Linear Regression

Uczenie maszynowe

uczeniemaszynowe@cs.uni.wroc.pl

21 October 2025, 10:00 – lab session

General Notes

- All data are provided in `.csv` files.
- We recommend **not copying complete solutions** from language models. However, you may ask for feedback or additional materials on the given topic.
- During **the lab session**, your **understanding of each step** and **familiarity with your own solution** will be verified. You should have an environment ready with access to all data to make small code modifications on the spot.
- You do not have to write down detailed answers to the discussion questions, but it is strongly advised to think about them—even, say, while on the tram on your way to class.

1. Task 1 [2 points]

Is linear regression an appropriate model for the data in `data1.csv`? How should the data be prepared for modeling?

(1 pt) Perform the necessary calculations and visualizations. Record your observations and conclusions.

(1 pt) Propose how to prepare or transform the data so that a linear regression model can be used. Provide the solution as a single function:

```
preprocess_data(X_train, y_train)
```

whose behavior should be invariant under permutations of the training data.

Hints

- To load data, use `pd.read_csv()` from the `pandas` library—check the file header.
- Use `matplotlib` or `seaborn` for plotting.

- Choose a plot type by asking yourself what question you want to answer. For example, recall the assumptions of linear regression and verify if the data meet them. Decide which variables to visualize on the X and Y axes, and then choose an appropriate plot (`lineplot`, `scatterplot`, `histogram`, `KDEplot`, etc.).
- Do all calculations on the computer; use functions from common libraries.
- Start by describing what you see in the plot (e.g., data points form a circular pattern)—these are **observations**. Then move to **conclusions** (e.g., therefore, it is likely that the relationship between X and Y is not functional).
- Try to identify a simple dependency between X and Y and write its equation. This will help you build the full solution for the second part of the task.

Evaluation criteria or reference solution: TBA

2. Task 2 [3 points]

Using the `sklearn` library, fit a linear regression model to `data1.csv` and check its performance using the Mean Squared Error (MSE). Test how the preprocessing proposed in Task 1 affects the results.

(0.5 pt) Propose a suitable data split. Should it be the same when building a single model and when selecting the best model among several?

(0.5 pt) Propose a **baseline model**, i.e., a simpler reference model to compare performance and assess whether results are acceptable.

(0.5 pt) Build both the linear regression model and the baseline model on the considered data variants (one for raw data and one for preprocessed data).

(0.5 pt) Evaluate the models and present results in a **table**.

(1 pt) Evaluate the model on the dataset provided during the lab session. *This part's score will depend on your results.*

Evaluation criteria or reference solution: TBA

3. Task 3 [2 points]

Use the data `advertising.csv`.

(0.5 pt) Perform data analysis and preparation, following the pattern of Tasks 1 and 2.

(0.5 pt) Build linear regression models on data with **original scale** and on **normalized data**.

(0.5 pt) Compare the two models and record your observations and conclusions.

(0.5 pt) Identify the best model to evaluate on the test dataset provided during the lab.

Hint: Reuse experience (and code) from previous tasks. Avoid repeating code fragments.
Evaluation criteria or reference solution: TBA

4. Task 4 [3 points]

Use the data `data4.csv`.

(0.5 pt) Analyze the provided data.

(1 pt) Build two models:

a) standard linear regression, and

b) regression with **L2 regularization** (ridge regression) using an appropriate hyperparameter λ^2.

(1 pt) Compare the obtained regression parameters and performance on the test set. Which model performed better and why? Provide both **observations** and **theoretical justification**.

(0.5 pt) Evaluate the model on the dataset provided during the class.

Evaluation criteria or reference solution: TBA

Discussion for the Lab

What techniques and good practices did you learn or apply while solving this list? Which of them are specific to linear regression, and which are transferable to other topics or branches of machine learning?