial intelligence is a field of study and technology focused on creating machines and computer systems that can exhibit human-like intelligence.” – Yoshua Bengio

“I believe that deep learning will allow us to build AI systems that can understand and interact with the world in a way that is far beyond what is possible with current technology.” – Geoffrey Hinton

Intelligence

Definition 1 (human intelligence). *mental quality that consists of the abilities to learn from experience, adapt to new situations, understand and handle abstract concepts, and use knowledge to manipulate one's environment.*

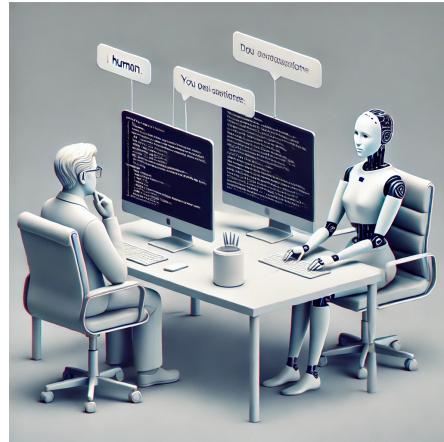
Source: Britannica <https://www.britannica.com/science/human-intelligence-psychology>

Definition 2 (artificial intelligence (AI)). *the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.*

Source: Britannica <https://www.britannica.com/technology/artificial-intelligence>

Todo [4]

Turing Test



Source: Generated by ChatGPT

Eliza

```
Welcome to
      EEEEEE  LL      IIII    ZZZZZZ  AAAAAA
      EE      LL      II      ZZ      AA  AA
      EEEEEE  LL      II      ZZZ    AAAAAAAA
      EE      LL      II      ZZ      AA  AA
      EEEEEE  LLLLLL  IIII    ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Source: Public Domain, https://en.wikipedia.org/wiki/ELIZA#/media/File:ELIZA_conversation.png

Eliza built into Emacs

Try out Alt M-x doctor

Simulating Eliza with ChatGPT – I /II

- » Please simulate an Eliza session as of the 60's (Weizenbaum)
- « Sure! Here's a simulated ELIZA session in the style of the original 1960s program by Joseph Weizenbaum, which used Rogerian therapy techniques to reflect back and encourage users to elaborate.

» ELIZA: Hello. I am ELIZA. How do you do? Please tell me your problem.

» User: I've been feeling anxious lately.

» ELIZA: I am sorry to hear that you are feeling anxious. Can you tell me more about it?

» User: I just feel overwhelmed with work and life in general.

» ELIZA: What about work and life makes you feel overwhelmed?

» User: There's just so much to do, and I don't have enough time.

» ELIZA: You say you don't have enough time. Could you explain that further?

» User: Well, between my job, family, and personal projects, I can't seem to keep up.

» ELIZA: Your job, family, and personal projects seem to be important to you. How do you feel about them?

» User: I love them, but they also stress me out.

Simulating Eliza with ChatGPT – II /II

» OK well done, can you assume for the time being ELIZA and I do the user prompt?

» Sure! I'll act as ELIZA, and you can be the user. Go ahead and type your response.

» ELIZA: Hello. I am ELIZA. How do you do? Please tell me your problem.

» You said: Men are all alike

» ELIZA: In what way?

» You said: They are always bugging us about something or the other

» ELIZA: Can you think of a specific example?

» You said: Well, my boyfriend made me come here

» ELIZA: Your boyfriend made you come here?

» You said: Yes

» ELIZA: I see. Do you often let others influence your decisions?

» You said: Sometimes, Do you think it's bad?

» ELIZA: I am not here to judge. How do you feel about it?

» You said: My boyfriend says I am depressed much of the time

» ELIZA: I am sorry to hear that you are feeling depressed. Do you think your boyfriend's observation is accurate?

» You said: Yes I am unhappy

» ELIZA: I am sorry to hear that you are unhappy. What do you think is causing this unhappiness?

Data, Information, Knowledge

We are drowning in information but starved for knowledge.

John Naisbit, "Megatrends" (1982)

Data

Definition 3 (Data). *Data refers to raw, unprocessed facts and figures without any context or interpretation. It is the basic building block from which information and knowledge are derived.*

Examples:

- Numbers, text, symbols, images, and measurements
- A list of temperatures recorded by a weather station
- A database of customer transactions

Characteristics:

- Unprocessed and unorganized
- Needs to be processed to become meaningful
- Can be quantitative or qualitative

Information

Definition 4 (Information). *Information is data that has been processed, organized, and structured in a way that makes it meaningful and useful. It provides context and relevance to data.*

Examples:

- A weather report that interprets temperature data to provide forecasts
- A sales report that analyzes customer transaction data to identify trends
- An article that explains scientific findings based on experimental data

Characteristics:

- Processed and organized data
- Provides context and relevance
- Useful for decision-making and understanding

Knowledge

Definition 5 (Knowledge). *Knowledge is the application of information, experience, and expertise to make decisions, solve problems, and take actions. It involves understanding and insight gained through learning and experience.*

Examples:

- A meteorologist using weather information to predict future weather patterns
- A business analyst using sales information to develop strategies for increasing revenue
- A doctor using medical information to diagnose and treat a patient

Characteristics:

- Applied information and experience
- Involves understanding and insight
- Used for decision-making, problem-solving, and action

Data, Information, Knowledge – Summary

Summary

- Data is the raw material.
- Information is processed data that has meaning and context.
- Knowledge is the application of information, experience, and expertise to make decisions and take actions. Understanding these distinctions is crucial in fields like artificial intelligence, where data is collected, processed into information, and then used to generate knowledge that can drive decision-making and problem-solving.

AI Fields

Reasoning

- Theorem Proving
- Automated Reasoning
- Reasoning with Uncertainties

Planning and Decision-making

- Agents
- Games
- Markov decision processes

Knowledge

- Knowledge Representation
- Knowledge Graphs
- Ontologies

Perceiving

- Computer Vision
- Speech recognition

Learning

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Learning
- (Big) Data Mining

Communication

- NLP
- Speech Synthesis

Reasoning Fundamentals

- based on Logical Reasoning with Propositional and First-Order Logic
- uses Rule-Based Systems
- Fuzzy Logic and Probabilistic Reasoning incl. Bayesian Networks to deal with uncertainty

Will be (partially) covered in lectures on Symbolic AI and Semantik/Ontology

Knowledge Fundamentals

- based on also Logical Reasoning with Propositional and First-Order Logic
- captures Semantics in Semantic Networks such as Knowledge Graphs and Ontologies
- uses Case-Based Reasoning (CBR) for adapting solutions to known solutions

Will be (partially) covered in lectures on Symbolic AI and Semantik/Ontology

Perceiving Fundamentals

- Computer Vision includes techniques such as filtering, edge detection, and segmentation
- Feature Extraction with convolution networks
- Object detection
- Speech detection uses Hidden Markov Models (HMMs), N-Grams and Deep Learning

Will be (partially) covered in lectures on Gen AI, Computer Vision and CNN

Communication Fundamentals

- Natural Language Processing
- Tokenisation
- Word Embeddings
- Bag of Words
- Transformers
- and many more

Will be (partially) covered in lectures on Gen AI, NLP

Machine Learning Fundamentals

- Is firmly grounded on probability theory and statistics
- Based on Statistical Learning Theory [6] (Supervised ML)
- Based on Markov Decision Processes and Bellman Equations (Reinforcement Learning)

Will be (partially) covered in this lecture and lectures on Regularisation, Bayes, Testing, and Validation, KNN, SVM, Random Forest, Trees, and NN and Reinforcement Learning

Machine Learning Fundamentals

The remainder of this (and the next) lecture will exclusively focus on

Machine Learning \ Reinforcement Learning

Remark: For a good general overview on AI please have a look into [4].

1.2 Examples of Machine Learning

Machine Learning Examples

Ideas for examples of Machine Learning applications?

Example One: Mortgage Approval

- Assume we have to decide to provide mortgages to potential customers
- In this problem, given some features of clients such as age, gender, annual salary, years in job, ...
- The goal of the model is to approve a new customer or not based on his/her data.

Customer	Age	Gender	Salary	Years in Job	...
Müller	45	m	80000	23	...
Meyer	25	f	40000	2	...
Scholz	19	d	120000	0	...
...

Example Two: Handwriting Recognition

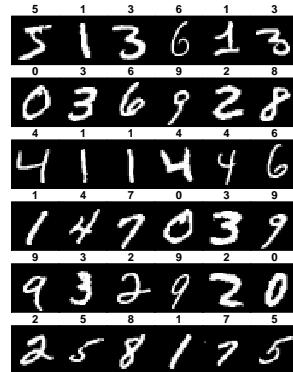


Figure 1.1: MNIST data set

Source: Suvanjanprasai, CC BY-SA 4.0, https://en.wikipedia.org/wiki/MNIST_database#/media/File:MNIST_dataset_example.png

Example Three: Fashion Classification

Source: [7]

Example Four: Image Classification

Example Five: Image Labeling

```
bazel-bin/tensorflow/examples/label_image/label_image
59.026866: examples/label_image/main.cc:250] military uniform (653): 0.834306
59.026906: examples/label_image/main.cc:250] mortarboard (668): 0.0218692
59.026920: examples/label_image/main.cc:250] academic gown (401): 0.0103579
59.026933: examples/label_image/main.cc:250] pickelhaube (716): 0.00800814
59.026944: examples/label_image/main.cc:250] bulletproof vest (466): 0.00535088
```

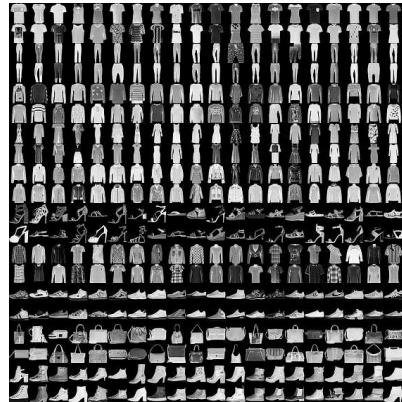


Figure 1.2: MNIST Fashion data set

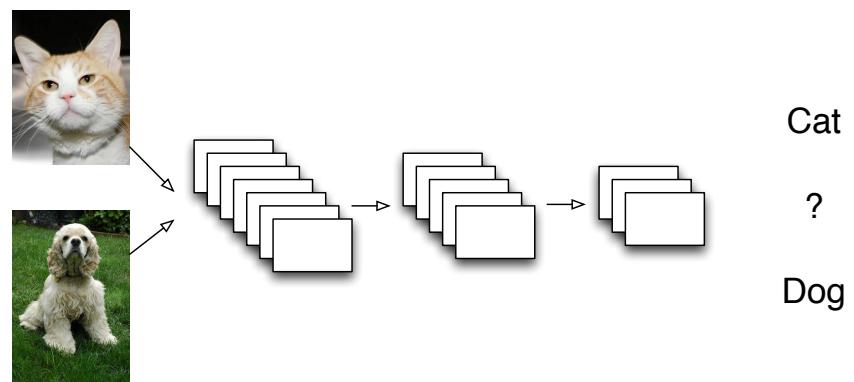


Figure 1.3: Cats or Dogs



Figure 1.4: Grace Hopper

Source: James S. Davis - Naval History and Heritage Command https://de.wikipedia.org/wiki/Grace_Hopper

#/media/Datei:Grace_Hopper.jpg

Example Six: Train Machines to Play Soccer



For more information, see <https://www.frankfurt-university.de/de/hochschule/fachbereich-2-informatik-und-ingenieurwissenschaften/einrichtungen-services/roboterfussball-team-fra-united/-FRA-UAS-has-scored-runners-up-at-world-championship-RoboCup-2017>.

Example Seven: Predict (Prostate) Cancer Example 2

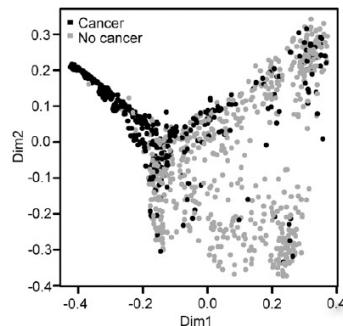


Figure 1.5: Multidimensional scaling plot of patients with prostate disease. Black dots indicate individual patients with prostate cancer; gray dots, patients with noncancerous disease.

Source: Li-Hong Xiao et al. Asian J Androl. 2017 Sep-Oct; 19(5): 586–590

Example Eight: Predict Breast Cancer

One can calculate ML algorithms online, see <https://www.kaggle.com/jschae02/kernels/notebooks/new?forkParentScriptVersionId=1288066> or <http://scikit-learn.org/stable/>.

Below an example of comparison of performance of several ML-Algorithm using the Breast Cancer Wisconsin (Diagnostic) Database, see [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)):

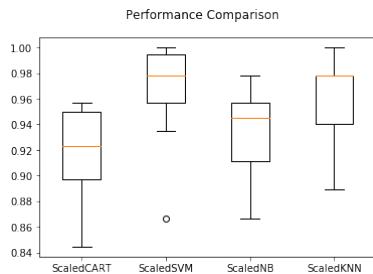


Figure 1.6: Performance of Several ML-Algorithm using the Breast Cancer Wisconsin (Diagnostic) Database

Example Nine: Autonomous Driving



Figure 1.7: Autonomous Driving

Source: Michael KR https://de.wikipedia.org/wiki/Autonomes_Fahren#/media/File:Daimler_2014_Mercedes_Autonomes_Fahren_Magdeburg_5430.jpg

1.3 Definition of Machine Learning

Machine Learning – Definitions

Definition 1. Machine learning is the science of getting computers to act without being explicitly programmed. (Andrew Ng, Stanford)

Definition 2. Machine learning answers the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” (Tom M. Mitchell, Carnegie Mellon University)

Definition 3. The learning process is a process of choosing an appropriate function from a given set of functions. (Vladimir N. Vapnik, Statistical Learning Theory)

Machine Learning and Statistics

Table 1.1: Glossary, Larry Wasserman

Statistics	Machine Learning
Estimation	Learning
Classifier	Hypothesis
Data Point	Example/ Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label
R	Matlab ¹

“Machine learning is glorified statistics”, Robert Tibshirani, Stanford

1.4 Machine Learning Types

Nomenclature

Machine learning approaches are categorized into

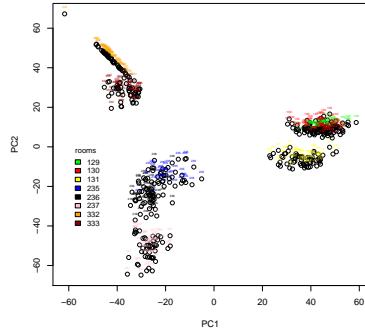
1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

based on whether there is a learning signal, teacher, expert, or feedback available to the learning system.

Supervised Learning

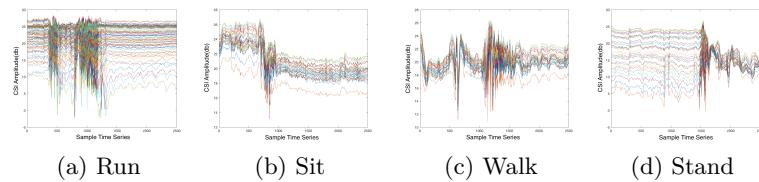
- In supervised learning we have (enough) data with correct outputs.
- These outputs can be used to train the model.
- For example, for mortgage rate approval we have data of a lot of customers with Age, Gender, Salary, Years in Job, etc. called *features* or *feature vector*. In addition, we *know* which of the customers defaults (from e.g. history).
- There are two main categories of Supervised Learning:
 1. Regression, where the output is continuous, i.e. $y \in \mathbb{R}$ or
 2. Classification, where the output is non-continuous and discrete and represented as class labels, i.e. $y \in L \subset \mathbb{N}$, where L is finite.

Supervised Learning Example – I/II



Source: [5]

Supervised Learning Example – II/II



Source: [2]

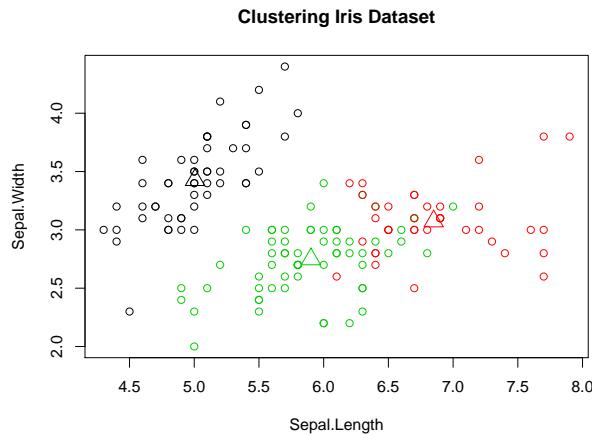
Unsupervised Learning

- In unsupervised learning we do not have data with correct outputs.
- Rather, we want the machine to learn structure by itself.

Whilst, this might seem an ill-defined problem, there are successful algorithms like e.g.

- Clustering: The goal of clustering is to group the given data automatically into groups/clusters. Each group has similar or related features
- Cocktail Party algorithm, separating sound sources, see <https://arxiv.org/abs/1504.04658>

Unsupervised Learning Example



Reinforcement Learning

- Reinforcement learning inspired by behaviorist psychology
- Concerned with how machine should ought to take actions in a specific environment to maximize some utility function
- Similar to how we teach children by reward and punishment

Similar to supervised learning but we do not have correct / incorrect labelling but rather provide *feedback* on actions chosen.

Reinforcement Learning Example: Robot Soccer

<http://www.tgabel.de/index.php?id=publications>

More details in separate lectures later!

1.5 Machine Learning Model

The (generic) Learning Model

The dataset $D = (x_i, y_i)_{i=1}^N$ is known, but f is not. The goal is to learn f from the data D :

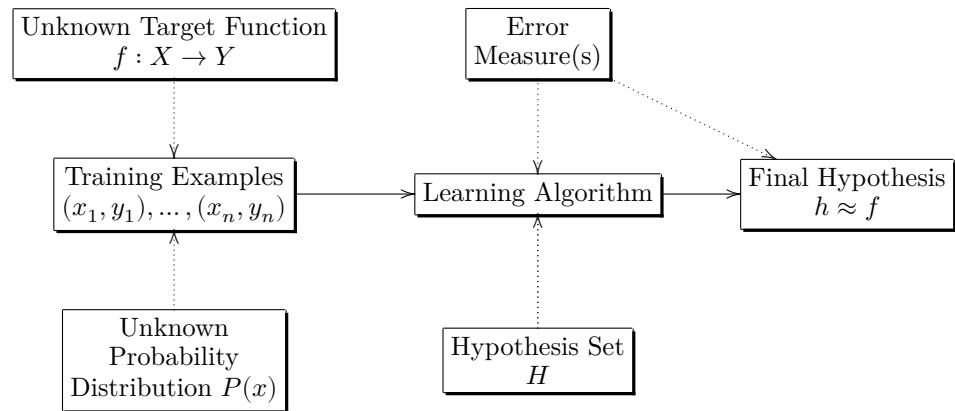


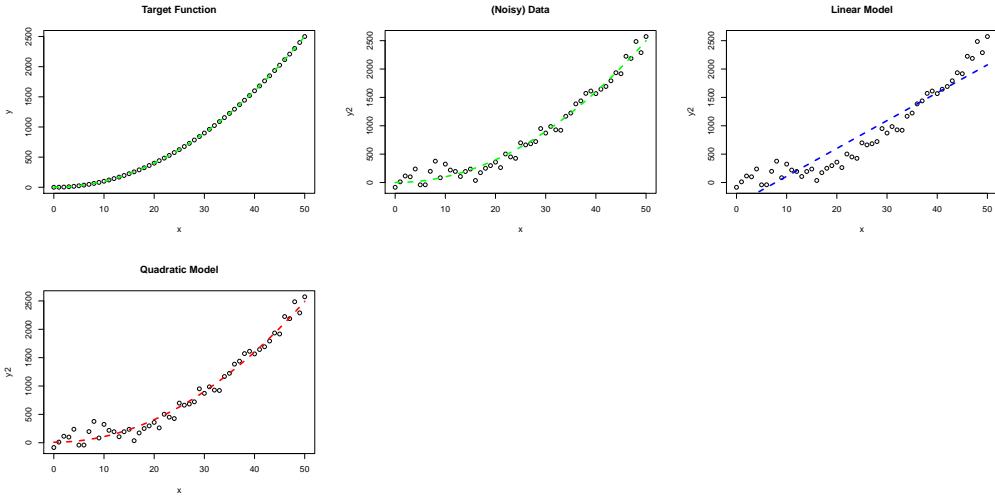
Figure 1.8: Learning Model, source: [1]

Probability is at the Core of ML



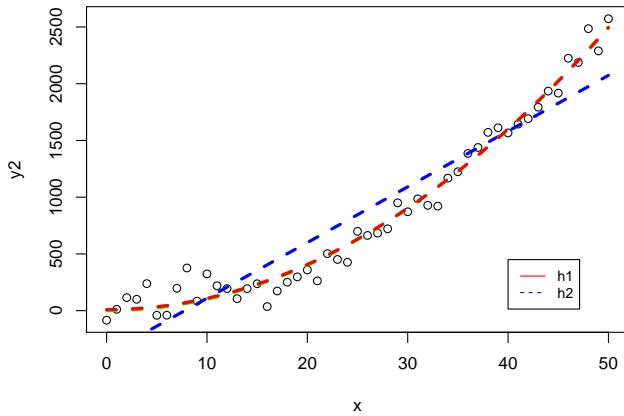
Source: Ralf Roletschek <https://commons.wikimedia.org/wiki/File:13-02-27-spielbank-wiesbaden-by-RalfR-094.jpg>

The Learning Model Applied



The Learning Model Applied

Two Hypothesis



How to choose? Remember, we do *not* know f !

The Learning Model Applied

Every learning model consists of two components:

1. The learning algorithm such as Linear or Non-linear Regression or Neural Network Backwards Propagation.
2. The hypothesis set $H = \{h_1, \dots, h_n\}$.

A selection of one $h_i \in H$ is often called the trained model.

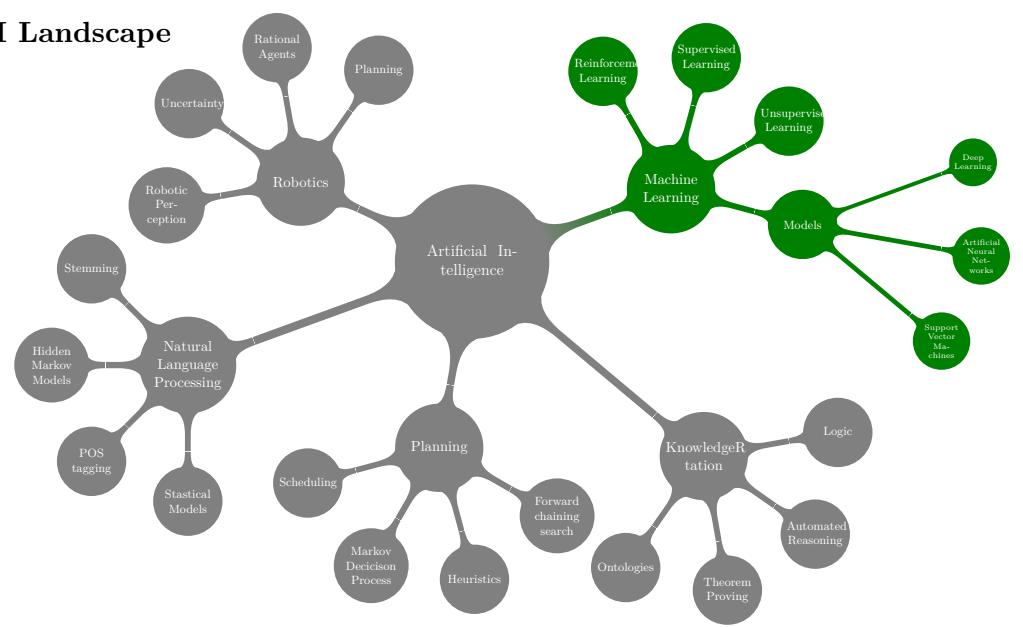
We want to select an $h_i \approx f$ in a certain sense.

Question: Is this even possible?

Answer: We will see in the next lecture (bear with me!).

1.6 Summary AI Landscape

AI Landscape



Chapter 2

Statistical Learning, PAC and VC Theory

Statistical Learning is deeply based on statistical concepts. It has been pioneered by Vapnik [6] and is also known under the term Probably Approximately Correct (PAC) Theory to indicate that it is (only) probably correct and only approximately i.e. in general never exactly correct. These terms will be explained in the sequel.

2.1 Statistical Learning Model

We will follow the exposition in [1] and use the following established model cf. Fig. 2.1, the details are explained in the sequel.

Assume we have given the following data:

- *Input:* $x \in X = \mathbb{R}^d$, d is the number of *features*
- *Output:* $y \in Y = \mathbb{R}$ or $y \in Y = \{-1, +1\}$ is the *expected* output. Can be continuous (e.g. regression) or discrete (classification)
- *Target function:* $f: X \mapsto Y$. Is *unknown*; thus we need to learn it from *the data*
- *Data:* $D := \{(x_1, y_1), \dots, (x_n, y_n)\}$, where $y_i = f(x_i)$
- *Hypothesis:* $g: X \mapsto Y$ (best hypothesis/best guess – the formula to be used for *prediction*)

2.2 Problem Statement

x, y and D are known, but f is not. The goal is to learn this function from the data D .

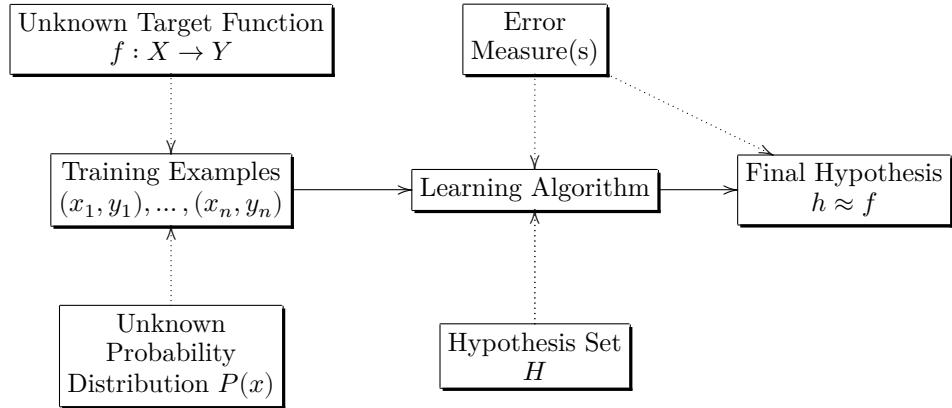


Figure 2.1: Learning Model, source: [1]

2.2.1 Features

We assume the $x \in X$ are properties that characterise the data. We assume that they are randomly sampled from an unknown probability distribution together with their corresponding labels $y \in Y$, i.e., we assume that we are given a number n of training data (supervised): $D := \{(x_1, y_1), \dots, (x_n, y_n)\}$

2.2.2 Labels

We assume the $y \in Y$ are so-called labels characterising the data by either a categorically or numerically (Classification or Regression).

2.2.3 Model

The (machine learning) model is any algorithm that has access to the data D , a hypothesis set H (see below) and tries to “learn” i.e., select the “best” hypothesis by making use of the data. In order to do so it needs to be given criteria for “best”, i.e., error measures or KPIs.

2.2.4 Hypothesis

The hypothesis set is a fixed set of functions H that we chose the “best” from. It can be shown that we cannot learn if we do *not* restrict this set carefully. taking *all* functions does not work and in general, the “more” (complex) functions we use, the more difficult learning becomes.

2.2.5 Error Measures (KPIs)

Error Measures or KPIs define how well we approximate (PAC!) the unknown function f by our hypothesis h chosen. For categorial data we define the so-called in-sample error as

$$E_{in}(h) := \frac{1}{n} \sum_{x_i \in D} \mathbb{1}|f(x_i) \neq h(x_i)|. \quad (2.1)$$

And for numerical data we define

$$E_{in}(h) := \left(\frac{1}{n} \sum_{x_i \in D} (f(x_i) - h(x_i))^2 \right)^{\frac{1}{2}}. \quad (2.2)$$

Please note, that other error measures or KPIs are possible and used.

The Out-of-Sample Errors E_{out} are defined similarly where the data set is taken from out-of-sample data.

2.2.6 Confusion Matrix

It is useful to display predictions vs. actual (true) values in tabular format – a confusion matrix:

	Predicted Positive	Predicted Negative
Actual Positive	90 (TP)	10 (FN)
Actual Negative	5 (FP)	95 (TN)

Table 2.1: Example Binary Confusion Matrix

2.2.7 Learning or Training

Learning or training refers to *any* algorithm that selects a “best” – based on the KPIs chosen – hypothesis and the training data provided. It is very often but not in any case an iterative computation n-intense process. For example, simple ML models such as KNN or Linear Regression might not require heavy computation.

2.3 Examples

2.3.1 Credit Approval

Context: Bank wants to assess credit worthiness from customer data. We have the following data as features¹

¹Exercise: Model X , note that Gender is not numeric – how can it be made to fit into $x \in \mathbb{R}^3$?

1. Age
2. Gender
3. Salary

The goal is to predict $y \in Y := \{0, 1\}$, where 1 indicates good credit risk, i.e., customer most likely will satisfy the loan and 0 otherwise.

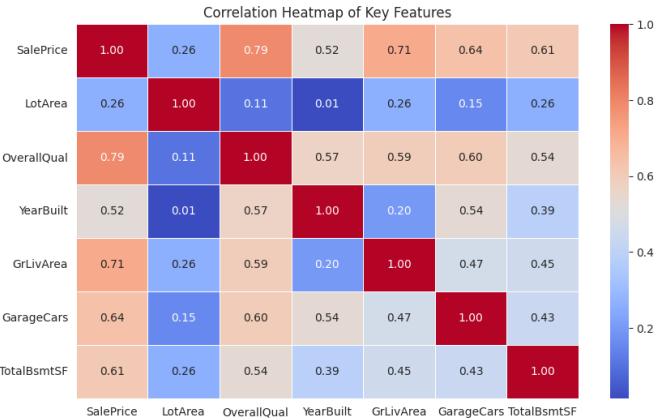
2.3.2 Cats and Dogs

Context: We have images of cats and dogs. The features are pixel (values), i.e., if the images are HD we have $x \in \mathbb{R}^{1024 \times 768}$. The labels are $y \in Y := \{0, 1\}$, where 1 encodes a cat and 0 a dog.

2.3.3 Regression

One can predict house prices (sales) from features, see e.g. <https://www.kaggle.com/code/gregorychernyavskiy/lab-3-ai2010-advanced-regression-ml-kaggle-hous>.

Figure 2.2: Regression Housing Prices



2.4 Probably Approximately Correct (PAC) Theory

2.4.1 Problem Statement

Remember, we want to train the model on train data. As it turns out, it usually will not exactly match the data, yielding E_{in} . We want to predict $f(x)$ on unseen data, and want to ensure that $E_{out} \approx E_{in}$ in other words we

want to control E_{out} by E_{in} , in other words we want to control the following probability:

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq \delta \quad (2.3)$$

for any chosen pair of ϵ (accuracy) and δ (tolerance).

As it turns out, this is a difficult problem.

2.4.2 Generalization Bound

Suppose that we only have *one* hypothesis h , i.e., that we do not learn but fix the hypothesis in advance. Then Hoeffding's inequality [3] tells us the following:

Lemma 4 (Hoeffding's Inequality).

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N} \quad (2.4)$$

Now, if we have a *finite* set of hypothesis, i.e. $H = \{h_1, \dots, h_M\}$ we get

Lemma 5 (Union Bound).

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2M e^{-2\epsilon^2 N} \quad (2.5)$$

As it turns out, almost any interesting ML model has an infinite set of hypothesis, thus this does not help. As it turns out, we can replace the constant M with a function, the so-called growth function $m_H(N)$ that characterises the “complexity” of the hypothesis set. This is the famous generalisation bound:

Theorem 6. *Let H be a hypothesis set with finite VC dimension d_{VC} . Assuming we draw sample data D of size N independently from the same unknown distribution. Then for any $\epsilon > 0$ we can bound the out-of-sample error E_{out} by the in sample error E_{in} as follows:*

$$P[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 4m_H(2N) e^{-\epsilon^2 N/8}, \quad (2.6)$$

where $m_H(N)$ denotes the growth function and $h \in H$ denotes the learned hypothesis.

The generalisation bound makes use of the following definitions and concepts.

Definition 7. Let H be a set of hypothesis. Let $x_1, \dots, x_n \in D$. Then the *dichotomies* generated by H are defined by

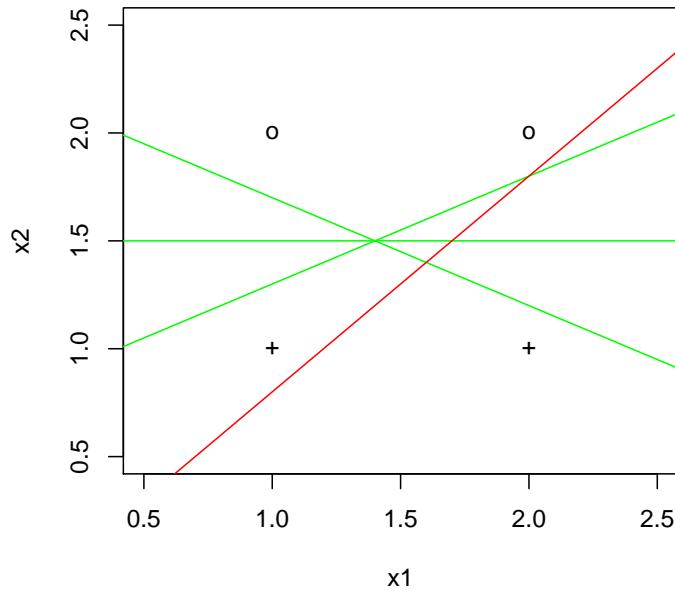
$$H(x_1, \dots, x_n) := \{(h(x_1), \dots, h(x_n)) \mid h \in H\}. \quad (2.7)$$

Example:

1. Let us assume that H is rich enough, so that we always find an $h \in H$ given x_i so that $h(x_i) = \pm 1$. Then
 - (a) Let $N = 1$. How many dichotomies exist? Answer: 2 (Why?)
 - (b) Let $N = 2$. How many dichotomies exist? Answer: 4 (Why?)
 - (c) In general there exists 2^N dichotomies! (Why?)
2. Let us assume four points as depicted in Fig. 2.3. If we assume the right hand side of the lines classifies as a $+$ and the left otherwise, all three green are “as good” in explaining the data, so we should not count them as three distinct hypothesis. This can be formalised as follows.

Remark 1. You find a nice animation at Mostafa Samir’s blog:
<https://mostafa-samir.github.io/ml-theory-pt2/>

Figure 2.3: Counting Hypothesis



Definition 8. The growth function defined for a set of hypothesis is defined as a map from $\mathbb{N} \rightarrow \mathbb{N}$

$$m_H(N) := \max_{x_1, \dots, x_N \in X} |H(x_1, \dots, x_N)|, \quad (2.8)$$

where $|\cdot|$ denotes the cardinality of the set.

Lemma 9. For any growth function we have:

$$m_H(N) \leq 2^N \quad (2.9)$$

Definition 10. Let H be a set of hypothesis. The *Vapnik-Chervonenkis dimension* of the set H , denoted by $d_{VC}(H)$ is the largest value of N such that $m_H(N) = 2^N$. If no such number exists, then $d_{VC}(H) := \infty$

2.4.3 Feasibility of Learning

The generalisation bound and equation 2.6 ensure that learning is feasible, i.e., for any given accuracy and tolerance level ϵ and δ if we have enough data N iff the RHS of equation 2.6 can be made arbitrarily small. As it turns out, if the VC dimension d_{VC} is finite, i.e., if the model is not too complex, the growth functions polynomially bounded, i.e.,

$$m_H(N) \in O(N^{d_{VC}}). \quad (2.10)$$

This implies, that $\lim_{N \rightarrow \infty} m_H(2N) e^{-\epsilon^2 N/8} \rightarrow 0$, and henceforth

$$\lim_{N \rightarrow \infty} P[|E_{in}(h) - E_{out}(h)| > \epsilon] \rightarrow 0, \quad (2.11)$$

i.e., that learning is feasible, provided we have enough training data!

2.4.4 KPIs and Validation

Equation 2.6 can be rewritten as follows (homework!). For any fixed tolerance level δ we have with probability of at least $1 - \delta$:

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \log \frac{4m_H(2N)}{\delta}} \quad (2.12)$$

$$E_{out}(h) \geq E_{in}(h) - \sqrt{\frac{8}{N} \log \frac{4m_H(2N)}{\delta}}. \quad (2.13)$$

This enables – at least in principle to calculate the number of training examples for a given accuracy and tolerance level ϵ and δ , provided the growth function is known. In practice this is not usually the case and we need other methods to ensure successful learning: Validation!

Bibliography

- [1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012.
- [2] Neena Damodaran, Elis Haruni, Muyassar Kokhkhovova, and Jörg Schäfer. Device free human activity and fall recognition using wifi channel state information (CSI). *CCF Transactions on Pervasive Computing and Interaction*, 2(1):1–17, 1 2020. Best Paper Award 2021.
- [3] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [4] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.
- [5] Jörg Schäfer. Practical Concerns of Implementing Machine Learning Algorithms for W-LAN Location Fingerprinting. In *Proceedings of ICUMT 2014, 6th International Congress on Ultra Modern Telecommunications, St. Petersburg*, pages 410–417. IEEE Computer Society, 10 2014.
- [6] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.
- [7] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.