University of Stuttgart | ANALYTIC COMPUTING

Machine Learning (SS 23)

Assignment 04: Linear Regression

Mojtaba Nayyeri
Mojtaba.Nayyeri@ipvs.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ipvs.uni-stuttgart.de

Rodrigo Lopez Portillo Alcocer
rodrigo.lopez-portillo-alcocer@ipvs.uni-stuttgart.de

Tim Schneider
timphillip.schneider@ipvs.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

Submit your solution in ILIAS as a single PDF file.[1] Make sure to list the full names of all participants, matriculation number, study program, and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g. PPTX files). If you have any questions, feel free to ask them in the exercise forum in ILIAS.

Submission is open until Monday, 22.05.23, 12:00 noon.

[1] Your drawing software probably allows exporting as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

1. Linear Regression

(a) Linear regression can include nonlinear features. Why is it still called linear regression? In what sense is it linear?

(b) For calculating optimal parameters $\hat{\beta}$ the inverse of $X^\top X$ has to be calculated. When would this matrix be singular?

(c) Suppose that attempting to optimize the weights $\hat{\beta}$ is unsuccessful because the matrix $X^\top X$ is singular. Describe how you would alter the matrix X to prevent this issue.

2. Regression for Time-Series Prediction

In this task, we will explore a time-series regression problem. The model under investigation simulates a lake ecosystem, where we aim to predict the dissolved oxygen and algae content based on the following parameters: water temperature, water conductivity, water alkalinity, NO3 content, and total hardness of the water.

Our input, denoted as $X \in \mathbb{R}^{T\times 5}$, signifies the collection of these five parameters, where T represents the number of time steps, with each step corresponding to a month. The output, represented as $Y \in \mathbb{R}^{T\times 2}$, signifies the values for dissolved oxygen and algae content that we aim to predict.

Our prediction for the output at time step $t+1$ is dependent on the input and output at time step $t \in \{0,\dots,T-1\}$. Hence, we need to learn the function f, given by:

$$\hat{y}_{t+1} = f(x_t, y_t)$$

where x_t is the t-th row of X and y_t is the t-th row of Y.

(a) Formally define a linear regression model to estimate the function f.

(b) Dynamic systems, such as our lake ecosystem, often exhibit characteristics that are challenging to measure directly. Nevertheless, these dynamics can be inferred from long-term time dependencies. For instance, the water temperature recorded two months prior might still influence the present algae content in the lake.

In modeling dynamic systems, lag variables are often introduced. In this task, we will augment our linear regression model from part (a) to include the inputs and outputs from the previous two months (t and $t-1$) to predict the output at month $t+1$. Hence, the features at time step $t-1$ would be considered lag variables.

Update the function signature of f to match the described prediction task. Then, formally define the linear regression model including the lagged variables.

(c) Please proceed with this task by using the provided Jupyter notebook.