# Machine Learning Course Workbook

## Introduction

**ML is everywhere!**

*Where (else) do you use ML in your everyday life incl. work?*

- 

**ML history: Why now?**

*What accelerated the rise of ML in the last few years?*

- 
- 
- 

*What is the difference between ANI and AGI?*

– Part 1 –

## The Basics

**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 <u>not</u> 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.*
- Dimensionality Reduction:
- Anomaly Detection:
- Clustering:
- Regression:
- Classification:
- Generative AI:
- Recommender Systems/Information Retrieval:
- Reinforcement Learning:

*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

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

*With what KPI could your department's goal be quantified? Is this a leading or lagging KPI?*

## 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 x 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:*
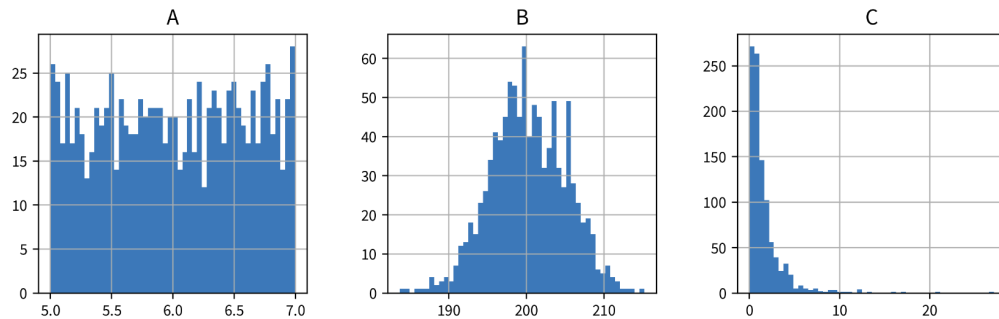


*How would you characterize their distributions (Gaussian, exponential, uniform) and which kind of transformation (StandardScaler, MinMaxScaler, PowerTransformer) might be best suited for which of the variables?*

➔ 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)?*

-

-

*What are the steps to transform a corpus (i.e., dataset with text documents) into a TF-IDF feature matrix?*
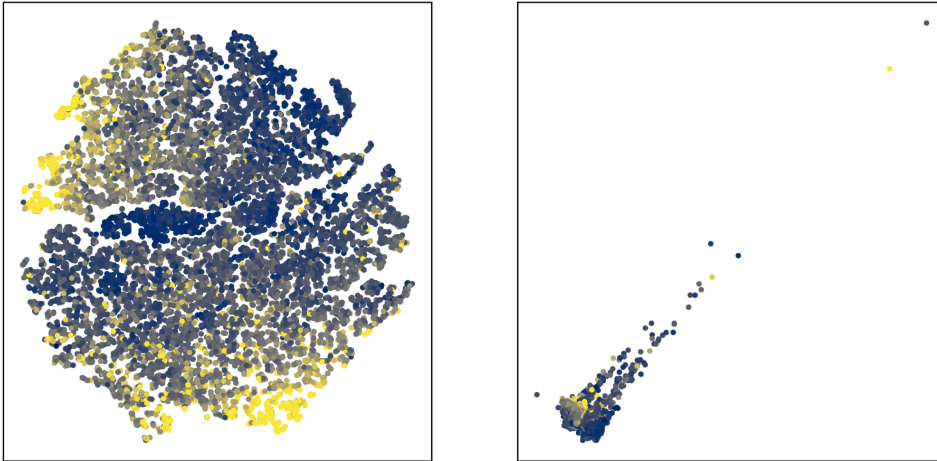
*What are the disadvantages of TF-IDF feature vectors?*

-

-

## Unsupervised Learning

### Dimensionality Reduction

*Guess: Which plot was generated with PCA and which with t-SNE?*



*How does PCA work?*

*Is PCA using the original input features for the computation or does it first compute a similarity matrix for the data points? What about Kernel PCA?*

*How does t-SNE work?*

*Is t-SNE using the original input features for the computation or does it first compute a similarity matrix for the data points?*

*When would you use PCA and when would you use t-SNE?*

*In the notebook, what did you observe about the PCA eigenvalue spectrum for the data with and without outliers? How do you interpret this?*

**Outlier/Anomaly Detection**

*What factors should you consider when choosing an outlier detection method?*

*Why can a data point still be an outlier even if it is within normal ranges of the individual features?*

*How does the $\gamma$-index work and how could you set the parameter k if you suspect clusters of outliers?*

**Clustering**

*How does the k-means algorithm work?*

*True or False: One disadvantage of k-means is that it assumes spherical clusters?*

*How does the DBSCAN algorithm work?*

*What are the advantages of DBSCAN?*

# Supervised Learning Basics

**Different types of models**

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

*How can you tell if a classification or regression dataset is linear or nonlinear (e.g., with one input x)?*

*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?*

*Name three regression evaluation metrics:*

- 
- 
- 

*Name two classification evaluation metrics:*

- 
- 

*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?*

## 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?*

## Avoiding Common Pitfalls

### 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?*

*Why can it hurt the performance if you (aggressively) reduce the dimensionality of the data with PCA?*

**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?*

*What can you do to get a fair model?*

**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?*

**Data & concept drifts**

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

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

## Advanced Topics

### Deep Learning

*What type of neural network architecture would be a natural choice for sequential data like text or time series data? What about image data?*

*How does self-supervised learning work (e.g., using text data)?*

*How does transfer learning work and when can it help?*

### Information Retrieval (Similarity Search)

*What is the most important (and difficult) step when trying to solve an information retrieval task?*

### Recommender Systems (Pairwise Data)

*What kind of problems (in terms of inputs and outputs) can you solve with recommender systems?*

*What is the "cold start problem" and how can you circumvent it?*

### Time Series Forecasting

*What kind of input features could you use to forecast how many pretzels a bakery will sell tomorrow?*

*What is the difference between stateless and stateful models and which conditions need to be fulfilled so it makes sense to use a stateless time series forecasting model?*

### Reinforcement Learning

*For which kinds of tasks does it make sense to use reinforcement learning and when does a normal optimization suffice?*

*How does the Epsilon-Greedy Policy manage the trade-off between exploration and exploitation?*

*What is a Q-value and how does Q-learning for tabular RL work?*

*How can Q-learning be extended to work with an infinite number of states?*

*Which factors can complicate the use of reinforcement learning?*

# Conclusion

*What can you do if you have "big data"?*

### AI Transformation of a Company

*According to Andrew Ng, what are the 5 steps for a successful AI transformation of a company and where do you think your organization stands in this process?*

1.
2.
3.
4.
5.