TASK 0

Ungraded Dummy Task

You are in the group Accept consisting of lrasmus (lrasmus@student.ethz.ch (mailto://[u'lrasmus@student.ethz.ch'])) and nmojado (nmojado@student.ethz.ch (mailto://[u'nmojado@student.ethz.ch'])).

1. READ THE TASK DESCRIPTION | 2. SUBMIT SOLUTIONS | 3. HAND IN FINAL SOLUTION

1. TASK DESCRIPTION

This is an **ungraded** dummy task to prepare you for the five graded tasks later in the semester. After this task, you should know how to:

- register for a task,
- form a team of up to three students,
- read the task and data description,
- download the provided data sets,
- make a submission with your predictions and source code,
- see how your submission scores with regards to the baselines and the other students, and
- hand in the task by both choosing which submission should be graded and writing an **individual** task description.

While this task is ungraded and the problem itself is easy, we strongly recommend that you complete it as we will be **unable to extend** the deadlines for the graded tasks if you experience issues related to the points above.

TASK

This task is a trivial form of regression: Your goal is to predict a value **y** based on a vector **x**. While the exact relationship is usually not known, in this task, **y** is the mean of **x**. You may verify this on the provided training set. Your task is to make predictions for **y** on the provided test set.

DATA DESCRIPTION

Download handout (/static/task0_sl19d1.zip)

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

- **train.csv** - the training set
- **test.csv** - the test set (make predictions based on this file)
- **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 double for y and 10 doubles for the vector x1-x10:

```
Id,y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10
0,738.02,1764.05,400.15,978.73,2240.89,1867.55,-977.27,950.08,-151.35,-103.21,410.59
...
```

The test set file (test.csv) has the same structure except that the column for y is omitted:

```
Id,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10
10000,-483.79,1288.05,-129.87,-198.07,-334.48,-391.44,-612.40,-676.52,1327.22,-448.69
...
```

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

```
Id,y
10000,0.0
10001,1.0
10002,2.0
```

Note that, for each prediction, you need to include the same sample id (in the Id column) as specified in test.csv.

SUBMISSION FORMAT

For every data instance in the test set, submission files should contain two columns: *Id* and *y* where *y* should be a double with your prediction.

The file should contain a header and have the following format:

```
Id,y
10000,0.0
10001,1.0
10002,2.0
..
```

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

EVALUATION

The evaluation metric for this task is the **Root Mean Squared Error** which is the square root of the mean/average of the square of all of the error.

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

How to compute it in Python:

```
from sklearn.metrics import mean_squared_error
RMSE = mean_squared_error(y, y_pred)**0.5
```

GRADING

This task is ungraded and will not count towards your project grade. However, in order to prepare you for the graded tasks, we provide you with the following information on how the subsequent tasks, except for the next task, will be graded.

We provide you with **one test set** for which you have to compute predictions. We have partitioned this test set into two parts and use it to compute a *public* and a *private* score for each submission. You only receive feedback about your performance on the public part in the

form of the public score, while the private leaderboard remains secret. The purpose of this division is to prevent overfitting to the public score. Your model should generalize well to the private part of the test set.

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.