Lecture 2: *"End-to-end Machine Learning Project"*

Tasks in the meeting

1. Create a checklist for the data set. What points should you pay attention to?
2. What types of encoders are there? For what kind of data are they used?
3. What methods are offered to simplify the finding of the best parameter combination? Explain how the two methods differ.
4. How does the scaling work on the train and test data?

Solution:

1. Create a checklist for the data set. What points should you pay attention to?
- All data points non-null, no NaNs? -> fill in
- All of type float or int? -> convert e.g. with one-hot-encoding
- Delete all doubled data points
- Exist outliers? -> remove
- Are input features strongly correlated? -> remove them, but for some models not necessary
- Balanced according to the output feature?

2.

Categorical features	No ordering possible	OneHotEncoder	One column for each value
Ordinal features	Ordering possible	OrdinalEncoder	Only one column, Every value gets a number according to its position in the order

3. GridsearchCV: Define a list of parameters to optimize together with a list of the values which should be tried. GridSearchCV tries all the different combinations and give the best estimator for a defined scoring

RandomizedSearchCV: Same definition as for GridSearchCV, but it tries a finite number of combinations which are randomly chosen

Ensemble Methods mean that we combine different good models to one. From my opinion it brings just 2-3% more accuracy.

4. The scaling is defined for the train set. Then the same scaler is applied to the test set to get the same scaling.