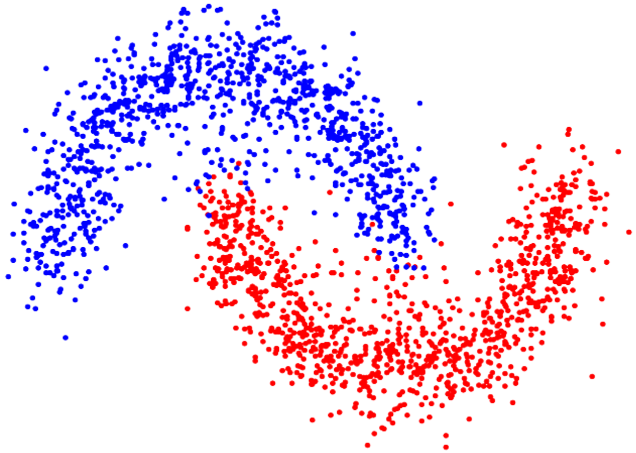
R-course:
Machine Learning using R

Course organization and introduction



Yannick Rothacher

Zürich, 2021

Course material

- Where to find it…
  - Print out
  - USB stick
  - Github

Who am I?

- **Yannick Rothacher**
- Original background: Biology/Neuroscience (PhD in Neuroscience)
- Further education in "applied statistics" at ETH Zürich
- Currently working as a Post-Doc at the Professorship for Psychological Methods, Evaluation and Statistics (Prof. Carolin Strobl)
  - Doing research on Random Forests and interpretable machine learning
  - Teaching introductory courses to machine learning and R
- yannick.rothacher@psychologie.uzh.ch

Who are you?

A list of questions I would be interested in:

- What is your **work**?
- Experience with **R**? How often used?
- Experience in **statistics**?
- Experience in **machine learning**?
- What are your **expectations for this course**?

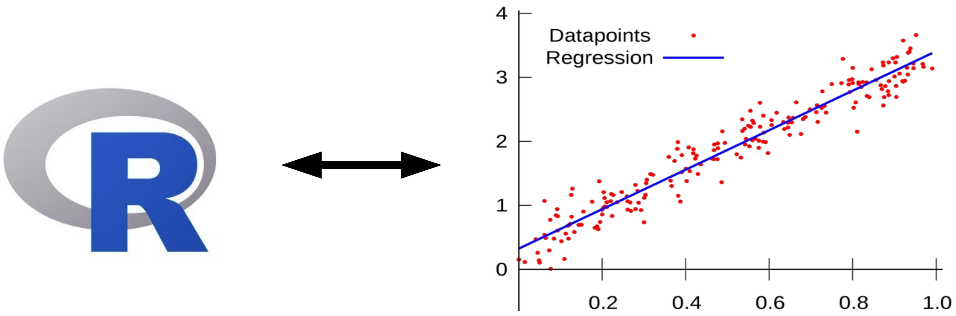
Course goals and organization

Goals:

- Give an overview of different machine learning methods
- Explain working principle of the presented methods
- Practice application of presented methods in R
- Discuss general issues in machine learning

Organization:

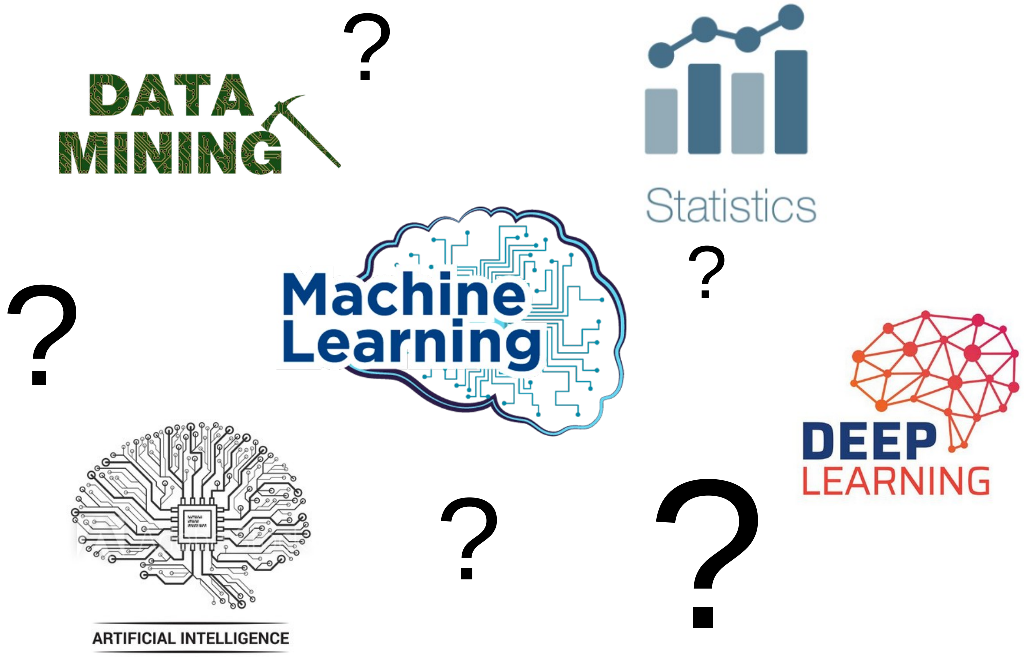
- Two day course
- Alternation between **lectures and exercises**



Course timetable

- See PDF "**RKurs2021_ML_Program.pdf**" ...

What is Machine Learning?



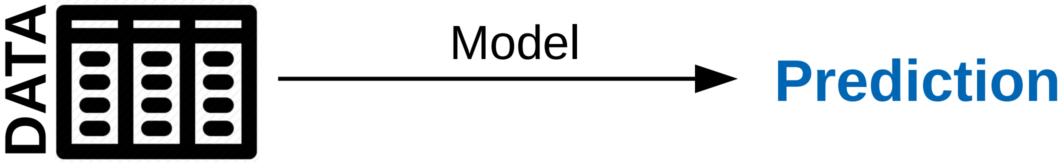
- Distinction from Machine Learning to other statistical methodology not always clear
- When comparing Machine Learning with “classical” statistics:
- Statistical models are generally designed for **inference**
- Machine Learning models are generally designed for **prediction**

Application of Machine Learning

Being able to **predict** certain outcomes based on data can be important in many different areas in **research and industry**

Examples:

- Predict the winner of a basketball game
- Predict the weather of tomorrow
- Predict whether a medical scan shows an image of a tumor
- Predict whether an email is spam or not
- Predict how likely a person is about to develop depression



DATA

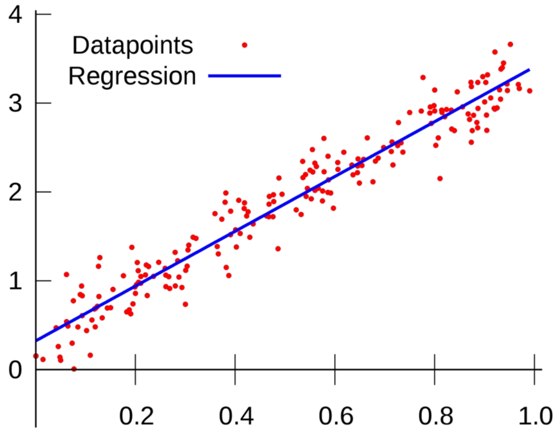
Model

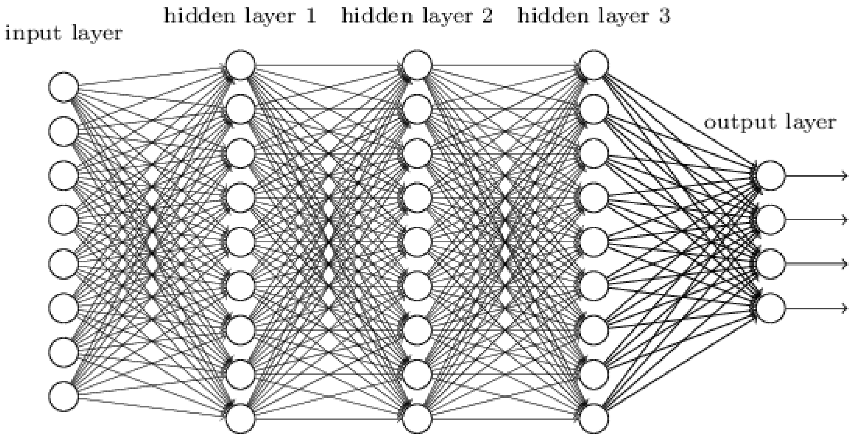
Prediction

In all cases: **Predictions are based on data !**

Prediction models don’t have to be complicated

- Simple linear regression can also be used to predict values of new observations





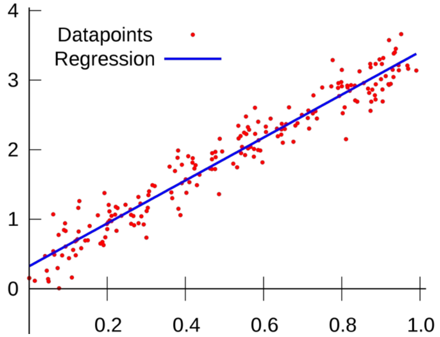
- However, sometimes statistical models have limited prediction accuracy, but allow **inference about the relation** between predictors and target variables (e.g. showing a significant influence of a treatment).
- In many Machine Learning models, the prediction accuracy is very good but it is difficult to infer the variables’ relations (e.g. neural network)

Application of Machine Learning

- Again: In general one tries to predict a target variable based on predictor variables

target variable ~ predictor variables

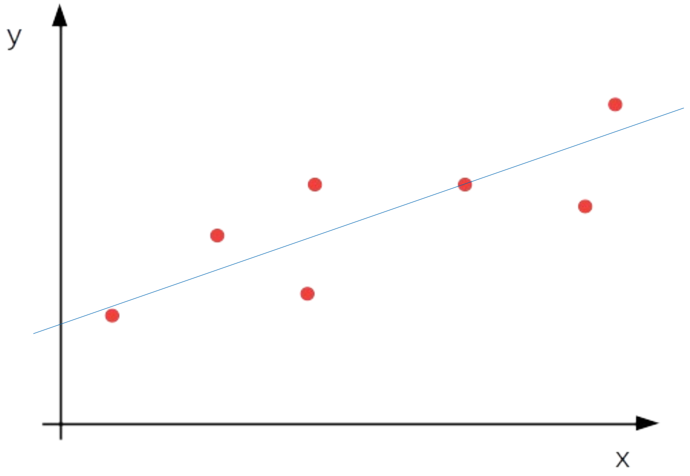
y ~ X



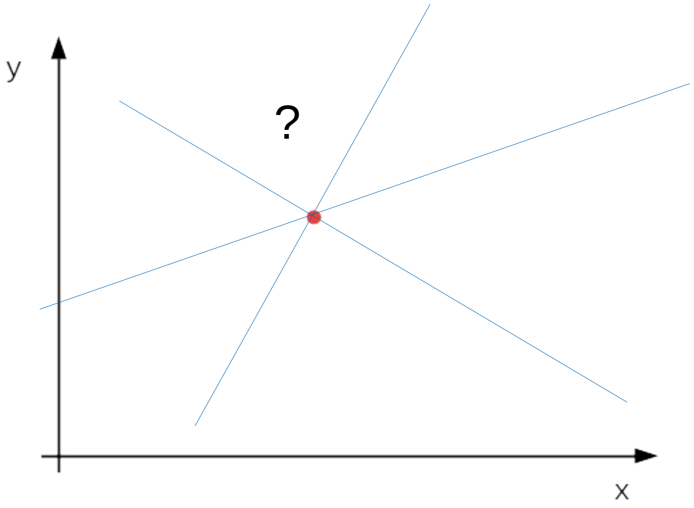
- Target variable can be a certain category, a number, a probability, ...
- In real-life data, there are often many predictor variables (genetic data: up to 10’000 predictors)
- Can even be n << p (much more variables (p) than data points (n))
- This case can be difficult to handle with conventional methods (for example linear regression)

Challenges of high-dimensional data

- For example linear regression only works for n > p :



n > p

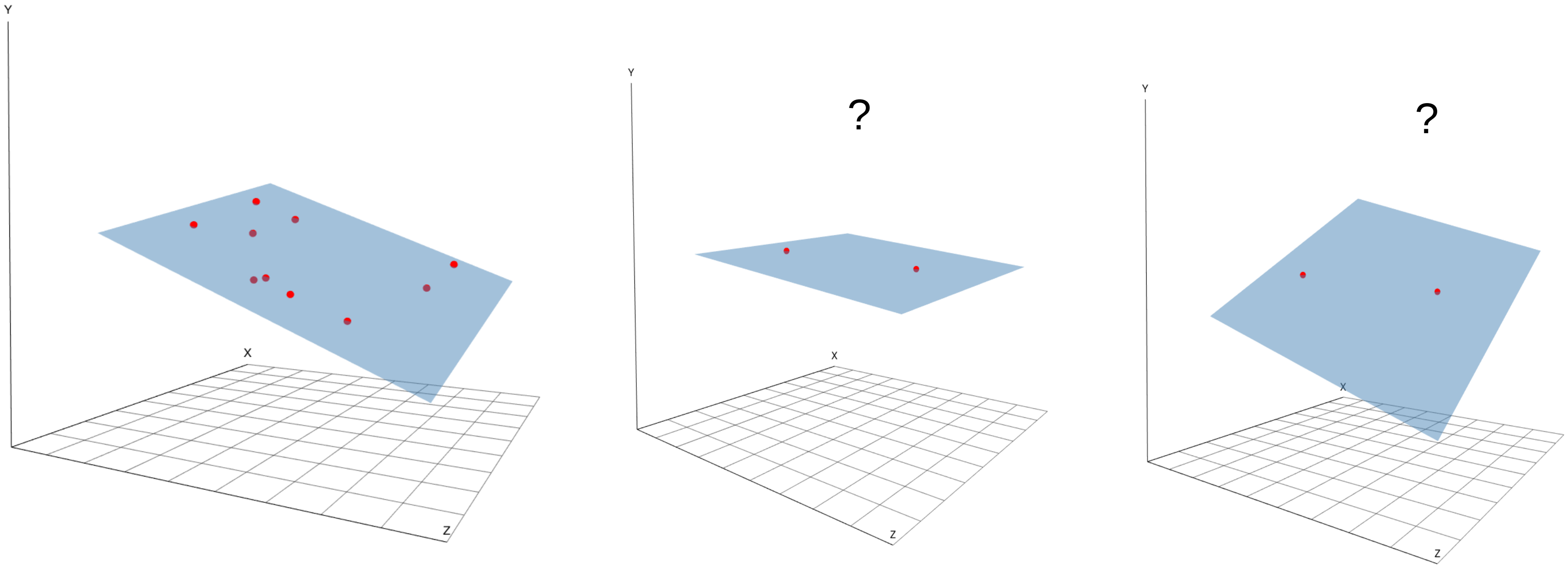


n = 1

- We need methods for situations with n < p
- Machine Learning methods are usually able to handle n < p situations

Challenges of high-dimensional data

- For example linear regression only works for n > p :

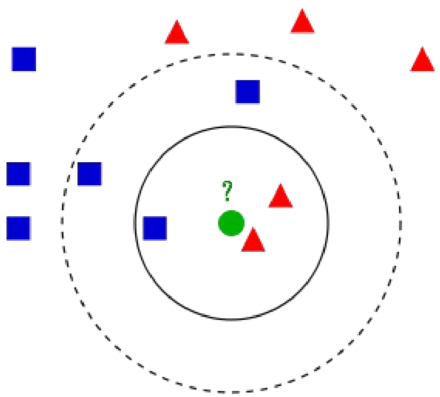


n > p

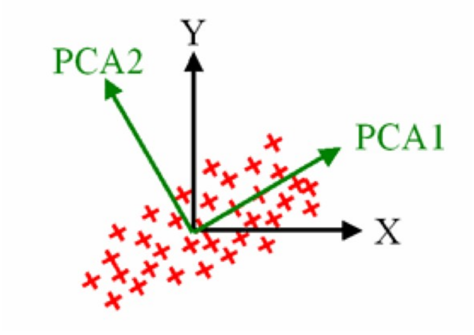
n = 2

- We need methods for situations with n < p
- Machine Learning methods are usually able to handle n < p situations

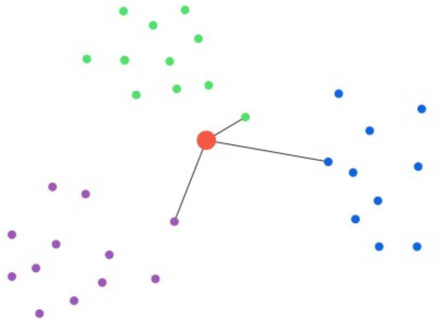
Outlook: Machine Learning methods



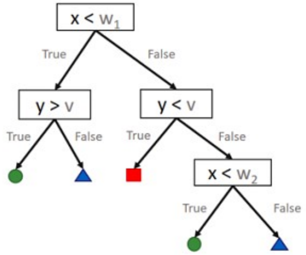
K-nearest neighbor



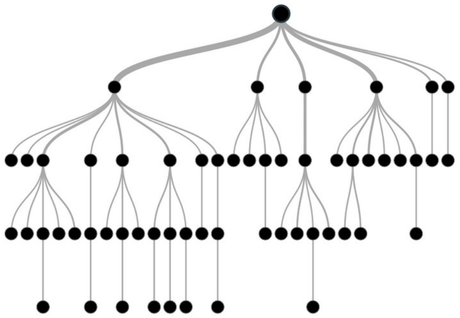
Principal Component Analysis



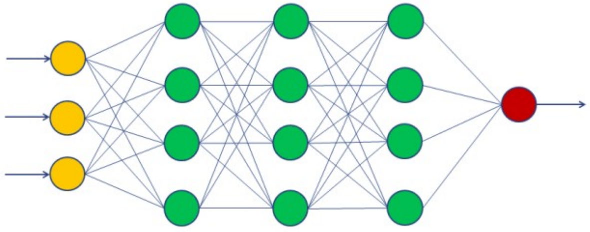
K-means clustering



Decision trees



Random Forest



Neural networks