Assignment 2: Regression and Classification

Due September 11, 2023 at 13:00

1 Instructions

In this assignment you will work with two different data sets: one from Hemnet[1] and one from Scikit-learn[2]. Instructions for downloading each dataset, along with the accompanying questions regarding each dataset, are given below.

1.1 Hemnet Dataset

The Hemnet dataset associated with this assignment was downloaded from `www.hemnet.se` on 2020-10-18. You can download it from the Module 2 Assignment page on Canvas; the file is called *'hemnet.csv'*. The data contains information about the selling prices of villas in Landvetter that were sold in the previous 12 months.

Answer the following questions about this dataset:

1. Find a linear regression model that relates the living area to the selling price. If in doing so, you performed any data cleaning step(s), describe what you did and explain why.
2. What are the values of the slope and intercept of the regression line?
3. Use this model to predict the selling prices of houses which have living area 100 m^2, 150 m^2, and 200 m^2.
4. Draw a residual plot. Discuss some potential strategies for improving the model.

1.2 Iris Dataset

The Iris dataset can be imported from Scikit-learn as follows:

```
from sklearn.datasets import load_iris
```

Answer the following questions about this dataset:

1. Use a confusion matrix to evaluate the use of logistic regression to classify the Iris data set.
2. Use k-nearest neighbors 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?
3. Compare the classification models for the Iris data set that are generated by k-nearest neighbors (for the different settings from question 2) and by logistic regression. Calculate confusion matrices for these models and discuss the performance of the various models.

1.3 What To Submit

- A PDF report that includes the figures produced and the descriptions/discussions that are requested in the questions.
- All Python code written. This may take the form of (i) individual Python files (submit the Python source code and append the code as an appendix in your PDF report), or (ii) a Jupyter notebook (submit both the Jupyter notebook and a separate PDF file exported from Jupyter).
- **Do not submit "zip" files**.

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

Ensure that all group members have joined the Assignment Group in Canvas before submitting your solution.

2 Self-Check: *Please read this before submitting your report*

In a data science project, it is usually not sufficient to write a program, run it, and then present a graph or table with results. We should also think carefully about the data that has been used, have a close look at the results, and consider whether these seem reasonable.

For this assignment in particular, it is important to check for yourself the following points:

- ☐ Did you do any data cleaning (e.g., by removing entries that you believe are not useful) for the task of plotting the data and answering the questions above? If so, be sure to explain what kind of entries you chose to remove, and why. If you chose not to do any data cleaning, please specify this in your report, and why.
- ☐ Did you present your arguments clearly, and describe your results in a pedagogical way? For instance, did you choose to present your results in a table or plot? If you chose a plot, what type of plot did you choose, and could there have been any better choices?
- ☐ Did you make sure that all your code is included in your report, and that it is well-documented (e.g., makes clear use of comments)?

Additionally, for all assignments, **always perform the following self-checks** before submission:

- ☐ Have you answered all questions to the best of your ability?
- ☐ Is all the required information on the front page (e.g., you and your partner's names, the hours each partner worked, the correct file name, etc)
- ☐ Anything else you can easily check? (e.g., clearly labeled axes in figures, clearly labeled units, clear terminology and arguments, clearly stated answers, etc)

Do not submit an incomplete assignment! We teachers are available to help you, and you can receive a short extension if you contact us.

References

[1] *Hemnet Dataset*. Retrieved October 18, 2020, from https://www.hemnet.se/

[2] *Iris Dataset*. Retrieved September 03, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html