Exercise 3 - Linear Regression

Winter term 2019/2020

name1, name2, name3

Important

Please solve the assignments in groups of 3 to 4 students. 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**. Please submit your solutions via Moodle **and** in printed form. Only one member of the group has to submit the solutions. Therefore, make sure to specify the names of all group members. Please do not submit hand-written solutions, rather use proper type-setting software like LaTeX or other comparable programs.

Your homework will be corrected and given back to you. Correct solutions are rewarded with a bonus for the exam (max. 10 percent, if all solutions submitted are correct). **Please note:** You have to pass the exam **without the bonus points**! *(i.e. it is not possible to turn 5.0 into 4.0)* The solutions have to be your own work. If you plagiarize, you will lose all bonus points!

Further remarks:

- Code assignments have to be done in `Python`
- The following packages are allowed: `numpy`, `pandas`
  (please ask, if you want to use a specific package not mentioned here)
- **Do not use already implemented models** (e. g. from `scikit-learn`)

1 Linear Regression

a) Ordinary Least Squares (5 points)

Implement an ordinary least squares regression model using gradient descent to predict the value of a house in Boston using the Boston Housing 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 $ 1000s). 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 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? 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: