# TASK 1A

## Cross-validation for Ridge Regression (graded)

You are in the group Stonks consisting of jahanna (jahanna@student.ethz.ch (mailto:///[u'jahanna@student.ethz.ch'])), relazzouzi (relazzouzi@student.ethz.ch (mailto:///[u'relazzouzi@student.ethz.ch'])) and shimmi (shimmi@student.ethz.ch (mailto:///[u'shimmi@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:

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

Consider the following set of regularization parameters:

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ |
|---|---|---|---|---|
| $0.1$ | $1$ | $10$ | $100$ | $200$ |

Your task is to perform **10-fold cross-validation** with **ridge regression** for each value of $\lambda$ 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 leaving out a different fold each time, and report the average of the RMSEs on the left-out folds. The RMSE is defined as

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

You should perform the linear regression on the original features, i.e., you **should not perform any feature transformation or scaling**.

## DATA DESCRIPTION

Download handout (/static/task1a_do4bq81me.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 represents one data instance by an id, its label y, and its features x1 to x13:

```
y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13
22.6,0.06724,0.0,3.24,0.0,0.46,6.333,17.2,5.2146,4.0,430.0,16.9,375.21,7.34
...
```

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 exactly 5 lines, each containing a floating point number that represents 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:

$$\text{score}(\mathbf{v}) = 100 \cdot \sum_{i=1}^{5} \frac{|v_i - v_i^*|}{v_i^*}$$

where $\mathbf{v} = \{v_1, \ldots, v_5\}$ are your submitted values, $\mathbf{v^*} = \{v_1^*, \ldots, v_5^*\}$ is the reference solution.

## GRADING

We provide you with a dataset for which you have to make prediction or compute other quantities. For each of your submissions, a score is computed. 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 our baseline. Each project task is graded with either **pass or fail**. To pass the project, you need to achieve a better score than the baseline. In addition, for the pass/fail decision, we consider the code and the description of your solution that you submitted. The following **non-binding** guidance provides you with an idea on what is expected to pass the project: If you hand in a properly-written description, your source code is runnable and reproduces your predictions, and your submission performs better than the baseline, you can expect to have passed the assignment.

⚠ 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.

### AM I ALLOWED TO USE MODELS THAT WERE NOT TAUGHT IN THE CLASS?

Yes. Nevertheless, the baselines were designed to be solvable based on the material taught in the class up to the second week of each task.

### 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.

### 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. Remember that collaboration with other teams is prohibited. 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.

### 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.