Exercise 3 - Linear Regression

Winter term 2019/2020

student1, student2, student3

General information

The assignments are voluntary. All students who choose to participate have to form groups comprising three to four students (not more and not less). The groups do not have to be static, you may form new groups for each assignment. You have **two weeks** to answer the questions and to submit your work. The solutions are going to be presented and discussed after the submission deadline. Sample solutions will **not** be uploaded. However, you are free to share correct solutions with your colleagues **after they have been graded.**

Formal requirements for submissions

Please submit your solutions via Moodle (as a .zip file) as well as in printed form. The .zip file must contain one .pdf file for the pen-and-paper tasks as well as one .py file per programming task. Only pen-and-paper tasks have to be printed, you do not have to print the source code. Only one member of the group has to submit the solutions. Please make sure to specify the matriculation numbers (**not the names!**) of all group members so that all participants receive the points they deserve!

Please refrain from submitting hand-written solutions or images of solutions (*.png* / *.jpg* files). Rather use proper type-setting software like LaTeX or other comparable programs. If you choose to use LaTeX, you may want to use the template files provided.

Code assignments have to be done in Python. Please submit *.py* files (**no jupyter notebooks**). The following packages are allowed for code submissions: `numpy`, `pandas` and `scipy`. Please ask **beforehand**, if you want to use a specific package not mentioned here. Finally, do not use already implemented models (e.g. from `scikit-learn`).

Grading details

Your homework is going to be corrected and given back to you. Correct solutions are rewarded with a bonus for the exam which amounts to at most ten percent of the exam, if all solutions submitted by you are correct (this corresponds to at most six points in the exam). It is still possible to achieve full points in the exam, even if you choose not to participate in the assignments (it is additional). The function which is used to compute the bonus is given by:

$$b(a) = \min\left(B, \left\lceil \frac{B}{A^2} \cdot a^2 \right\rceil\right) \tag{1}$$

- b denotes the number of bonus points you get for the exam (this is up to you)
- B refers to the maximum attainable bonus points for the exam (six points)
- A denotes the maximum attainable points in the assignments (40 points)
- a is the score you achieved in the assignments (this is up to you)

Please note: You have to pass the exam **without the bonus points!** This means that it is not possible to turn a failing grade ($= 5.0$) into a passing grade (≤ 4.0). The bonus points will be taken into account in case you have to repeat the exam (i. e. they do not expire if you fail the first attempt).

Important!

The solutions have to be your own work. If you plagiarize, you will lose all bonus points!

1 Linear Regression

a) Ordinary Least Squares (5 points)

Implement an ordinary least squares regression model optimized with gradient descent to predict the value of a house in Boston. Please use the data set stored in `/data/bostonhousingdataset.csv`. The last column contains the class label which is called `medv` (median value of owner-occupied homes in $1000). You can find an explanation of each attribute on Kaggle.[1] Split the data into train and test sets, evaluate and report the mean squared error (MSE) of your model on the test data set.

Solution:

b) Basis Function Features (3 points)

Compute polynomial or radial basis function features from the raw features in the data set. Optimize the basis functions for the task (i. e. tune the degree of the polynomials or the means and scale of the radial basis functions). Which basis functions worked best?

Solution:

[1] `https://www.kaggle.com/c/boston-housing`

c) Regularization (2 points)

Explain in your own words what *regularization* is, why it is beneficial and which kinds of regularization you could apply to a linear regression model. Finally, explain what *ridge regression* is.

Solution:

d) Bonus Question 1 (1 point)

What are the three most important features for the prediction according to your linear regression model (without basis functions)? Explain your answer.

Solution:

e) Bonus Question 2 (1 point)

Plot the residuals for your regression model (y-axis) and the predicted values (x-axis). What can the residuals tell you about the performance of your model?

Solution: