DAT405/DIT407 Introduction to Data Science and AI, SP4 22-23

Assignment 2: Regression and classification

1. The dataset associated with this assignment was downloaded from www.hemnet.se on 2020-10-18. The data contains information about selling prices of villas in Landvetter that were sold in the previous 12 months.
    a. Find a linear regression model that relates the living area to the selling price. If you did any data cleaning step(s), describe what you did and explain why.
    b. What are the values of the slope and intercept of the regression line?
    c. Use this model to predict the selling prices of houses which have living area 10 m^2, 100 m^2, 150 m^2, 200 m^2, 1000 m^2 .
    d. Draw a residual plot.
    e. Is this a useful model? Are there any limitations? What could you do to improve the models ability to predict selling prices? Can this model be used in other areas than Landvetter?

2. In this question, you will use the Iris data set ("from sklearn.datasets import load_iris").
    a. Visualise the data. Can you gain any insights from the visualisation?
    b. Use a confusion matrix to evaluate the use of logistic regression to classify the iris data set.
    c. Use k-nearest neighbours to classify the iris data set with some different values for k, and with uniform and distance-based weights. What will happen when k grows larger for the different cases? Why does this happen? What do you think is the best choice of k? Compute a confusion matrix for the best uniform and distance-based classifiers.
    d. Compare the logistic regression classifier in (a) with the k-nearest neighbour classifiers in (b). What do you observe? Are all classes equally challenging for the models to predict?

Submitting work

The **most convenient format for submitting your work is by extracting a pdf from your notebook**. This way, you can include both code, figures and text in one file, and we can easily view it directly in Canvas, meaning marking will be quicker and you get your feedback sooner. Write the name of the participants in the group as well as the number of hours each person has worked in the beginning of the notebook. Make sure all cells are run so that the output (plots, results, etc.) are all visible in the pdf.
The submission should contain:

- The notebook in pdf.
- The notebook in .ipynb (standard notebook) format