# Learning journal central document

## Key takeaways per chapter

## Homework 1:

* Machine learning is a branch of artificial intelligence focused on training algorithms to make predictions or decisions based on data.
* There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning.
* Scikit-learn is a library that provides machine learning models and tools for data preprocessing, feature engineering, model selection, and evaluation.

1. What is a machine learning pipeline? Please make a fictitious example that is appropriate for our data set.

A machine learning pipeline is a series of steps that handle raw data to create a model that can forecast on new data. The typical process includes preprocessing data, selecting relevant features, choosing a model, tuning hyperparameters, and evaluating the performance. For our dataset, a possible example of this could involve scaling numerical data, encoding categorical data, selecting meaningful features, training a linear regression model, and evaluating its performance with cross-validation.

2. Is our task a a supervised, unsupervised, semi-supervised, self-supervised, or reinforcement learning task? Is it a classification or regression task? Do we use batch or online learning?

Our task is a supervised regression learning task. We aim to predict the price of a house with the given features. The model uses batch learning, which means it is trained with the complete dataset at once.

3. What is a performance measure? Give two typical examples of performance measures in regression models. How can these be interpreted in a linear regression

A performance measure is a tool used to assess a machine learning model's effectiveness. Two typical regression performance measures are mean squared error (MSE) and the coefficient of determination (R-squared). MSE calculates the average of the squared differences between predicted and actual values, while R-squared estimates the percentage of variance in the target variable that the model explains. In linear regression, the MSE measures the average squared distance between the predicted and true values, while R-squared measures the percentage of the target variable's variance that the linear regression model explains. A higher R-squared value signifies a more accurate fit of the model to the data.

4. What is stratified sampling?

In statistical sampling, stratified sampling is a technique where the population is divided into homogeneous groups, called strata, and then a random sample is selected from each stratum. This sampling method ensures that each stratum is represented in the sample, which can be important when the strata have different proportions of the target variable or different distributions of important features.

For example, in the context of the housing dataset used in the book, a stratified sample might be used to ensure that the sample contains a representative proportion of houses from each neighborhood (stratum), rather than simply selecting a random sample of houses from the entire dataset. This can be important because different neighborhoods might have different price ranges or other features that are important for predicting housing prices. By using stratified sampling, the sample is more likely to be representative of the population and the resulting models are likely to be more accurate.

4. What are ordinal and categorical data? Do we have such data in our data set? What possibilities do we have to transform this data correctly? What are dummy attributes? Try to understand what the function *OneHotEncoder()* does with the variable waterfront.

Ordinal data is data that can be ranked or ordered, such as movie ratings (1 star, 2 stars, 3 stars, etc.). Categorical data is data that belongs to a specific group or category, such as color (red, blue, green, etc.) or type of animal (mammal, reptile, bird, etc.).

The kc\_housedata dataset contains ordinal data, such as the "grade" column that rates the quality of construction and design of the house on a scale from 1 to 13. The dataset also has categorical data, such as the "waterfront" column, which indicates if the house has a view of the waterfront or not.

To correctly transform ordinal data, it is essential to maintain the order in the model. For instance, the "grade" column can be converted into a numerical scale from 1 to 13, where higher values indicate better quality.

To correctly transform categorical data, one-hot encoding technique can be applied. This involves creating dummy attributes for each category, where 1 is assigned to the relevant attribute for each instance that belongs to that category, and 0 for all other attributes. For example, two dummy attributes can be created for the "waterfront" column, one for houses with a waterfront view and one for houses without a waterfront view.

The OneHotEncoder() function in scikit-learn can be used for one-hot encoding. When used on the "waterfront" column in the kc\_housedata dataset, it would generate two dummy attributes for "waterfront," one indicating "waterfront=0" and the other indicating "waterfront=1."

5. What is min-max scaling and standardization of features? What is a heavy tail in a distribution? What problem do we have with min-max scaling and standardization on a heavy-tailed distribution? Can you make an example?

Commonly used preprocessing techniques in machine learning include min-max scaling and standardization, which are employed to transform features into a common scale, improving the performance of various machine learning algorithms.

To perform min-max scaling, which is also known as normalization, feature values are rescaled within a specific range, typically [-1,1] or [0,1], by subtracting the feature's minimum value and dividing by the range of the feature (i.e., the difference between the maximum and minimum values).

On the other hand, standardization transforms feature values so that they have a mean of zero and a standard deviation of one. This is done by subtracting the mean of the feature and dividing by the standard deviation.

A heavy-tailed distribution is characterized by a long tail of infrequent values that are much larger than the majority of values in the distribution. For instance, a population's income distribution may have a small number of individuals with significantly higher incomes than the majority of people, leading to a heavy tail.

Applying min-max scaling or standardization to a heavy-tailed distribution may amplify the differences between the tail values and the majority of values, resulting in a loss of information in the majority of values. For example, min-max scaling on a feature with a heavy-tailed distribution will rescale the tail values to [0,1], which can be much larger than the rescaled values of the majority of the data. This makes it challenging for machine learning algorithms to differentiate between the tail values and the majority of the values, which can adversely impact performance.

6. Have a look at my suggestion of variable conversion. I ask you to understand the code in detail, what variable transformations have I made? What arguments are in favor, and what arguments might be against? How would you convert the variables if you compare it with your guess?

As I didn’t want to describe all lines of code, here are some examples:

* time function: datetime to seconds, standardized with StandardScaler
* ClusterSimilarity function: cluster lat and long into variables representing their cluster
* np.log function: applied to sqft\_living, sqft\_lot, sqft\_above, and price variables, standardized afterwards
* OneHotEncoder function: converts waterfront and renovated categorical variables into dummy variables
* StandardScaler function: applied to yr\_built, yr\_renovated, bedrooms, bathrooms, floors, view, condition, grade variables

7. Try to explain what the Grid Search and Random Search methods do. What are the advantages of each method. For which parameters would this make sense in our example?

Grid Search and Random Search are two methods used for tuning hyperparameters in machine learning models to find the best combination of hyperparameters.

Grid Search is an exhaustive technique that searches through a specified subset of the hyperparameter space, evaluating the model's performance for each possible combination of hyperparameters, and selecting the best one based on a scoring metric such as accuracy or mean squared error. It is best suited for smaller search spaces where the number of hyperparameters is limited.

In contrast, Random Search randomly samples from the hyperparameter space, which can be more efficient than Grid Search when dealing with a large number of hyperparameters or when the optimal hyperparameter values are not known. Random Search may find the optimal hyperparameters in a shorter amount of time and outperform Grid Search.

In our example, Grid Search or Random Search could be used to tune hyperparameters such as the regularization strength for the linear regression model or the number of estimators and maximum depth for the random forest model. These hyperparameters play a crucial role in achieving the best performance for these models.

## Homework 2

* Machine learning projects consist of various steps, such as data acquisition, data cleaning, data exploration and visualization, feature engineering, model selection, and evaluation.
* Data preprocessing involves dealing with missing values, outliers, encoding categorical variables, scaling numerical variables, and splitting data into training and testing sets.
* Data visualization can help understand the relationship between variables and detect patterns in the data.
* Often it is necessary to visualize the data to be able to build a comprehensive model

## Homework 3

* Linear regression is a powerful algorithm that predicts continuous values.
* Polynomial regression can detect nonlinear relationships between variables.
* How to avoid overfitting
* Logistic regression is a binary classification algorithm that forecasts the probability of a binary outcome.
* Softmax regression is a multiclass classification algorithm that predicts the probability of each class.

## Homework 4

* Decision trees are simple models that can handle both numerical and categorical data.
* Random forests are an ensemble of decision trees that can reduce overfitting and improve performance.
* Gradient boosting is another ensemble method that trains models sequentially, with each new model correcting the errors of the previous one.
* Support vector machines (SVMs) are binary classifiers that find the best separating hyperplane between classes.
* SVMs use kernels to transform data into a higher-dimensional space and identify nonlinear relationships between variables.
* Stochastic Gradient Descent, and Mini-Batch Gradient Descent

## Homework 5

1. What is a pre-trained NLP model?

A pre-trained NLP model is a type of machine learning model that has been trained on a large text corpus to learn the patterns and connections between words and phrases in natural language. These models can be utilized for various NLP tasks like sentiment analysis, named entity recognition, language translation, and more.

2. How do I load pre-trained NLP models?

There are multiple libraries and frameworks like Hugging Face's Transformers library, TensorFlow, PyTorch, etc., that you can use to load pre-trained NLP models. Depending on the library, you may have to download the model weights and configuration files and then load them into your code.

3. What is tokenization?

Tokenization is the process of dividing a piece of text into smaller units called tokens, such as words, subwords, or characters. Tokenization is an essential step in NLP tasks as it transforms text into a format that machine learning models can comprehend.

4. What does fine-tuning mean?

Fine-tuning is the process of taking a pre-trained NLP model and training it on a smaller, domain-specific dataset to adapt it for a particular task. Fine-tuning can improve the performance of the model on the target task since the model has already learned the general understanding of natural language from its pre-training.

5. What types of NLP models are there?

There are various types of NLP models, including Bag of Words models, Neural Networks (such as Recurrent Neural Networks and Convolutional Neural Networks), Sequence-to-Sequence models, Attention models, and Transformers (such as BERT, GPT-2, and T5).

6. What possibilities do I have with the Transformers package?

The Transformers package from Hugging Face is a widely-used and powerful library for working with pre-trained NLP models. Using the Transformers package, you can load and use pre-trained models for a wide range of NLP tasks, fine-tune pre-trained models on your own data, train your own models using the same architectures as popular pre-trained models, generate text using pre-trained models, visualize the attention and output of pre-trained models, and more.

7. How to use a transformer model in practise?

First, the input data needs to be preprocessed by splitting it into tokens, converting them into integer representations, and grouping them into batches. Depending on the model, special tokens and input masks might need to be inserted.

Next, the model processes the inputs and produces intermediate vector representations of the sentences.

Finally, the post-processing step involves converting the number representation of the final task result back into human-readable format, using different functions depending on the task and model.

The huggingface Transformers library provides a Tokenizer object for handling pre-processing and post-processing steps, as well as a Trainer object for fine-tuning models. Additionally, a pipeline can be created to handle all necessary steps from pre-processing to post-processing, as long as the model head is retained.

## Homework 6

1. Broadcasting is like a helpful assistant in numpy and PyTorch that makes working with arrays of different shapes much easier. It automatically adjusts the dimensions of arrays so they can be combined and computed together without any hassle.

Imagine you have two arrays with different shapes, and you want to perform some calculations on them. Broadcasting steps in and follows specific rules to make sure the arrays are compatible. It copies and stretches the values along the smaller dimensions, so they match the corresponding dimensions of the other array. This way, you can perform computations smoothly without needing to manually reshape or loop through the arrays.

The great thing about broadcasting is that it simplifies coding. It allows developers to write shorter and more readable code by eliminating the need for complicated loops or reshaping operations. So, when you're dealing with complex math operations, like working with matrices or performing calculations on each element, broadcasting comes to the rescue. It not only makes your code easier to understand but also improves its performance by avoiding unnecessary repetitive work.

In the world of deep learning models, broadcasting is especially valuable. It helps with various operations involving tensors of different shapes, such as calculating gradients or applying activation functions to each element. It also plays a key role in batch processing, where you can process multiple data samples at once, making the training process more efficient.

## Homework 7.1

1. What problem does collaborative filtering solve?

Collaborative filtering solves the problem of personalized recommendation by predicting user preferences or interests based on the behavior and preferences of similar users or items.

2. How does it solve it?

It solves this problem by analyzing the historical user-item interactions or ratings to identify patterns and similarities among users or items. The idea is that users with similar tastes or preferences in the past are likely to have similar preferences in the future, and items that are preferred by similar users are likely to be preferred by the target user as well.

3. Why might a collaborative filtering predictive model fail to be a very useful recommendation system?

collaborative filtering relies on the availability of sufficient user-item interactions or ratings data. If there is sparse or insufficient data, it becomes challenging to accurately identify similarities and make meaningful recommendations.

Secondly, collaborative filtering suffers from the cold-start problem. When new users or items are introduced to the system with limited or no interaction history, it becomes difficult to establish meaningful recommendations as there is insufficient data to identify similarities.

Additionally, collaborative filtering tends to recommend popular items or rely on mainstream preferences, which may lead to a lack of diversity in recommendations. It can result in the model overlooking niche or less-known items that could be of interest to certain users.

7. What is a dot product? Calculate a dot product manually using pure Python with lists.

can be use to calculate the similarity of user preferences to items in recommendation systems.

A dot product, also known as an inner product or scalar product, is a mathematical operation that takes two vectors of equal length and returns a scalar value. It is calculated by multiplying the corresponding elements of the two vectors and summing the products.

vector1 = [1, 2, 3]

vector2 = [4, 5, 6]

dot\_product = sum(x \* y for x, y in zip(vector1, vector2))

9. What is an embedding matrix?

An embedding matrix is a matrix that maps discrete categorical variables, such as words or user IDs, to continuous vectors of fixed dimensions.

11. Why do we need Embedding if we could use one-hot-encoded vectors for the same thing?

The use of embeddings over one-hot-encoded vectors offers several advantages. Firstly, embeddings can capture relationships and similarities between categories, as they are represented in a continuous vector space. This allows the model to generalize better to unseen data and handle out-of-vocabulary cases. Secondly, embeddings reduce the dimensionality of the input, making it more efficient to process and train models. Lastly, embeddings can be learned from the data itself, enabling the model to automatically derive meaningful representations for the categories.

This doesn't work for encoded cetegories, as they are only represented by binary 1 or 0.

18. What is the use of bias in a dot product model?

In a dot product model, biases play an important role in capturing additional information and improving the model's ability to make accurate predictions.

The use of bias terms allows the model to account for factors that cannot be fully captured by the dot product of the latent factors alone. While the latent factors capture the underlying interactions between users and items, biases provide an additional adjustment or offset to the prediction based on specific user or item characteristics. In our case the bias is used to take the rating tendency and/or how good a movie is preceived to be in a general perspective.

12. What does an embedding contain before we start training (assuming we're not using a pretained model)?

Randomly created tensors.

To initialize the embeddings, random values are assigned to each dimension of the embedding vector. The number of dimensions in the embedding vector is a hyperparameter chosen by the model designer, and it determines the expressive power of the embeddings. Typically, the dimensionality of embeddings can range from tens to hundreds, depending on the size and complexity of the dataset.

After initializing the embeddings, the recommendation model is trained using techniques such as matrix factorization or neural networks to optimize the embeddings based on the observed user-item interactions. The model learns to predict user preferences or item relevance based on the learned embeddings, and the embeddings are updated during the training process to improve the accuracy of the predictions.

23. What does argsort do in PyTorch?

The argsort function is particularly useful for tasks such as ranking or sorting elements based on their values. By utilizing the indices returned by argsort, you can retrieve the original tensor elements in sorted order or perform further operations based on the sorted indices. so you can find the coresponding indexes of a tensor soreted for example. This can be very useful to f.e. filter tensors to a ranking etc.

26. What is the "bootstrapping problem" in collaborative filterina?

The "bootstrapping problem" in collaborative filtering pertains to the predicament of generating personalized recommendations for novel users or items with limited or no historical data available. Collaborative filtering algorithms heavily rely on past user-item interactions to make accurate predictions and suggestions. However, when confronted with a new user joining the system or the introduction of a fresh item, the absence of data or insufficient user-item interactions poses a significant challenge.

The quandary arises from the fact that collaborative filtering approaches typically rely on discerning patterns and similarities from prior interactions to infer user preferences. Without substantial historical data, accurately assessing the preferences and characteristics of new users or items becomes arduous, impeding the provision of relevant and personalized recommendations.

27. How could you deal with the bootstrapping problem for new users? For new movies?

To ask the user questions to categorize him initially when he joins the platform, or initialize him with values for interests like the average person on the platform

## Homework 7.2

User-Based vs. Item-Based Collaborative Filtering: Collaborative filtering can be implemented in two main ways: user-based and item-based. The user-based approach recommends items based on users with similar tastes, while the item-based approach suggests items based on item similarities. The choice between these approaches depends on factors like dataset size, sparsity, and computational efficiency.

Scalability and Efficiency: Efficiently handling large datasets is crucial for collaborative filtering, given the numerous users and items in real-world systems. Optimization techniques, such as data partitioning, parallelization, and approximation algorithms, enhance scalability and efficiency.

Evaluation Metrics: Evaluating the performance of collaborative filtering algorithms is vital. Metrics like precision, recall, mean average precision, normalized discounted cumulative gain, and root mean square error are commonly used. The choice of metrics depends on the specific goals of the recommendation system.

Cold Start Problem: The cold start problem arises when new users or items join the system, lacking sufficient data for accurate recommendations. Collaborative filtering struggles in such scenarios. Hybrid approaches that combine collaborative filtering with content-based or knowledge-based methods help mitigate this issue by leveraging additional information.

Long-Tail Recommendations: Collaborative filtering can address the long-tail problem, where a few popular items overshadow niche items. Leveraging user preferences and item similarity, collaborative filtering algorithms recommend long-tail items, enhancing diversity and user satisfaction.

Sparsity Handling: Collaborative filtering often deals with sparse data, where users rate only a fraction of available items. This sparsity poses a challenge in predicting user preferences accurately. Techniques like matrix factorization, neighborhood-based methods, and dimensionality reduction address sparsity to improve recommendation quality.

Overcoming Bias and Personalization: Collaborative filtering can inadvertently reinforce biases present in the data, leading to homogeneous recommendations and excluding certain user segments. Techniques like diversity-aware recommendation algorithms and hybrid models help mitigate bias and enhance personalization.

## Additional Projects to the class assignments

## Project 1

Since we began with linear regression models, I chose to construct a linear regression model independently (assignment 3.1). However, I didn't want to use pre-existing data; instead, I desired a dataset that I could consistently update and extend. My interest in real-time financial data led me to build a Linear Regression model with the scikit library and yfinance. I aimed to predict the price of Bitcoin by assessing its correlation with other assets in the market.

To utilize the data consistently, I developed a database that stores the predictions, allowing me to create a trading bot that can make informed decisions based on the model's predictions. I also calculated the correlation between assets over various time periods, dynamically implementing a time range that adapts as I retrain the model. To diversify the assets, I included those with positive correlations, such as SPY to Bitcoin, those with negative correlations, such as Gold to Bitcoin, and a broad index of bonds. I utilized as many parameters as possible in the model to account for biased predictions due to the differences in base values of the assets, and converted the dataset to work on percentage change of the assets values.

Project 2

For my second project (assignment 3.2), I constructed a basic LSTM (long short term memory) neural network that predicted Bitcoin's price based on sequential analysis. I utilized historical data from Yahoo Finance via the yfinance library to focus specifically on BTC. I designed and trained an LSTM-based neural network model to predict future BTC prices based on past data. The model's efficiency was evaluated on a test set, and I created functions to determine the optimal training sequence length and visualize the model's predictions. The ultimate goal of this project was to develop a model that can accurately predict BTC prices based on historical data, potentially allowing for more informed cryptocurrency trading decisions.

### Project 3

As a third project (assignment 3.3), I used to create a linear regression model and a random forest to predict the probability of survival for passengers of the titanic. Firstly, I analyzed the dataset, whereas already some clear patterns emerged, converted the data, and selected a couple features for the models. I was quite happy with my third project, although I tried to keep it simple, as it was mainly a review of last week’s lesson.

## Project 4

In the context of my fourth project, specifically assignment 4.1, I employed a pretrained Convolutional Neural Network (CNN) to detect facial features within images sourced from an online platform characterized by rating-based interactions, commonly known as "hot or not." By extracting landmark coordinates from the detected faces, I computed scores based on various facial metrics, including symmetry and adiposity. These metrics served as inputs for constructing a linear regression model aimed at predicting the ratings associated with individuals, as inferred from the aforementioned facial landmarks. To delve deeper into the project and gain comprehensive insights, I encourage perusal of the associated presentation on CLIP and the Tinder-bot, which elaborates further on its intricacies and implications.

### Project 5

In the context of the fifth project, the Tinder API endpoint was reengineered, and a multifaceted bot was developed, incorporating several pretrained neural networks to facilitate swiping and communication within the Tinder application. The deployment of this bot was executed through a docker container within an Ubuntu runtime environment. For further insights into the project and its intricacies, a presentation was delivered jointly by myself and Damiano, shedding light on the various aspects and dimensions of the aforementioned undertaking.

### Project 6

I implemented a version of the Tinderbot, which used a cocoa model, in order to predict what can be found in the image.   
The Cocoa model typically consists of two main components: an image encoder and a language decoder.

1. Image Encoder: The image encoder is responsible for processing the input image and extracting high-level features that capture its visual content. This is typically achieved using a convolutional neural network (CNN). The CNN processes the image through multiple layers of convolution and pooling operations, gradually reducing the spatial dimensions while preserving important visual features. This process results in a compact representation of the image that encodes its visual content.
2. Language Decoder: The language decoder takes the encoded image representation from the image encoder as input and generates a textual description in the form of a caption. This component usually employs a recurrent neural network (RNN) or a transformer-based architecture. The RNN or transformer model processes the encoded image features in a sequential manner, generating words or tokens one at a time. At each step, the model predicts the most probable next word given the previously generated words and the image context. This process continues until an end-of-sentence token is generated, indicating the completion of the caption.

Training the Cocoa model involves leveraging large datasets of images paired with their corresponding captions. During training, the model learns to associate the visual features extracted from the image with the textual descriptions. This is achieved by minimizing the discrepancy between the predicted captions and the ground truth captions using techniques like maximum likelihood estimation or reinforcement learning.