Mandatory Assignment 2

Classification & Clustering

Markus Kinn

Tuning Strategy:

I use this strategy for all tuning, regardless of model, unless specified different in the notebook (I do test out a sequential method in Oblig 3).

I always start the first grid with values which I know I in the range where I usually find the best values. After running the grid search, I find the best values which was represented in the grid, and then I create a new grid in the cell below where I test out new values. The new values I choose are always: between the value to the left of the best value and the best value, the best value, and between the best value and the value to the right of the best value.

Ex.: a grid contains the learning rate values 0.05, 0. 1, 0.2 and 0.3. Let's say the grid search finds out that the best value which was represented was 0.1. In the next cell I will create a new grid search with: one value above 0.05, but below 0.1, one value which represent the best value from the previous grid(0.1), and one value above 0.1, but under 0.2. I do this for every hyperparameter I test for each grid.

Optimally I should only change the values of one parameter at the time, but that would take too long.

Classification:

Dataset 1:

For the first dataset I chose to continue with the titanic dataset from the first mandatory assignment. The goal is the same as before; predict whether a person was going to survive the titanic disaster or not.

The has dataset has 12 columns and 891 rows.

The final dataset (after preprocessing) contains 15 columns:

Survived(the ground truth), Sex, FamilyMembersCount, Pclass_1, Pclass_2, Pclass_3, Embarked_C, Embarked_Q, Embarked_S, AgeGroup_0, AgeGroup_1, AgeGroup_2, AgeGroup_3, AgeGroup_4, AgeGroup_5.

I got the best result with: XGBoost

Dataset 2:

The second classification dataset I used was: Heart Attack Analysis & Prediction Dataset. (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset). This dataset is used to predict if a person is likely to have a heart attack given current metrics.

The has dataset has 14 columns and 303 rows.

There are 10 columns (including ground truth):

Age: age of patient, Exng: exercised induced angina, Ca: number of major vessels, Cp: Chest Pain type, Trtbps: resting blood pressure, Chol: cholestoral, Fbs: fasting blood sugar, Rest_ecg: resting electrocardiographic results, Thalach: max heart rate achieved, Target: 0 = lower chance of heart attack and 1 = higher chance of heart attack

I got the best result with: Support Vector Machine

Clustering:

Dataset 1:

For the clustering part of the assignment I chose to use the same clustering dataset form the first mandatory assignment: Airline Passenger Satisfaction (https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). This dataset is usually used as a classification dataset with the ground truth: Satisfaction, but for this task the ground truth has been removed.

The has dataset has 25 columns and 103904 rows.

Due to the dimensionality, I ended up removing 90.000 rows for both the Birch algorithm and Agglomerative.

I got the best result with: Agglomerative clustering.