Machine Learning Course Workbook

– Before the Course –

## Introduction

### ML is everywhere!

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

– Part 1 –

### ML history: Why now?

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



#### What is the difference between ANI and AGI?

### What is ML?

#### Define ML:

#### What do you need to create an ML-powered product (i.e., value)?



#### AI and ML Researchers, Statisticians, and Data Scientists all use a certain set of tools. What is the difference between...

* ML vs. AI?
* ML vs. Deep Learning?
* ML vs. Statistics?
* ML vs. Data Science?

### 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 is the drawback of unsupervised learning methods?

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

### When should you use ML?

#### In what ways can ML create value?

#### When should you not use ML?

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

### Solving problems with ML: Workflow

#### What are the 3 main steps to create value with ML?



#### What should you check before starting an ML project?



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



#### Which tasks take up most of a Data Scientist’s time?

## ML with Python

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

import numpy as ...

import pandas as ...

## Data & Preprocessing

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

* Features:
* Labels:

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

* Structured Data:
* Unstructured Data:

#### 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:*

### What constitutes one data point?

#### 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 is…

* … 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?

### Feature Extraction

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

#### 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?



### Feature Engineering & Transformations

#### 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:

### Computing Similarities

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



### 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?

## ML Solutions: Overview

#### What does the output of the different algorithm categories look like for one data point?

* Dimensionality Reduction:
* Anomaly Detection:
* Clustering:
* Regression:
* Classification:
* Recommender Systems/Information Retrieval:

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

– Part 2 –

## 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?

#### How does the 𝛾-index work?

#### How could you set the parameter k of the 𝛾-index to detect a cluster of outliers?

### Clustering

#### How does the k-means algorithm work?

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

#### K-means: What would happen if you set k to a very large value, e.g., the number of data points?

#### 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?

– Part 3 –

## 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)?

### 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?

## Deep Learning & more

### Information Retrieval (Similarity Search)

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

### Deep Learning (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?

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

#### What type of neural network architecture would be a natural choice for image data?

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

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

### 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?

### Recommender Systems

#### 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?

– Part 4 –

## Avoiding Common Pitfalls

#### What are some common pitfalls that you should avoid?

#### In what ways can domain knowledge help you arrive at a better model?

#### 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 over- or underfits the data?

#### What can you do to improve the prediction performance in case of underfitting?

#### What can you do to improve the prediction performance in case of overfitting?

#### 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

#### In what ways can a biased model negatively affect users?

#### Why can it happen that a model discriminates?

#### 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?

#### Name two intrinsically interpretable models:



#### How can you explain an individual prediction of a linear model?

#### How is the permutation feature importance computed?

#### How is a partial dependence plot generated?

#### How can an intrinsically interpretable surrogate model be used to explain an individual prediction of a more complex model?

#### How can you generate optimal inputs and counterfactual examples for a neural network?

– Part 5 –

## 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?



#### Where do you think your organization stands in this AI transformation process?