DIT852 Introduction to Data Science and AI, SP1 2021

Assignment 2: Regression and classification

1. The dataset associated to 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 past 12 months. [4p]
    a. Find a linear regression model that relates the living area to the selling price. If you did any data cleaning step(s), describe and explain why you did that.
    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 100 m^2, 150 m^2 and 200 m^2.
    d. Draw a residual plot.
    e. Discuss the results, and how the model could be improved.

2. In this question, you will use the Iris data set ("from sklearn.datasets import load_iris"). [4.5p]
    a. Use a confusion matrix to evaluate the use of logistic regression to classify the iris data set.
    b. 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?
    c. Compare the classification models for the iris data set that are generated by k-nearest neighbours (for the different settings from question 3) and by logistic regression. Calculate confusion matrices for these models and discuss the performance of the various models.

3. Explain why it is important to use a separate test (and sometimes validation) set. [1.5p]

Submitting work

In each file that you submit, give the names of the people submitting the work. On the first page of the report state how many hours each person spent working on the assignment.

If you upload a zip file, please also upload any PDF files separately (so that they can be viewed more conveniently in Canvas).

Remember, we check for plagiarism and we are obliged to report suspected cases.

Deadline: Monday 13 September 2021 at 23:59.

Self-check

Is all the required information on the front page? Have you answered all questions to the best of your ability? Anything else you can easily check? (details, terminology, arguments, clearly stated answers etc.?)

Grading will be based on a qualitative assessment of each assignment. It is important to:

i. Present clear arguments
ii. Present the results in a pedagogical way
    i. Should it be table/plot? What kind of plot? Is everything clear and easy to understand?
iii. Show understanding of the topics
iv. Give correct solutions.
v. Make sure that the code is well commented.
    i. Important parts of the code should be included in the running text and the full code uploaded to Canvas.