



University of Stuttgart



ANALYTIC
COMPUTING

Machine Learning (SS 23)

Assignment 04: Linear Regression ([Solution](#))

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.

¹Your 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?

Solution:

- It is linear in its weights β .
- The input data X can be replaced by arbitrary non-linear features without impacting the linearity of the model in its weights.

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

Solution:

The matrix $X^\top X$ is singular if and only if the columns of X are not linearly independent, e.g. $X[:, 1] = 3 * X[:, 0]$.

- (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.

Solution: To do this, compute the pairwise linear correlation between each column in X and remove any columns that display strong correlation. This will help to mitigate issues stemming from singularities in the $X^\top X$ matrix.

- (d) We are given data $D = \{(x_i, y_i)\}_{i=1}^n$ and we use a model $f(x) = \phi(x)^\top \beta$ where $x \in \mathbb{R}^d$ be d -dimensional and $\beta \in \mathbb{R}^k$ be k -dimensional. What is the cost function $L^q(\beta)$ for general linear regression with regularization of order q including a regularization parameter λ ?

Solution:

$$L^q(\beta) = \sum_{i=1}^n (y_i - \phi(x_i)^\top \beta)^2 + \lambda \sum_{j=2}^k |\beta_j|^q$$

- (e) Derive the minimizer β^* for ridge regression.

Solution:

The cost function for ridge regression can be expressed as:

$$L^{\text{ridge}}(\beta) = (X\beta - y)^\top (X\beta - y) + \lambda \beta^\top \beta$$

Taking the derivative of this cost function with respect to β and setting it to zero, we get:

$$\frac{\partial L^{\text{ridge}}(\beta)}{\partial \beta} = 2(X\beta - y)^\top X + 2\lambda \beta^\top = 0$$

Rearranging the terms, we get:

$$\beta^\top (X^\top X + \lambda \mathbb{I}) = y^\top X \Rightarrow \beta = (X^\top X + \lambda \mathbb{I})^{-1} X y^\top$$

- (f) What is the difference between Ridge and Lasso regularization? Provide a figure and briefly describe their effect on the optimal solution β^* .

Solution:

Lasso regularization can lead to sparsity in the coefficients, meaning that some coefficients are set exactly to zero. This can be useful for feature selection, as it can identify and remove irrelevant features from the model. In addition, computing the optimal solution for Lasso regularization can be computationally expensive.

Ridge regularization penalizes the length of the coefficient vector, which can help to prevent overfitting by discouraging large coefficients. However, it does not generally lead to sparsity in the solution, as

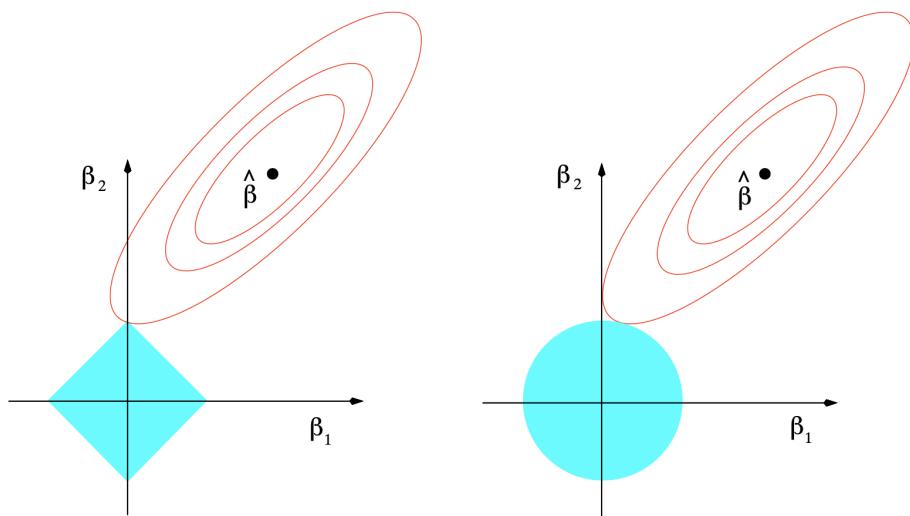


Figure 1 Estimation picture for the lasso (left) and ridge regression (right).

all coefficients are shrunk towards zero but are rarely set exactly to zero. In addition, computing the optimal solution for Ridge regularization is relatively easy and efficient.



2. Ridge 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 .

Solution:

For the formal definition of a linear regression model to estimate function f , we could define:

$$\hat{y}_{t+1} = A \begin{bmatrix} x_t \\ y_t \end{bmatrix} + b$$

with $A \in \mathbb{R}^{2 \times (5+2)}$ and $b \in \mathbb{R}^2$.

- (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.

Solution:

To accommodate this, we will update the function signature of f to be including lagged variables:

$$\hat{y}_{t+1} = f(x_t, x_{t-1}, y_t, y_{t-1}) \text{ with } t \in \{1, \dots, T - 1\}$$

The linear regression model, now incorporating lagged variables, could be formally defined as:

$$\hat{y}_{t+1} = A^{(0)} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + A^{(1)} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + b$$

with $A^{(i)} \in \mathbb{R}^{2 \times 7}$ for $i = 0, 1$ and $b \in \mathbb{R}^2$.

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