MTU
Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

MACHINE LEARNING

Assignment 3: Regression & optimisation

DUE DATE

This assignment should be submitted to Canvas before 11:59pm on **Friday 17/12/2021**.

Please submit a single ZIP file with your student number and name in the filename. Your submission should contain **exactly 2 files**:

- A detailed documentation of all code you developed, including the tests and evaluations you carried out. Please make sure that you include a .pdf document with every result you produce referencing the exact subtask and lines of code it refers to.
- All Python code you developed in a single .py file that can be executed and that generates the outputs you are referring to in your evaluation. The file needs to be readable in a plain text editor, please do NOT submit a notebook file or link. Please also make sure that you clearly indicate in your comments the exact subtask every piece of code is referring to.

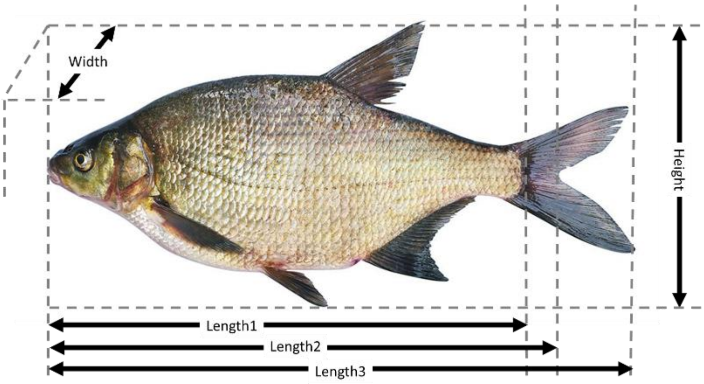
Please do NOT include the input files in your submission.

You can achieve a total of 35 points as indicated in the tasks.

OBJECTIVE

The Excel file "fish.csv" on Canvas contains measurements of the size and weight of different types of fish.

The goal of this assignment is to train a regression function to determine the weight of a fish based on its size.



TASK 1 (pre-processing, 6 points)

Create a function that loads the file [1 point] and extracts the different species of fish contained in the dataset [1 point]. For each of these species separately [1 point] extract the corresponding features [1 point] and targets [1 point]. Use the size measurements as features and the weight as target. Count the number of data-points for each species and select those for further processing that contain at least 20 samples [1 point].

TASK 2 (model function, 5 points)

Create a polynomial model function that takes as input parameters the degree of the polynomial, a list of feature vectors as extracted in task 1, and a parameter vector of coefficients and calculates the estimated target vector using a multi-variate polynomial of the specified degree [4 points]. Create a second function that determines the correct size for the parameter vector from the degree of the multi-variate polynomial [1 point].

TASK 3 (linearization, 2 points)

Create a function that calculates the value of the model function implemented in task 2 [1 point] and its Jacobian [1 point] at a given linearization point. The function should take the degree of the polynomial, a list of feature vectors as extracted in task 1, and the coefficients of the linearization point as input and calculate the estimated target vector and the Jacobian at the linearization point as output.

TASK 4 (parameter update, 3 points)

Create a function that calculates the optimal parameter update from the training target vector extracted in task 1 and the estimated target vector and Jacobian calculated in task 3. To do that start with calculating the normal equation matrix [1 point]. Make sure that you add a regularisation term to prevent the normal equation system from being singular. Now calculate the residual and built the normal equation system [1 point]. Solve the normal equation system to obtain the optimal parameter update [1 point]. The function should take the training target vector and the estimated target vector and Jacobian at the linearization point as input and calculate the optimal parameter update vector as output.

TASK 5 (regression, 3 point)

Create a function that calculates the coefficient vector that best fits the training data. To do that, initialise the parameter vector of coefficients with zeros. Then setup an iterative procedure that alternates linearization and parameter update [1 point]. Calculate the magnitude of the parameter update and choose a suitable threshold [1 point] to terminate the iteration [1 point]. The function should take the degree of the polynomial, the training data features, and the training data targets as input and return the best fitting polynomial coefficient vector as output.

TASK 6 (model selection, 5 points)

Setup a leave-one-out cross-validation procedure for all datasets extracted in task 1 [1 point]. Calculate the difference between the predicted weight and the actual weight for the test data point in each fold [1 point]. Compare different model functions by selection different polynomial degrees ranging from 0 to a maximal degree determined so that the number of training points is at least twice the number of estimated parameters [2 point]. Use the mean absolute weight difference as measure of quality for the different model functions to determine the best polynomial degree for each dataset [1 point].

TASK 7 (evaluation and visualisation of results, 11 points)

Estimate the model parameters for each species of fish [1 point] using the selected optimal model function as determined in task 6. Calculate the estimated weight for each fish in the dataset using the estimated model parameters [1 point]. Plot the estimated weights against the true weights [1 point] and observe how your estimation corresponds to the true weights. What is the maximum [1 point] and the average [1 point] estimation error and how does this compare to the maximum and average weight of the fish species considered [2 points]?

Inside the parameter update calculation in task 4 you calculated a residual. Observe how the residual changes in the iterations of the optimisation. What can you observe and why [2 points]? Inside this function you also applied a regularisation parameter. What happens if you increase/decrease this parameter and why [2 points]?