Machine Learning and its applications in natural sciences

Dr. Björn Brauer

Organizational

BBB link: https://webroom.hrz.tu-chemnitz.de/gl/bja-a0q-wat-bkr

Lecture name: Aspekte der modernen Physik: Machine Learning and ist applications in natural sciences (212000-720, 212000-721)

Lecture and exercise class: Every Wednesday 13:45 – 17:00

Exam prerequisites:

- Submit all your homework on time
- Finish all project milestones on time and hand in a research paper + code of your course project:
  - Pick a natural science topic that you like and apply ML techniques to it
  - Work on the project on your own or in a group of 2
- Sign up through SB-service by December or hand in form to ZPA (Prüfungsamt), select examiner: B. Brauer
  https://www.tu-chemnitz.de/zpa/formulare/allgemeineformulare/Formular_Pr%C3%BCfungsanmeldung_allgemein.pdf

Exam:

- Content: lecture, exercises, and homework problems (no programming questions)
- First question will be dedicated to your course project so that you have a good start into the exam
- *Bonus point system: collected through active participation in lecture and exercises*

Why should you take this class?

- To receive 5 credit points (approved for physics, chemistry, and afm students; others are based on request)
- To learn some basics of python programming
- To understand the algorithms behind AI applications and know when to use what algorithm
- To have better job perspectives
- To apply AI in your own academic and potential industrial projects and maybe publish or patent the results
- To maybe find a topic/interesting sub task for your thesis

Bonus point system

You can collect up to 5% Bonus Points for your final exam by:

- Asking questions during the lecture and exercise session – I would really appreciate if you turn on your video while you do it
- Answer quiz questions
- Help others with their questions and struggles in class, e.g. by answering their posted questions in OPAL
- Provide feedback to the lecturer: no matter if positive or negative

Actively participate and show your presence in class.

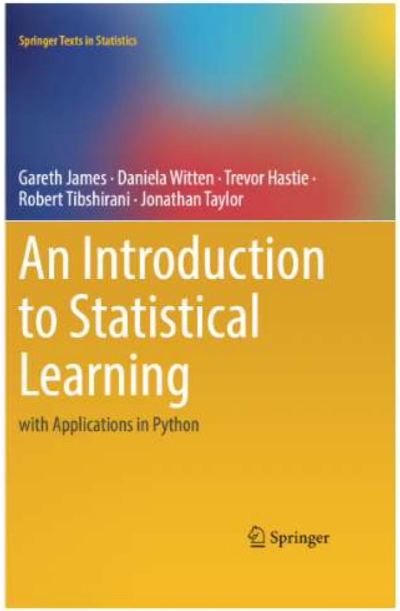
Course literature

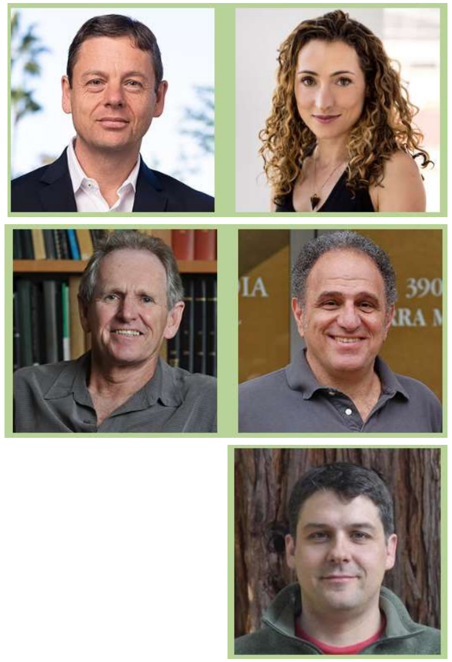
In our lectures we will follow the book:

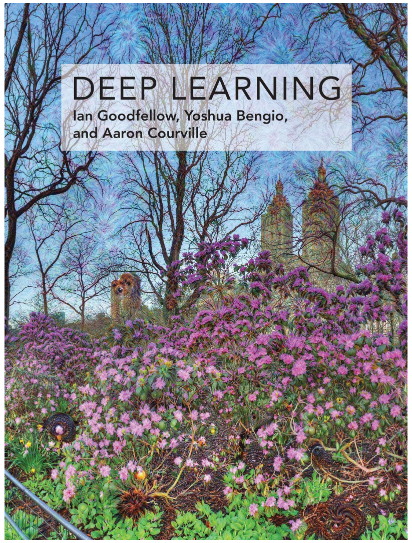
An introduction to statistical learning with applications in Python by G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, Springer 2023. (ISLP)

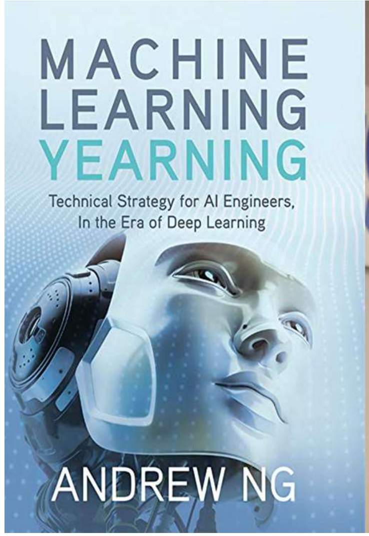
A free PDF version of the book can be found in Week01 folder in OPAL.

Other helpful literature which we will partially use:


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer






MACHINE LEARNING YEARNING
Technical Strategy for AI Engineers, In the Era of Deep Learning
ANDREW NG

Course schedule

Week	Lecture Topic	Reading, homework, and project dates
10/15/25	Introduction to Machine Learning, Notation	
10/22/25	Bias, Variance, kNN, Python and GPT introduction	Read chapters 1,2
10/29/25	Linear regression	Homework 1 due by end of 10/28/25, submit topic of course project by end of 10/29/25
11/05/25	Naïve Bayes, Logistic regression	Read chapters 3,4
11/12/25	LDA, QDA, Classification algorithm performance evaluation	Homework 2 due by end of 11/11/25, project proposal due by end of 11/12/25
11/26/25	SVM, Subset Selection, Regularization, Dimensionality reduction	Read chapters 5,6,9
12/03/25	Resampling (CV, Bootstrap), Decision trees	Homework 3 due by end of 12/02/25
12/10/25	Unsupervised learning, Reinforcement Learning	Read chapters 8,12
12/17/25	Non-linear models	Homework 4 due by end of 12/16/25
01/07/26	Neural Networks I	Read chapter 7
01/14/26	Neural Networks II	Homework 5 due by end of 01/13/26
01/21/26	Convolutional Neural Networks I	Read chapter 10 Research paper draft due by end of 01/21/26
01/28/26	Convolutional Neural Networks II	
02/04/26	Generative models, GPT	Final research paper due by end of 02/04/26
March 2026	Final Oral Exam	

The course project

- Solve a problem of your choice in the field of Natural Sciences through ML
- If you are already working on a research project or you know the direction you want to go, e.g. for your bachelor/master/PhD thesis, then just take it
- You don’t have to write a computer program yourself but you need to try to understand it. You can use one from the course or from the internet and modify parameters
- There are project milestones which make sure you stay on track
- You can reach out to the instructor to discuss your ideas and issues
- In the end you hand in a small research paper (2 – 4 pages) and the code you used
- The project is meant to be interesting, fun to work on, benefits you in many aspects, should take away any fear you might have to use ML. If you plan it well it should not take more than one hour per week
- You can work in groups of up to 2 persons

The course project

Some links to get some ideas:

https://www.tensorflow.org/tutorials/

https://colab.research.google.com/github/keras-team/keras-io/

https://www.kaggle.com/

https://data-flair.training/blogs/machine-learning-datasets/

https://b2find.eudat.eu/dataset

Feel free to get inspired here by typing in key words like “material science“ or „physics“ or „chemistry“ … https://paperswithcode.com/

Make use of GPTs (learn more about those later)

Submission of project topic, proposal, and reasearch paper

Please send to

To: Björn Brauer bjoern.brauer@physik.tu-chemnitz.de

Subject: Course project – *Your Name* (and *Name of Teammate*)

CC: (*Teammate e-mail address*)

Body: *Title of the project*

Some research projects from the last years

- ML-based SEM image analysis for evaluating the spray parameter dependent melting behavior of thermally sprayed particles
- Height profile analysis of Bi islands deposited on Epitaxial Graphene
- Classification and interpretation of UV-Vis spectra of dye solutions
- Scanning Electron Microscopy image colorization
- Crack detection in porous materials
- Star classification
- Architectural style detection of buildings
- Time series forecasting of energy consumption
- Apple sorting via image-based deep learning
- Material recognition in SEM images
- Enhancing high energy ball milling by machine learning
- Machine learning-based regression analysis of DFT calculated Mg/Si/Zr/Sc alloy diffusion on Al(111) crystal surface

The proposal of the course project

1) Problem specification/Introduction

2) What is the problem that you will be investigating? Why is it interesting?

3) What data will you use? If you are collecting new datasets, how do you plan to collect them?

4) What method or algorithm are you proposing? If there are existing implementations, will you use them and how? How do you plan to improve or modify such implementations?

5) What reading will you examine to provide context and background?

6) How will you evaluate your results? Qualitatively, what kind of results do you expect (e.g. plots or figures)? Quantitatively, what kind of analysis will you use to evaluate and/or compare your results (e.g. what performance metrics or statistical tests)?

Writing a research paper on your course project

- Title, your Name
- Abstract
- Introduction
- Background and related work
- Approach
- Experiments
- Conclusion
- References
- Supplementary Material

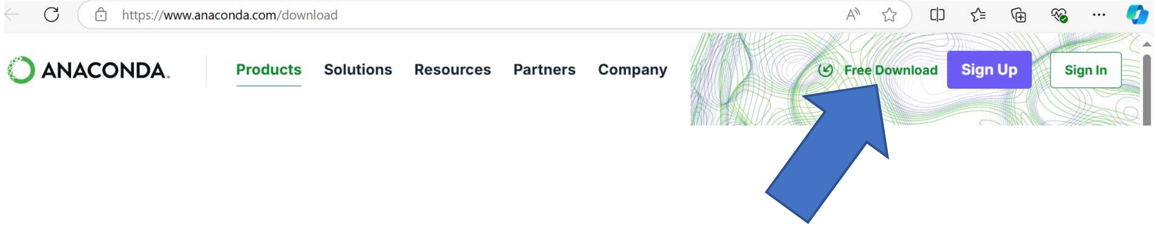
Find more details in week10 folder on OPAL

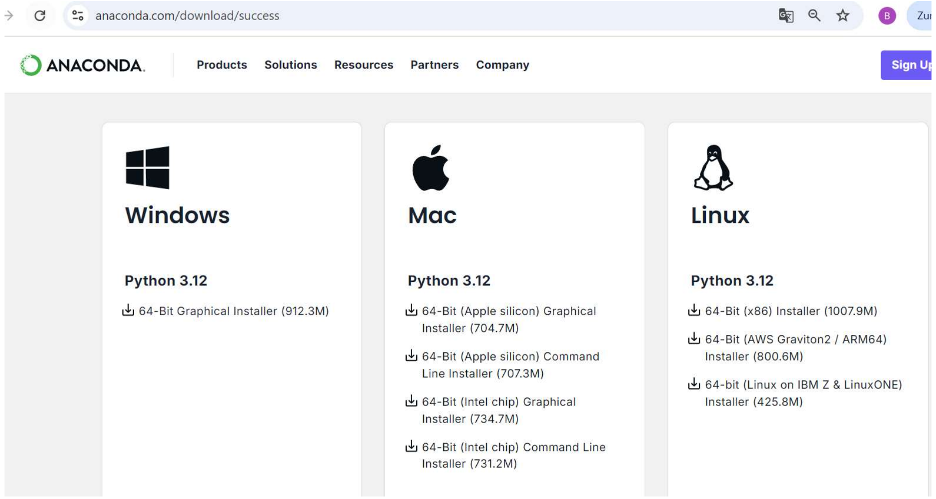
Download the following Software for the class

1) Please install Jupyter Notebook:
https://www.anaconda.com/products/individual
Click on download
Click on .exe file in download folder and go through the installation accepting the default settings
Type "cmd" in Search window and click on Anaconda Prompt, then type:

```
conda install -c anaconda jupyter
```

once installation is done you can type the following in anaconda prompt: jupyter notebook





Windows

Python 3.12

64-Bit Graphical Installer (912.3M)

Mac

Python 3.12

64-Bit (Apple silicon) Graphical Installer (704.7M)

64-Bit (Apple silicon) Command Line Installer (707.3M)

64-Bit (Intel chip) Graphical Installer (734.7M)

64-Bit (Intel chip) Command Line Installer (731.2M)

Linux

Python 3.12

64-Bit (x86) Installer (1007.9M)

64-Bit (AWS Graviton2 / ARM64) Installer (800.6M)

64-bit (Linux on IBM Z & LinuxONE) Installer (425.8M)

2) Please install orange:
https://orangedatamining.com/download/
Select your operating system, download orange, and install it

Working under Linux

Running Jupyter Notebook in Linux in TUC system:
Option 1:
Start Linux-Terminal
Type the following line and press enter afterwards:

```
/afs/tu-chemnitz.de/global/capp/anaconda/bin/jupyter-notebook &
```

Option 2:
Start Linux-Terminal
Type the following lines and press enter after each line:

```
/afs/tu-chemnitz.de/global/capp/anaconda/etc/profile.d/conda.sh && conda activate base
jupyter-notebook &
```

Creating a virtual environment (optional)

Please make sure to use the following procedure to create a virtual Environment. If you don‘t do this, you will face the problem that some Notebooks in this class won‘t run on your PC as the library versions don‘t match.

1) We need to create a virtual environment. Please do the following:
Open the anaconda prompt as an administrator. Navigate with cd … to the path where you want to have your jupyter notebook files later on, e.g.:
`cd \Users\Documents\Notebooks`   now type:

```
pip install virtualenv
```

Now you can create an environment in which you are going to work from now on. I am calling it env1 here but you could name it anything you like. Then type:

```
python -m venv env1
cd env1/Scripts/
activate
```

You are now seeing your path as follows: (env1) (base) C:\Users\... Note the (env1) in the very front. This means that everything you are pip installing will just go into this environment.
Every time you want to run jupyter notebook with the material of this ML class, you should go to this ...env1/Scripts/ folder and type activate

2) Go to Window explorer and copy the file requirements.txt (Week01 folder in Opal) into the folder …/env1/Scripts/ and then type:

```
pip install -r requirements.txt
```

after this installation is done, please type: `jupyter notebook`

Introduction to Machine Learning

Definitions

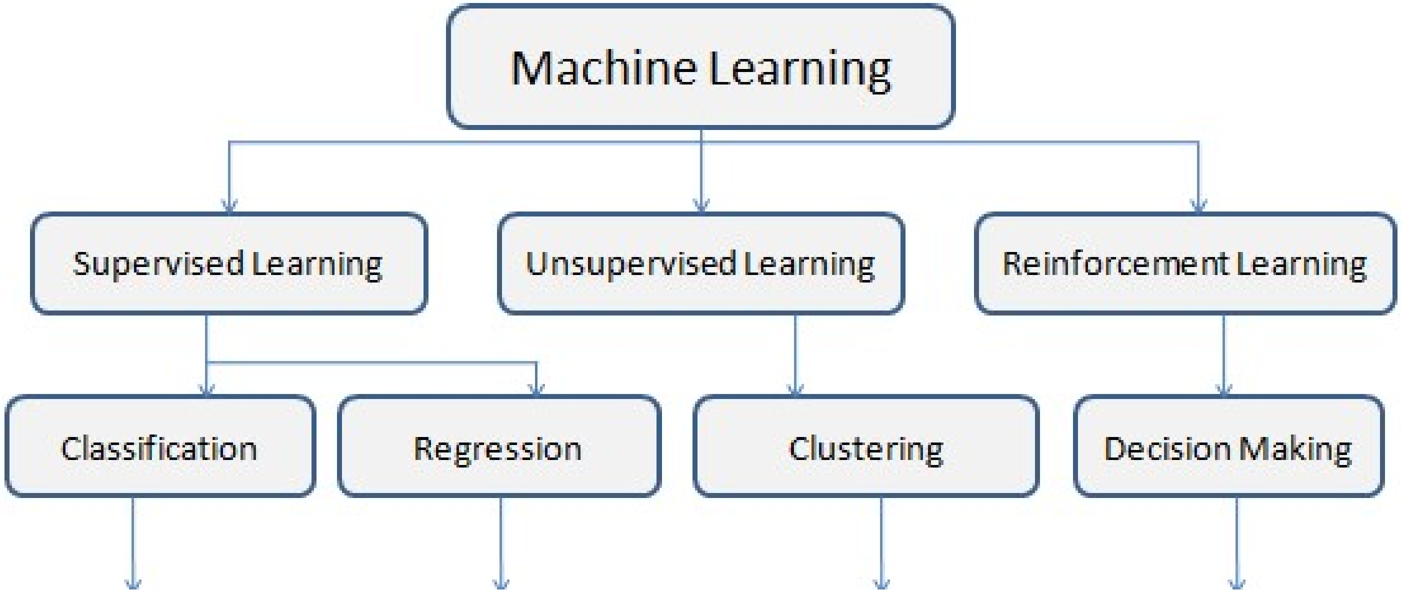
Artificial Intelligence (AI) refers to computer systems which perform tasks that typically require human intelligence, such as speech recognition, decision making, and natural language processing.

Machine Learning (ML)/Statistical Learning (SL) is a subset of AI. It uses statistical techniques to identify patterns in data during a training process to make predictions without being explicitly programmed. Learning through experience.

Since the overlap of ML and SL is very big we are going to use those as synonyms in this lecture. Main difference is that ML has the upper hand in marketing.

Deep learning (DL) is a subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data

Possible categorization of Machine Learning



Naive Bayes
Decision Trees
Support Vector M.
Random Forest
K-nearest neighbor

Linear regression
Regression trees
Lasso regression
Ridge regression

K-Means clustering
Hierarchical clustering
Gaussian mixture models
Principal component analysis

Value iteration
Q-Learning
TD-Learning

***There is no free lunch theorem:** Just because there is a certain tool to solve one learning problem, doesn‘t mean that it can solve a different problem as well.*
This is why we need to know about many algorithms.

Notation

+ Y is the outcome or measurement, response, target, dependent variables
+ X is a one dimensional vector, consists of lots of input such as x1...xp, vector of p predictor measurement - so-called features, or predictor, inputs, attribute, or observation, properties, regressors, independent variable

example: X = (melting point, boiling point, magnetism, electron config, proton count)
=(1538, 2862, ferromagnetic, [Ag] 3d64s2, 26)
thus p =5 (5 features) => Y = Fe

Quiz question

In the expression specific heat = f(heat change, temperature change, mass), "specific heat" is the:

a) Training Data

b) Independent Variable

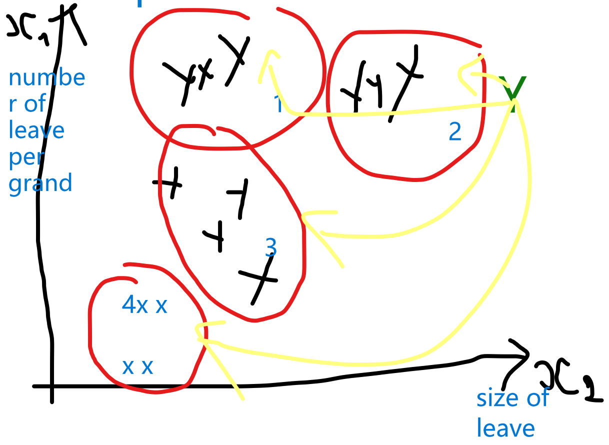
c) Response

d) Feature

Classification vs. Regression

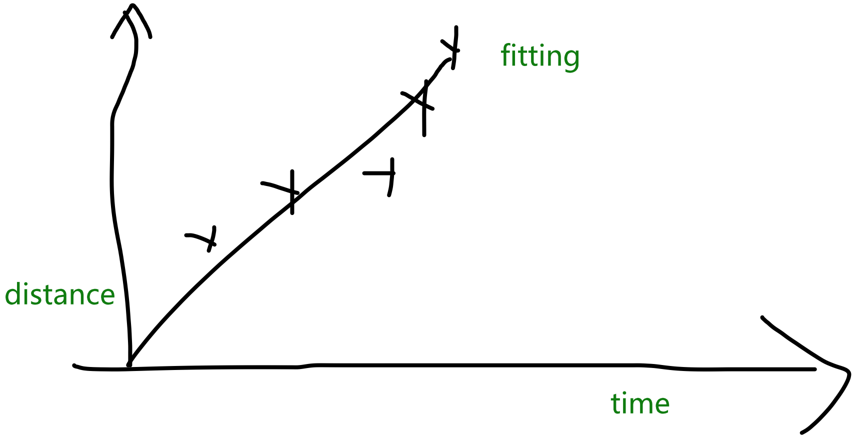
Classification problems: Y takes values in finite (restricted) unordered set

- example: survive/ dead, cancer class of tissue sample, ...
- Graphical illustration:



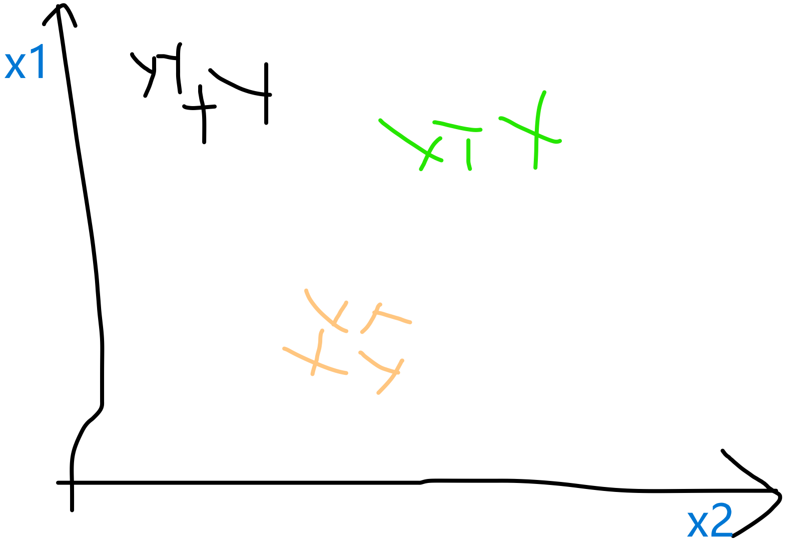
Regression problems: Y is quantitative

- Example: housing price, blood pressure, Temperature, distance..
=> often call is curve fitting



Supervised vs. Unsupervised Learning

Supervised learning: for every set of observati
the respective class is known



Unsupervised learning: For every set of observation the class is unknown

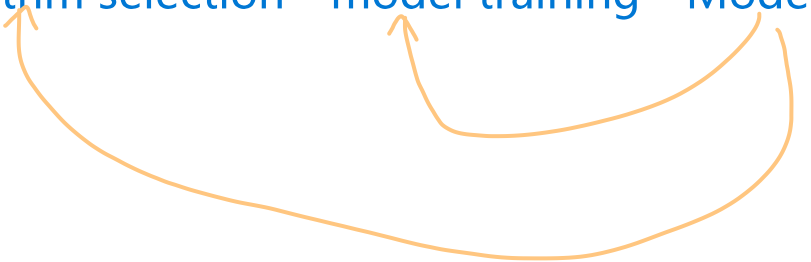
often use as pre-processing for the supervised learning

Programming vs. machine learning

While programming aims to answer a problem using predefined set of rules or logic

Example: if … then…

Machine learning: seeks to construct a model or logic for the problem by analysing the input data
ML pipeline:
- Task description - Data acquisition - Algorithm selection - model training - Model validation - prediction or inference

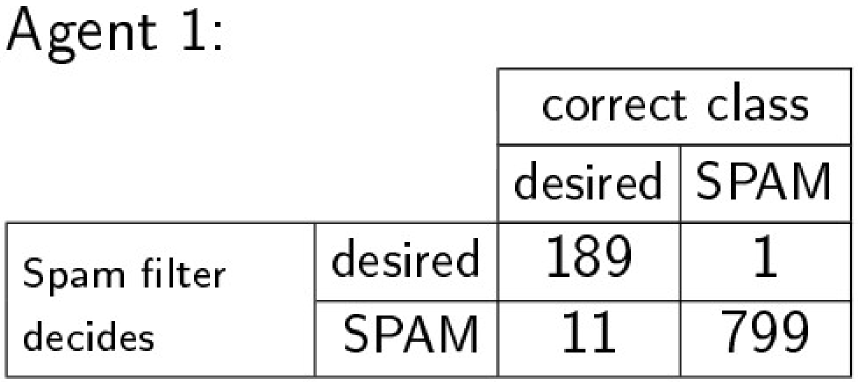


possibly a new ru

prediction: accurately predict unseen test cases and dont wanna know the exact form of the estimator function and treated as a blackbox, then it is prediction

Quiz question

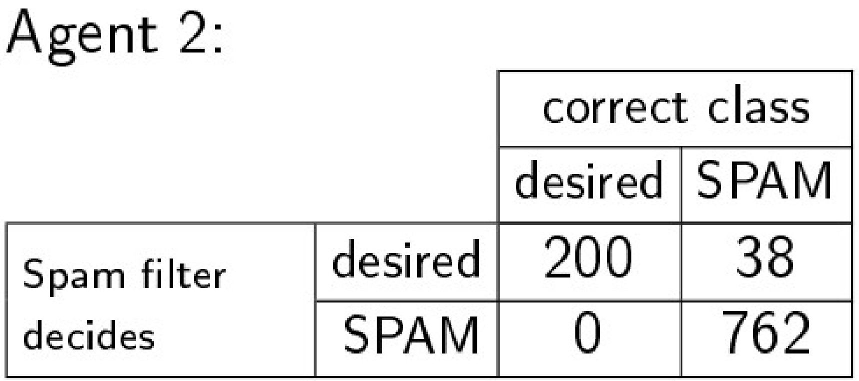
Imagine you have trained two agents to filter out spam e-mails from your inbox. The performance of the two agents is as follows. Which agent would you prefere as your spam filter?



Agent 1:

		correct class	
		desired	SPAM
Spam filter decides	desired	189	1
	SPAM	11	799

94,5% 99.8%



Agent 2:

		correct class	
		desired	SPAM
Spam filter decides	desired	200	38
	SPAM	0	762

100 95,25

Prediction vs. Inference

Prediction: accurately predict unseen test cases and don't wanna know the exact form of the estimator function and treat black-box, then it is prediction.

Inference: understand which input effects the outcome, and how it is impacted, want to understand the relation between Y. How does Y changes as a function of X (X1, X2, X3, ...Xp). which predictor associated with the response. Want to know e equation (Cannot treated as a blackbox)

Quiz question

Which of the following are supervised learning problems? Y: Y/N

[✓ a)] Predict whether a website user will click on an ad Y: Unknow

b) Find clusters in genes that interact with each other

[✓ c)] Classify handwritten digits as 0-9 from labeled examples Y: 0-9

d) Find stocks that are likely to rise Y: Rise/not Rise

Quiz question

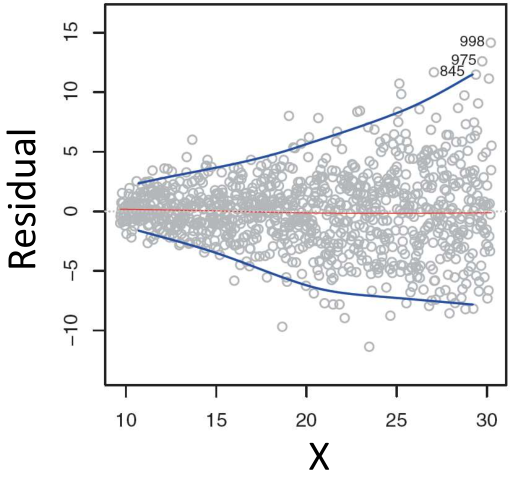
True or False:

The only goal of any supervised learning study is to be able to predict the response very accurately.

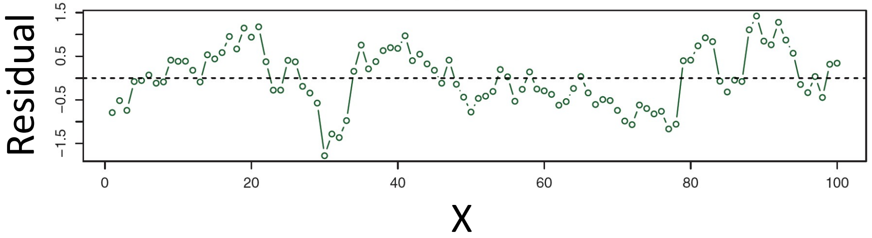
F

Notation and definitions

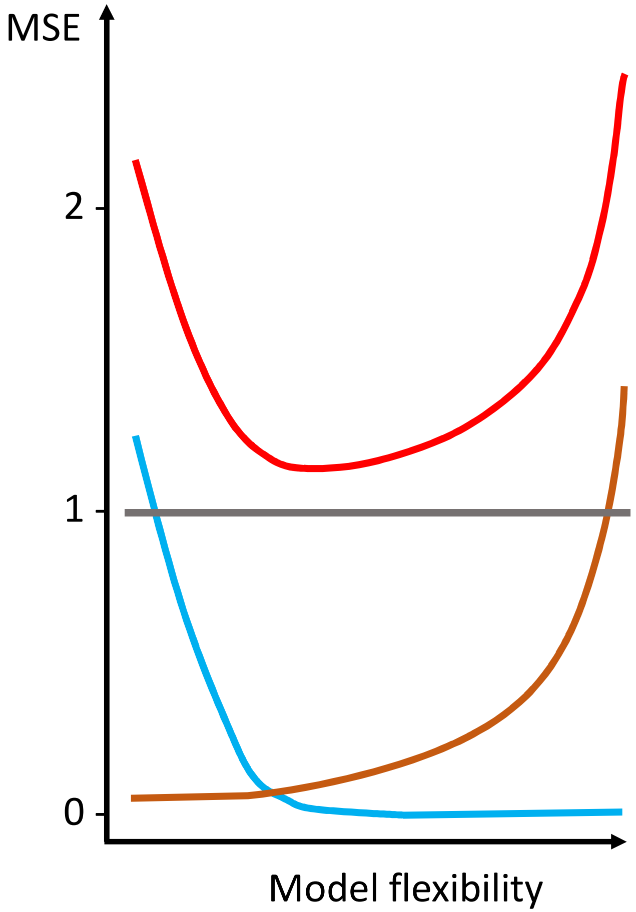
Non-constant variance (heteroscedasticity)



Correlated error terms



The Bias Variance trade-off



Quiz question

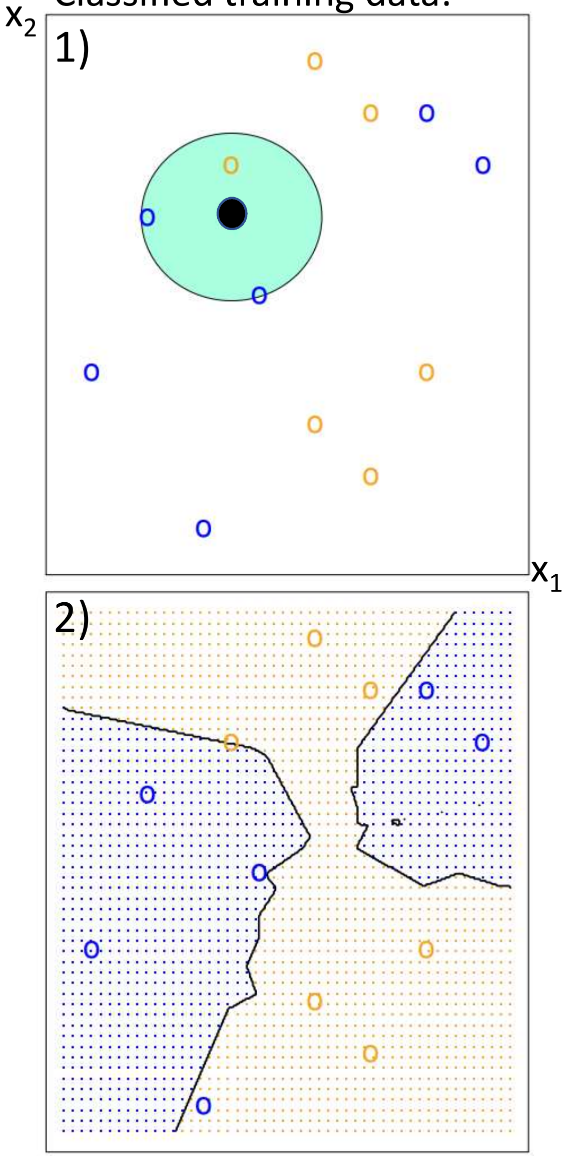
True or False:

A fitted model with more predictors will necessarily have a lower Training Set Error than a model with fewer predictors.

F

k-Nearest Neighbor: kNN

Classified training data:

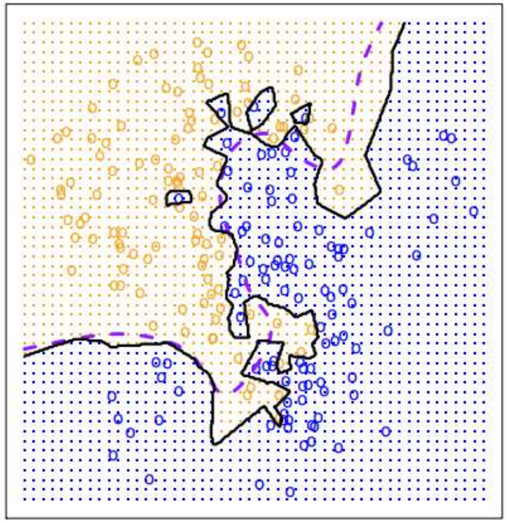


Figures from: An introduction to statistical learning, Hastie et al., 2021.

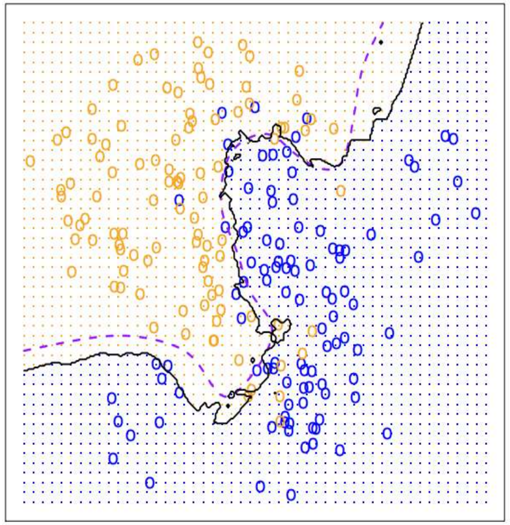
Quiz question

A nearest neighbor classifier was used here. What answer is correct?

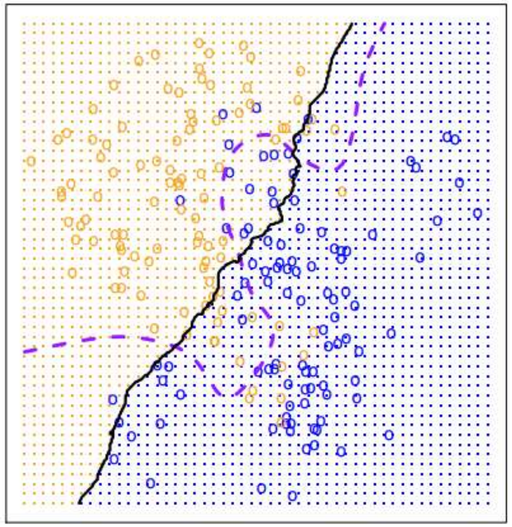
1)



2)



3)

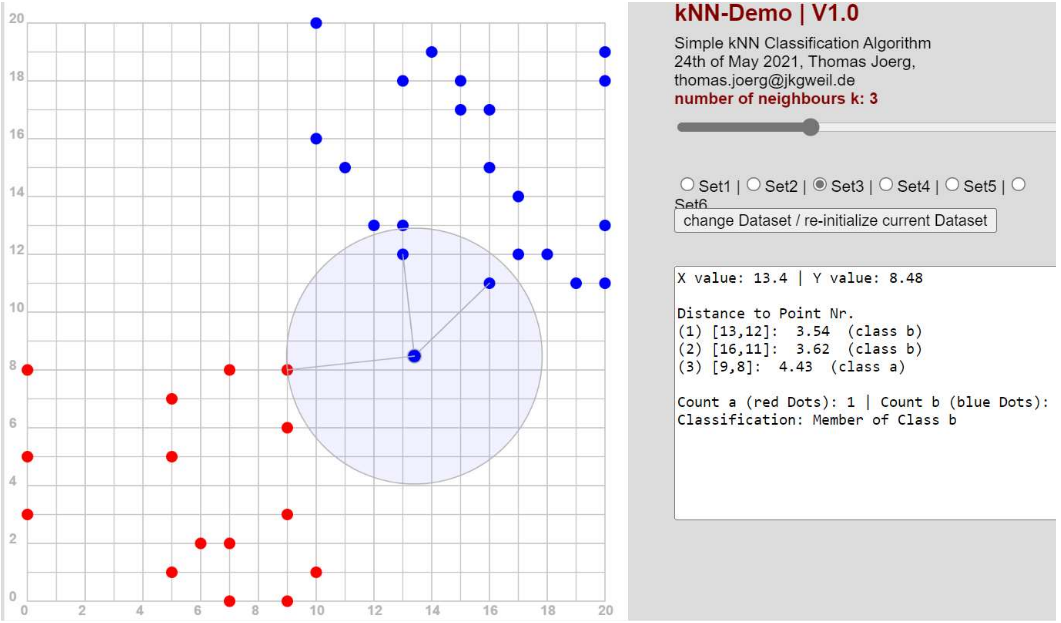


a)	$k_1 = 1$,	$k_2 = 100$,	$k_3 = 10$
b)	$k_1 = 10$,	$k_2 = 100$,	$k_3 = 1$
c)	$k_1 = 1$,	$k_2 = 10$,	$k_3 = 100$
d)	$k_1 = 100$,	$k_2 = 10$,	$k_3 = 1$

Figures from: An introduction to statistical learning, Hastie et al., 2021.

kNN simulator and programming exercise

https://iludis.de/kNNDemo/index.html


20
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
16
18
20
○ Set1 | ○ Set2 | ● Set3 | ○ Set4 | ○ Set5 | ○ Set6
change Dataset / re-initialize current Dataset

kNN-Demo | V1.0

Simple kNN Classification Algorithm
24th of May 2021, Thomas Joerg,
thomas.joerg@jkgweil.de
number of neighbours k: 3

```
X value: 13.4 | Y value: 8.48

Distance to Point Nr.
(1) [13,12]:  3.54  (class b)
(2) [16,11]:  3.62  (class b)
(3) [9,8]:  4.43  (class a)

Count a (red Dots): 1 | Count b (blue Dots):
Classification: Member of Class b
```

Run the kNN Jupyter Notebook in this weeks folder