Machine Learning (PG) M2021
Assignment-1

Max Marks: 100 **Due Date**: 24/09/2021, 11:59 PM

Instructions

- This assignment should be attempted individually.
- You can only use python as the programming language.
- Keep collaborations at high level discussions. Copying/Plagiarism will be dealt with strictly.
- Create a '.pdf' report that contains your approach, pre-processing, assumptions etc. Add all the analysis related to the question in the written format in the report, anything not in the report will not be marked. Use plots where ever required.
- Submit code, readme and report files in ZIP format with naming convention `A1_rollno_name.zip`. This nomenclature has to be followed strictly
- Remember to **turn in** after uploading on Google Classroom. No justifications would be taken regarding this after the deadline. If the deadline is 11:59 pm, submit before 11:58 pm to avoid late submissions.
- Start the assignment early. Resolve all your doubts from TAs in their office hours **two days before the deadline.**
- You should be able to replicate your results during the demo, failing which will fetch zero marks.

1. (25 points) **Data visualization** is an essential part of machine learning. In this question, you will use python libraries (such as matplotlib) to create different types of plots. You have to include the legends in the plots to denote the class labels and other relevant information.

   Datasets: `COVID-19| Face mask dataset|`

   1. Find the two CSVs present in the `COVID-19` folder and explore the files. Specifically, check the number of columns, their names, type of columns (categorical or continuous), and the possible data values/range for each column. Illustrate these in your report. **5 marks**
   2. Use the file `covid_19_india.csv` and show the trend of confirmed, and cured cases along with the number of deaths in India for the time period of the data in one plot. **5 marks**

3. Use the file `covid_vaccine_statewise.csv` and plot the total number of doses administered of Covishield, Covaxin, and Sputnik in the following cities: Kerala, Delhi, Rajasthan, Haryana, Uttar Pradesh, Tamil Nadu for the time period of the data. **5 marks**

4. Use the Face mask dataset for the following tasks:

   (a) Randomly select 5 images from each class and visualise them as images. **5 marks**

   (b) We can visualize only 2D or 3D data using scatter plots. For features dimensions higher than three, we may use T-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the number of features. Use the t-SNE to reduce the dataset to 2 dimensions and visualize the scatter plot. What is your inference regarding the class separation? **5 marks**

2. (30 points) For this question, you can use the decision tree classifier from sklearn.
Dataset: data (Use bank-additional-full.csv)
Target variable: As mentioned in the dataset description

Use the first 70% of the samples for training and the remaining 30% for testing. Implement the function to split the dataset and calculate accuracy. You cannot use any inbuilt version. Though, you can use Numpy, Pandas, Random, etc.

1. Take the Complexity parameter as a hyperparameter, and perform a grid search for finding its optimal value. You have to perform grid search for at least 10 values of the Complexity parameter. You have to implement calculating sum of impurity value of all leaf nodes for each complexity parameter value.
   Plot a curve between Complexity parameter and testing accuracy. Plot a curve between Complexity parameter and sum of impurity of leaf nodes. Comment on the effect of the Complexity parameter on the performance of the classifier. You have to implement grid search and cannot use any inbuilt implementation for the algorithm. **15 marks**

2. For part (1), prepare a table representing train accuracy and testing accuracy for each value of the Complexity parameter. Comment on overfitting, and underfitting for each entry in the table. **10 marks**

3. Replicate part (2) with sklearn Decision Tree Classifier's 'cost_complexity_pruning_path' function. Is there any deviation between the results from your implementation and the inbuilt function? **5 marks**

3. (35 points) For this question, you can use the decision tree classifier from sklearn.
Dataset: data
Target Variable: Month
You will have to handle null values in the data.

Split the data into training and testing set (75:25 ratio) using the function you created in previous question.
Use the same training set for training the following models. You can not use sklearn for splitting the dataset.

- Train a decision tree using both gini index and entropy. Don't change any of other default values of the classifier. In the following models, use the criteria which gives better accuracy on test set. **5 marks**
- Train decision trees with different maximum depths [2, 4, 8, 10, 12, 30]. Find the best value of depth by using testing and training accuracy. Plot the curve between training and testing accuracy and depth to support your analysis. **10 marks**
- Ensembling is a method to combine multiple not-so-good models to get a better performing model (more in upcoming lectures). Create 150 different decision stumps (max depth 4). For each stump, train it on randomly selected 40% of the training data, i.e., select data for each stump separately. Now, predict the test samples' labels by taking take majority vote of the output of the stumps. How is the performance effected as compared to part (a) and (b)? **10 marks**
- Now, try to tune the decision stumps by changing the max-depth [5, 7, 13, 15, 25, best achieved from (b)] and number of trees. Analyze the effect on the training and testing accuracy. Use majority vote for final prediction on the test data. **10 marks**

Compare the results of the classification models created above on the test set. Rank the models and analyze if there is a statistically significant difference. Add all the analysis to the report.