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

Machine Learning 2022 Exercise 1

Topics:

- **EDA – Essential Data Analysis**
- **Logistic Regression**
- **MLP – Multi Linear Perception**
- **Train, Validation, Test split**
- **Accuracy Metric**

For this exercises you will participate in a Kaggle competition – "Titanic Machine Learning from Disaster".

You will perform a classification task.

You are required to classify between to labels: survived or not.

The completion dataset provides a small amount of features to play with. You can remove features or calculate new ones based on the given features.

The data description is available at: https://www.kaggle.com/c/titanic/data . Please make sure to read it carefully, make sure you understand the meaning of each column and how each features is likely to effect the survival chance of a passenger.

In case 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/titanic/discussion .

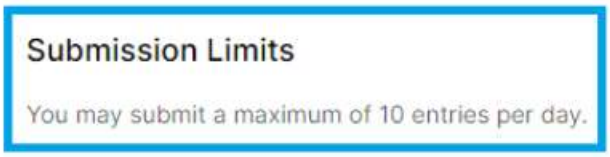
You may take inspiration from the work of other participator in the notebook section of the competition: https://www.kaggle.com/c/titanic/notebooks.

Your predictions to the competition will be measured using the accuracy metric, which is the percentage of passengers you correctly predicted rather the will survive or not.

Requirements:

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

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 a 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. The name of your files should begin with your full name and ID number.
5. 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.
6. 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 and there effect on the classification result.
    b. In this section you may create new features.
7. 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.
8. Divide your Dataset to train and validation using random_state = 42.
9. You may use logistic regression (both Sklearn LogisticRegression implementation and or the Sklearn SGDClassifier Implementation), MLP or both. Other models are not allowed to be used for this exercise.
10. Evaluate your model on the validation set before making a submission to the competition.
11. Please notice that you are limited to up to 10 submissions a day



    Therefore you should test your theories on the validation set prior to making a submission.
12. Try different hyper parameters, different sub groups of features.
13. Show graphs of the training loss and of the validation loss as a function of different hyper parameters
14. Show graphs of the training accuracy and of the validation accuracy as a function of different hyper parameters
15. Compare between the train and validation and make conclusions.
16. Use a MD cell to explain your comparisons between the train and the validation And your conclusion

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

17. Attach a screenshot of up to your 10 most recent submissions. Emphasize your best submission
18. 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) and a short summary on the bottom of your notebook. Use the TL;DR and the conclusion to explain your work.
19. 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 on different features, models and different hyper parameters
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

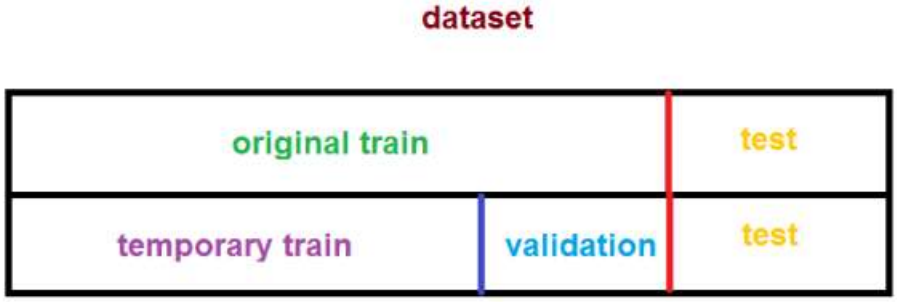
Grade Structure

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

Remarks

1. Your advised to split the data as follows:

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



2. Use the purple temporary train set as temporary train set for your experiments used to determine the optimal hyper parameter and the cyan validation as test set in order to validate your results.
3. Once you chose the hyper parameter you found optimal train again on the green original train and make a submission to evaluate your model on the yellow test set.
4. You may use different split sizes for your train validation set, explain shortly why you chose each ratio.
5. You should implement function in order to avoid code duplication.
6. Functions you implement should contain explanations of what they do.

GOOD LUCK!

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