ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Machine Learning
Summer 2023

Prof. Dr.-Ing. Klaus Berberich
Phone: +49 681 58 67-243
klaus.berberich@htwsaar.de

Programming Assignment 1

The programming assignment will be discussed on **May 22nd, 2023**. To obtain bonus points, you have to submit your solution as a Zip-file via Moodle by **May 18th, 2023 23:59**. Teams of up to **three** students are allowed. Please indicate your names and your student numbers in your submission.

1.1 Predicting the Popularity of Songs (2.5)

For this exercise, we'll work with data from Spotify provided via Kaggle. More details about the dataset can be found at

https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019

You can download the dataset there or from Moodle. We are interested in predicting the feature POP (Popularity) based on other features.

(a) First, we want to eyeball the data and see whether there are any obvious relationships between the features BPM (Beats per Minute), NRGY (Energy), DNCE (Danceability), and VAL (Mood) and our target feature POP (Popularity). To this end, for each feature under consideration, **create a scatter plot** having the current feature on the x-axis and our target feature POP on the y-axis. Also, **compute the correlation coefficient** for each pair of features.

(b) Now, we want to use ordinary least squares to predict the POP (Popularity) based on one of the other four features. For each of the four features, determine **optimal coefficients**, add the obtained **regression line** to the respective scatter plot from (a), and compute the **mean squared error** (MSE).

(c) Finally, we want to use multiple linear regression and predict the POP (Popularity) based on all four features. In addition, we want to take the feature GENRE (Genre) into account – a *nominal* feature which needs some encoding upfront. **Randomly split** the data into 80% training and 20% test data, determine **optimal coefficients**, and compute the **mean squared error** (MSE) on the test data.

1.2 **Predicting the Price of Used Cars** (2.5)

Our second exercise looks at a bigger dataset about prices of used cars. More details about the original dataset can be found at:

https://www.kaggle.com/datasets/harikrishnareddyb/used-car-price-predictions

You can download the dataset there or from Moodle.

(a) First, we want to use multiple linear regression to predict PRICE based on the features YEAR and MILEAGE. Use **5-fold cross-validation** and report the **mean of the root mean squared errors** (RMSE) as measured on the five test folds.

(b) Now, try to improve your model by adding some of the other available features except VIN. Note that all of the other available features need to be encoded first. You are also free to compute additional (derived) features (e.g., combining mileage and year). Evaluate your model again using **5-fold cross-validation** and report the obtained estimate of RMSE.

(c) Often, feature engineering is crucial to achieve good prediction quality. One way to obtain additional features is to bring in another dataset that includes background knowledge. Please describe three additional features that could be added if you had the English Wikipedia available in a suitable format. You do not have to implement this!

(d) Discuss how you could achieve a RMSE of zero. Which additional feature would you need to bring in? Why wouldn't this be a practical solution? You do not have to implement this!