Phase 4: Feature importance and reduction (For Bonus Points)

The overall objective is to identify non-informative input features and remove them from the dataset.

Pre- Requisites:

- Before working on this phase, please practice “Activity 10”.
- Refer A recipe of supervised learning development
- Refer the “Instructions for Projects” under Module 0
- For this phase, you need to split your data into training and validation.
- This is an extension of Phase 1, Phase 2 and Phase 3 , so build and perform below tasks on the same idea that you started your project with.
- Refer the Example Report

Tasks

- Shuffle your dataset
- Split the dataset into a training set and a validation set.
    - For small datasets, selecting a random 20%-30% of the rows as the validation set and leaving the rest as the training set works well.
- Train the NN models by taking only 1 input feature/column at a time.
    - ***Method 1:*** You can also write a loop to implement this, which will take only one feature at a time. For example: if you have 6 input features/columns, the python loop which will take only one feature at a time and train all your models.
    - ***Method 2:*** If you have 6 input features/columns, train 6 models where each model only receives one feature at a time.
- Check the validation accuracies of these models that will indicate the relative importance of all the input features.
- Plot these validation accuracies in the form of a bar diagram. If all your accuracies are more than 80%, your plot’s y-axis should be limited to 80-100.
- From the previous step you have noticed the significance/importance of each feature. The feature that yields the highest accuracy is the most important feature.
- Train your model by removing least ranked features(the features which are giving you low accuracy)
    - ***Method 1:*** You can also write a loop to implement this, which will eliminate one feature at a time

    - ***Method 2:*** You can iteratively train various models by removing more and more unimportant features.
- Plot these validation accuracies of these models in the form of a bar diagram.
- Finally, compare your feature-reduced model with the original model with all input features and discuss the difference in accuracy.

Report must include the following:

- Report the accuracies of the models to include in your report.
- You can summarize your findings in the form of a table and the table should contain the accuracy and loss on the training set and the validation set
- You can also mention about the input feature that was highly ranked and least ranked.
- Your report should also include the plots/diagrams of your experiments

Final report

- Please submit a PDF of your final report.
- It should contain the important findings from each phase of your project.
- Your report should not be very long; 10/12 pages at most.
- Your tables and figures should be numbered and captioned (labeled) appropriately. Please resize the figures appropriately and ensure that none of your figures flow outside of the border.
- If you are copying images from a Notebook, please remember to turn off the ‘dark mode’ in Notebook before you copy images/plots.
- Your report should include abstract, introduction and conclusion (each 250 words minimum).
- Please also submit a link to your final Notebook along with the Overleaf PDF file.

Sample final reports

For your reference, I have listed below some final reports by students who took this course in earlier semesters. Please understand that these examples are only meant to be references and your focus should be on meeting the requirements mentioned above instead of preparing a report similar to these example reports.

- Report 1
- Report 2
- Report 3