AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING אפקה המכללה האקדמית להנדסה בתל־אביב

Machine Learning 2022 Exercise 2

Topics:

- **EDA – Essential Data Analysis for Regression**
- **Linear Regression**
- **Cross Validation: LPOCV, K-Fold and LOO**
- **Regularization: Lasso Ridge and Elastic Net**
- **Feature Selection**

For this exercises you will participate in a Kaggle competition – "House Prices - Advanced Regression Techniques".

You will perform a regression task.

You are required to house prices in the city Ames, Iowa, USA

The competition dataset provide a large amount of features for you to play with. You are welcome to remove features you deem irrelevant, or calculate new features.

The data description is available at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data . Please make sure to read it carefully, make sure you understand the meaning of each column and how each feature might affect the price of a house.

If you have question about the competition or about the data, you are advised to refer to the discussion section of the competition: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/discussion.

You may take inspiration from the work of other participators, in the notebook section of the competition: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/notebooks.

Your predictions to the competition will be measured using the RMSE metric.

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

AFEKA אפקה המכללה האקדמית להנדסה בתל-אביב TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING

Requirements:

1. Your entire solution has to be written in a single python notebook (ipynb file)
2. Your submission to the Moodle has to include your solution file as an ipynb and as an html notebook (please refer to the attached guidance how to download your jupyter notebook as html) zipped into a single zip file.
3. How to download your ipynb as html: https://stackoverflow.com/a/64487858
4. When using Plotly Graphs, you need to make sure that they appear properly in the html. See Appendix 1.
5. The name of your files should begin with your full name and ID number.
6. The first MD –Mark Down cell of your submitted notebook has to include your full name, id number and a link to your Kaggle account.
7. You are required to analyze the data – EDA – Essential Data Analysis using graphs and tables. Use MD cells to explain your analysis.
    a. Analyze the effect of features on the regression result.
    b. In this section you may create new features.
8. You should refer to this exercise as a riddle, hence you should start with understanding the data and drawing conclusions. Afterwards you may begin solving the riddle.
9. Divide your Dataset using Cross Validation with random_state = 42.
10. For your regression task you are obligated to use Linear Regression (Linear Regression and or SGDRegressor).
11. Use Regularization techniques (Ridge, Lasso, Elastic Net).
12. Try different Features selection algorithms (Forward, Backward, Hybrid)
13. Explain and analyze your results.
14. Please notice that you are limited to up to 5 submissions a day.
    Therefore you should test your theories using cross validation, prior to submission of the results of a model which has been trained on the entire training set.
15. Show graphs of the loss function as a function of different values of hyper parameters.
16. Show graphs of the training loss (RMSE) and of the validation loss as a function of different hyper parameters.
17. You may compare different loss functions between the training and the validation and analyze and explain the differences.
18. At last, make a submission to the competition.

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

19. Attach a screenshot of up to your 10 most recent submissions. Emphasize your best submission. In addition attach a picture of your location in the leaderboard to your notebook.
20. Use a MD cell to summarize your work (5-20 lines), explain your work, what worked well and what did not work well.
21. Write a short TL;DR – To Long Didn't Read (5-10 lines) at the second MD cell of your notebook (under your full name and your ID number). Use the TL;DR to explain what you will be doing make, mainly what worked well.
22. Your last cell should hold references to resources you used including notebooks you took inspiration from, links, books etc.

Notebook Structure:

1. Your name ID Number and link to your Kaggle account.
2. TL;DR – explanation of the competition and what are you trying to do and how.
3. EDA – Essential Data Analysis
4. Experiments – different regularizations, feature selections, hyper parameters etc. (test your experiments using CV- Cross Validation)
5. Graphs of the loss and accuracy as function of different hyper parameters and analysis of the results.
6. Screenshot of up to 10 most recent submissions including your best submission and your place on the leaderboard.
7. Summary
8. References

Grade Structure

1. Simple, organized explained and clean code 10%
2. Organized, understandable and explained notebook 10%
3. Effort and self-learning 10%
4. Correct implementation of the requirements and valid notebook structure 70%
5. 10% bonus for extreme effort.

Remarks

1. You are adviced to try different CV methods: (K fold, Leave P out, etc.)

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

2. Chose Feature selection algorithms and regularizations that you think would fit for the data and the model, in addition make sure to explain your choices.
3. You should implement function in order to avoid code duplication.
4. Functions you implement should contain explanations of what they do.
5. Try and error is part of machine learning, make sure to understand what you are doing and why.
6. Make sure to show your understanding, what has worked and what hasn't worked and try to explain it.

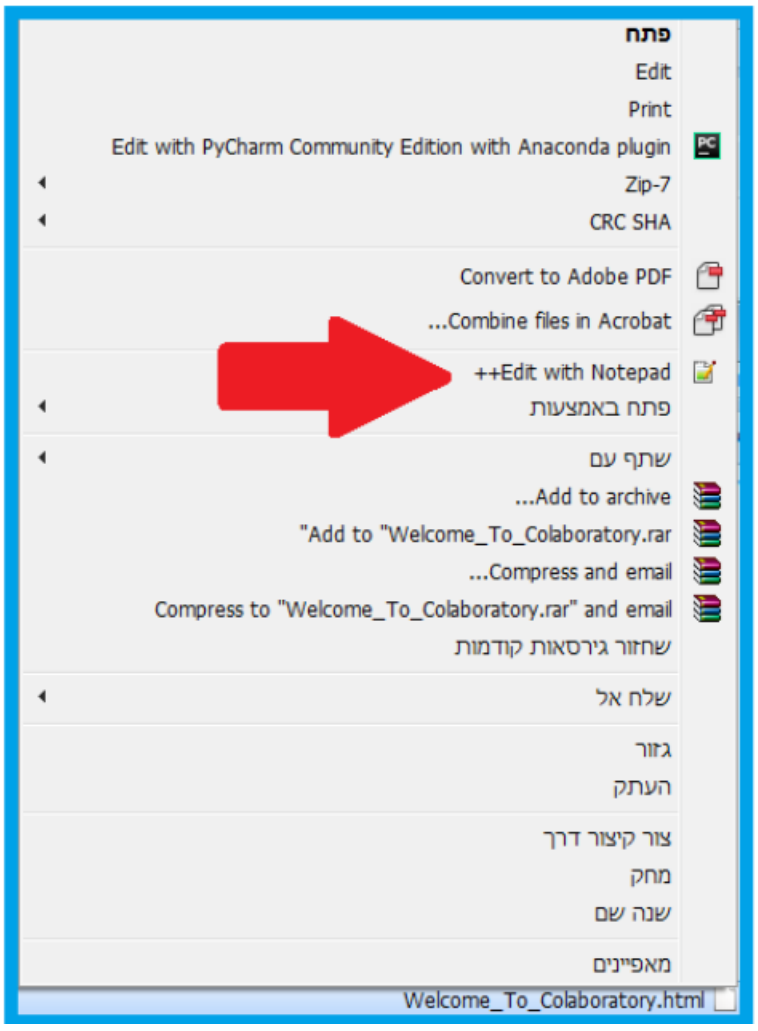
GOOD LUCK!

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

Appendix 1: How to show Plotly graph properly in the HTML file.

Given an html file which has been created from an ipython notebook with Plotly graphs, it is possible to force the plotly graphs to appear in the html as well by adding a script before the rest of the scripts in the html file.

You need to install notepad++ (freely available to download from their website). Once you downloaded **Notepad++.** Right click on your html file and chose: "**edit with Notepad++**" as follows:



Paste the following line:

<script src="https://cdn.plot.ly/plotly-latest.min.js"></script>

Before the rest of the scripts in the html file, as follows:

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi

Hit save and the graphs should appear in the html file as well.

Credit to Dima a former student of the course for this solution.

Written by: Ido Meroz, IdoMe@afeka.ac.il
Based on the work of Mr. Aviad Malachi