

# Machine Learning Course Workbook

– Day 1 –

## The Basics

### **Data is the new oil!?**

*What does Goodhart's Law warn us about?*

*With what KPI could your department's performance be quantified? What would be the target state, alert threshold, and what could be possible corrective actions?*

### **What is ML?**

*What is the difference between Machine Learning, Artificial Intelligence, and Deep Learning?*

*What are the benefits of ML compared to traditional software?*

*When should you not use ML?*

*Which kind of ML problems have a high chance of success and when is the outcome uncertain?*

### **How do machines “learn”?**

*Describe the different learning strategies and what their requirements (in terms of data) are:*

- Unsupervised Learning:
- Supervised Learning:
- Reinforcement Learning:

*What are “features” and what are “labels”?*

- Features:
- Labels:

*What is the goal of a supervised learning algorithm and how is it accomplished?*

## ML use cases

*What does structured and unstructured data look like? Which of them is homogeneous and which (usually) heterogeneous?*

- Structured Data:
- Unstructured Data:

*Take another look at the [ML algorithm cheat sheet](#) & try to find examples where you could be using each of these algorithms to improve your organization's products or processes.*

- Anomaly Detection:
- Clustering:
- Regression:
- Classification:

*What is the drawback of unsupervised learning methods?*

*What are the benefits of breaking down a complex input-output problem into simpler subproblems?*

*What is the downside of a system composed of multiple ML models?*

## Solving problems with ML

*What are the two deployment options for an ML model and when should you use which?*

## ML with Python

*What are the standard abbreviations used when importing the numpy and pandas libraries?*

```
import numpy as ...  
import pandas as ...
```

## Data Analysis & Preprocessing

### Data Analysis

*You want to pick a restaurant for dinner. Your data source is Google Maps. What information do you consider when making a decision and what makes you choose one restaurant over another?*

### Garbage in, garbage out!

*Think about some of the datasets you've encountered in the past: In what ways were they messy?*

*Which concrete next steps should your organization take to improve their data quality?*

### Data Preprocessing

*What is the difference between feature extraction and feature engineering?*

- Feature Extraction:
- Feature Engineering:

*A feature matrix  $X$  has the shape  $(n \times d)$ . What do  $n$  and  $d$  stand for?*

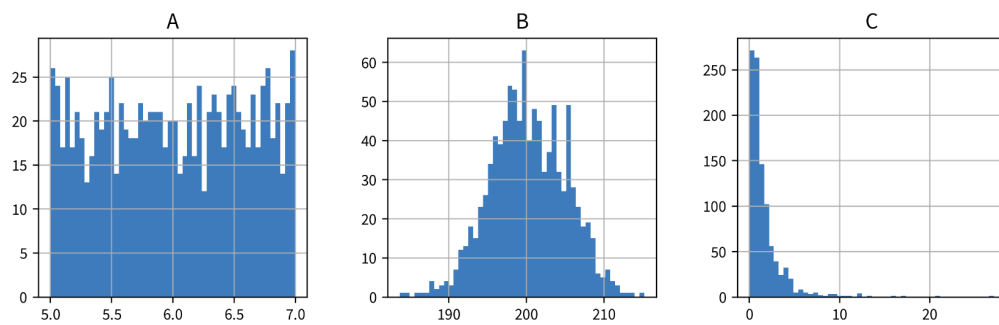
- $n$ : number of ...
- $d$ :

*You are given a dataset with time series data, consisting of measurements from  $d$  sensors for  $n$  time points. What would your feature matrix look like, if your task was...*

- ... to make a prediction for each time point?
- ... to categorize the different sensors?
- ... to predict the quality of each of the 100 products produced during this time span?

*What is one way to transform categorical features into a meaningful numerical representation?*

These are the histograms of three different variables A, B, and C:



How would you characterize their distributions (Gaussian, exponential, uniform) and which kind of transformation (StandardScaler, MinMaxScaler, PowerTransformer) might be best suited for which of the variables?

- A:
- B:
- C:

What preprocessing steps can be helpful to compute a more meaningful similarity or distance between the data points' feature vectors (especially for heterogeneous data)?

- 
- 

## Supervised Learning Basics

### Different types of models

What is the difference between a regression and a classification problem?

When should you use a features-based and when a similarity-based model and what are their respective drawbacks?

### Model Evaluation

With which stupid baseline should you compare regression and classification models respectively?

When is it a really bad idea to evaluate a classification model with the accuracy metric?

How does a cross-validation work? What are the advantages and disadvantages compared to using a fixed validation set?

– Day 2 (Part 2) –

## Supervised Learning Models

### Linear Models

*How does a linear model compute the prediction for a new data point?*

*What happens when you use a regularized model and set the regularization parameter to a high value (e.g., `alpha` for a linear ridge regression model in `sklearn`)?*

### Neural Networks

*How does a feed forward neural network (FFNN) compute the prediction for a new data point?*

*How could a multi-layer FFNN be simplified, if it did not contain any nonlinear activation functions between its layers?*

*In what way could you manipulate the parameters (i.e., weight matrices) of an existing FFNN without changing its predictions?*

### Decision Trees

*How does a decision tree compute the prediction for a new data point?*

*For a decision tree with `max_depth=2`, how many different features can be used at most for the prediction?*

### Ensemble Methods

*What are the different strategies for creating ensemble models?*

*How does a random forest compute the prediction for a new data point?*

### **k-Nearest Neighbors (kNN)**

*How does a kNN model compute the prediction for a new data point?*

*Why is it better to use an odd number of nearest neighbors for kNN for a binary classification problem?*

### **Kernel Methods**

*How does a kernel ridge regression (KRR) model compute the prediction for a new data point?*

*Why is it more efficient to compute the prediction for a new data point using a support vector machine (SVM) model compared to KRR?*

– Day 3 –

## Avoiding Common Pitfalls

*What is the difference between data and concept drift?*

*What could be reasons for data or concept drift in your domain / next project?*

### **Model does not generalize**

*How can you tell whether a model underfits the data and what can you do to improve the model's performance if this is the case?*

*How can you tell whether a model overfits the data and what can you do to improve the model's performance if this is the case?*

*Why can the performance on the training set get worse as the size of the training set increases?*

*Why should you not use a univariate feature selection approach? What are better alternatives?*

### **Model abuses spurious correlations**

*Why can a model still be wrong, even though it generates correct predictions for data points from the testset?*

*What are "Adversarial Attacks"?*

### **Model discriminates**

*Why can it happen that a model discriminates and in what ways could this negatively affect users?*

*How can you check whether a model discriminates?*

## **Explainability & Interpretable ML**

*What is the difference between local and global explainability?*

*How can you explain an individual prediction of ...*

- a decision tree?
- a linear model?
- a neural network?

*How can you identify the features that are overall the most important for a model?*

*How can you determine (approximately) how an individual feature influences the model prediction overall?*

*What model-agnostic approach can you use to explain an individual prediction of any model?*