Assignment-2

Decision Trees

Points: 100 **Due Date: 13/06/2021**

Submission Window: 11:00 p.m. to 11:59 p.m. (June 13, 2021)

Do not submit the assignment outside the submission window.

About the Data

The data has been split into two groups:

training set (train.csv) and test set (test.csv)

The training set should be used to build your machine learning model.

For the training set, we provide the outcome (also known as the "ground truth") for each passenger.

Your model will be based on "features" like passengers' gender and class.

The test set should be used to see how well your model performs on unseen data.

Even for the test set, we are providing the ground truth for each passenger, so that you can compute the required deliverables yourself.

But it is your job to predict these outcomes also.

For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

Variable Notes

pclass: A proxy for socio-economic status (SES)

1st = Upper

2nd = Middle

3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

Deliverables

The Python Notebook containing the complete, organized and well documented code with the following statistics:

Accuracy, Losses and F1 Scores for both Testing and Training datasets.

Instructions:

Create a Google Colab / Jupyter Notebook.

Explore the relationships between the output label and features. Plots are always helpful for such kinds of analysis.

Create a Decision Tree, which classifies the data into two classes (survived (1) or not-survived (0)).

(It is always helpful to plot losses and accuracy against these quantities to better understand how changing them is affecting your training process.)

(I recommend using Google Colab to avoid all the installation-related issues. You don't need to install anything. You can work right from your browser.

For more info, please go through this link: https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c)

How to submit the Assignments?

1. Go to https://github.com/vanshbansal1505/ICG-Summer-Program-2021-DS/ and fork this repository.
2. Push all the files that you want to submit in the **appropriate** folder of the forked repo. Name your folder "`<Roll Number>_<Name>`" (without quotes.) Strictly follow this convention in each of your submission otherwise it will be rejected.

   (For example, if I want to submit the files `x`, `y` and `z` for the n^{th} assignment, then I'll put them in folder named `190941_VanshBansal` and I'll put this folder in the `Assignment-n` folder of my forked repository.)
3. Create a pull request **only during the submission window**. All other pull requests will be rejected.
4. Before the next assignment, fetch upstream your forked repo.
5. Go to Step 2.

(If you are new to GitHub and are using

- Windows, look at this: https://www.youtube.com/watch?v=_NrSWLQsDL4
- Linux, look at this: https://blog.scottlowe.org/2015/01/27/using-fork-branch-git-workflow/

)

Honor Code of Ethics

1. Be Regular and complete all the assignments timely.
2. Read the instructions carefully.
3. Maintain the highest standards of honesty and integrity.
4. Use the internet to learn, not to copy.
5. Always cite the source if you are taking some idea or code snippet from there; otherwise, we will consider it as copying.

Any unethical/unprofessional conduct or violation of any of the above-mentioned points will be dealt with very strictly.