

TASK 1A

Cross-validation for Ridge Regression (graded)

You are in the group 🧑‍🤝‍🧑 Mygroup consisting of 🧑 Irasmus (Irasmus@student.ethz.ch) and 🧑 nmojado (nmojado@student.ethz.ch).

📖 1. READ THE TASK DESCRIPTION

📁 2. SUBMIT SOLUTIONS

✉ 3. HAND IN FINAL SOLUTION

📖 1. TASK DESCRIPTION

This task is about using **cross-validation** for **ridge regression**. Remember that ridge regression can be formulated as the following optimization problem:

$$\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

We provide you with the following set of regularization parameters:

λ_1	λ_2	λ_3	λ_4	λ_5
0.1	1	10	100	1000

Your task is to perform **10-fold cross-validation** with **ridge regression** for each of the lambdas given above and report the **Root Mean Squared Error (RMSE) averaged over the 10 test folds**. In other words, for each lambda, you should train a ridge regression 10 times by leaving out a different fold each time, and report the average of the RMSEs on the left-out folds. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

The dataset consists of 500 points having 10 features. You should not use any feature transformations.

DATA DESCRIPTION

[Download handout \(/static/task1a_lm1d3z.zip\)](/static/task1a_lm1d3z.zip)

In the handout for this project, you will find the the following files:

- **train.csv** – the training set
- **sample.csv** – a sample submission file in the correct format

Each line in train.csv is one data instance indexed by an Id. It consists of one float for y and 10 float for the feature vector x1-x10:

```
Id, y, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10
0, 508.451969694, 0.886864686583, 0.400530684599, 1.10369395746, 0.13545466319, 8.1480691138
...
```

For your convenience, we further provide a sample submission file:

```
13.1
9.6
6.0
14.5
22
```

SUBMISSION FORMAT

The submission file format should be identical to the format of sample.csv, i.e. it should have 5 lines containing a float each, representing the average RMSE scores obtained for $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ (in this order).

Please keep in mind that, as a group, you have a limited number of submissions as stated on the submissions page.

EVALUATION

Your submission is evaluated based on the following formula:

$$\text{score}(\mathbf{v}) = \sum_{i=1}^5 \frac{|v_i - v_i^*|}{v_i^*} \cdot c_i$$

where $\mathbf{v} = \{v_1, \dots, v_5\}$ are your submitted values, $\mathbf{v}^* = \{v_1^*, \dots, v_5^*\}$ is the reference solution and $c_i = 2^{8-i}$ for $i \in \{1, \dots, 5\}$.

GRADING

When handing in the task, you need to select which of your submissions will get graded and provide a short description of your approach. This has to be done **individually by each member** of the team. We will then compare your selected submission to three baselines (easy, medium and hard). Your final grade depends on the public score and the private score (weighted equally), on your submitted code and on a properly-written description of your approach. The following **non-binding** guidance provides you with an idea on what is expected to obtain a certain grade: If you hand in a properly-written description, your source code is runnable and reproduces your predictions, and your submission performs better than the easy baseline, you may expect a grade exceeding a 4. If it further beats the medium baseline, you may expect that the grade will exceed a 5. If in addition your submission performs equal to or better than the hard baseline, you may expect a 6. If you do not hand in a properly-written description of your approach, you may obtain zero points regardless of how well your submission performs.

▲ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

FREQUENTLY ASKED QUESTIONS

WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library. However, **we strongly encourage you to use Python**. You can use publicly available code, but you should specify the source as a comment in your code.

IN WHAT FORMAT SHOULD I SUBMIT THE CODE?

You can submit it as a single file (main.py, etc.; you can compress multiple files into a .zip) having max. size of 1 MB. If you submit a zip, please make sure to name your main file as *main.py* (possibly with other extension corresponding to your chosen programming language).

WILL YOU CHECK / RUN MY CODE?

We will check your code and compare it with other submissions. We also reserve the right to run your code. Please make sure that your code is runnable and your predictions are reproducible (fix the random seeds, etc.). Provide a readme if necessary (e.g., for installing additional libraries).

SHOULD I INCLUDE THE DATA IN THE SUBMISSION?

No. You can assume the data will be available under the path that you specify in your code. For example, you could read in the dataset as:

```
import pandas as pd
df_train = pd.read_csv('train.csv')
```

CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a graded part of the class, **we cannot help you solve them**. However, feel free to ask general questions about the course material during or after the exercise sessions.

CAN YOU GIVE ME A DEADLINE EXTENSION?

⚠ We do not grant any deadline extensions!

CAN I POST ON PIAZZA AS SOON AS HAVE A QUESTION?

This is highly discouraged. Instead,

- Read the details of the task thoroughly.
- Review the frequently asked questions.
- If there is another team that solved the task, spend more time thinking.
- Discuss it with your team-mates.

If you still consider that you should contact the TAs, you can post a **private** question on Piazza. Remember that collaboration with other teams is prohibited.

WHEN WILL I RECEIVE THE PRIVATE SCORES? AND THE PROJECT GRADES?

We will publish the private scores, and corresponding grades before the exam the latest.