Assignment on k-Nearest Neighbors

Machine Learning

Prof. Dr. Marc Pouly & Dr. Tim vor der Brück
Lucerne University of Applied Sciences and Arts

Deadline: March 12, 2018.

Exercise 1: Theory Questions (1 Point)

Fill in the online theory quiz on ILIAS.

Acceptance criterion: This is an individual exercise. You can participate in the quiz as many times as you want. ILIAS registers your best result over all runs. You passed this exercise if your best score reaches at least 80%.

Exercise 2: Breast Cancer Classification with k-NN (0 Points)

We provide a Jupyter notebook with k-NN applied to a classical dataset on breast cancer detection that is often used by the machine learning community. Note that a lot of data cleaning and wrangling has already been done on this dataset, which makes it a good starting point for machine learning classes. However, keep in mind that coming across such a well-behaved dataset in practice is rather unrealistic. Carry out the notebook, study the code and read carefully our comments. Consider this as a preparation for the next exercises, i.e. you do not need to hand in anything.

Exercise 3: Car Price Prediction for AutoScout24 (1 Point)

This is a team exercise; hand in your solution as a Jupyter notebook by e-mail to Tim vor der Brück no later than the above deadline.

AutoScout24 provided us with a dump of their productive database. You are assigned to train a prediction model[1] for the selling price of second-hand cars, which in the future shall enable the platform to automatically suggest an adequate selling price whenever a customer uploads a new sale advertisement. Carry out the following steps:

- Import and inspect the data; transform categorical attributes
- Search for anomalies and suggest appropriate measures
- Inspect correlations between the price and other attributes
- If necessary, perform data cleaning actions
- Split your data into 80% train and 20% test set and normalize your data
- Train a k-NN regression model and determine the best choice for k
    - Measure prediction quality with R^2 and consult your favorited text book in order to correctly interpret this measure of explained variance
    - Measure prediction quality as the (mean) absolute price difference between your model and the ground truth

[1] *In the real estate sector, similar models are frequently employed to estimate selling prices of properties.*