ASSIGNMENT - 1
Machine Learning - I
Topic: Evaluation Metrics

Total Marks : 60 marks + 10 report Deadline: January 29th, 2022

Instructions :

1. Do not copy from other students. Any case of plagiarism will result in zero marks.
2. Strictly follow the submission guidelines.
3. Allowed languages: python
4. Do not use any inbuilt library other than NumPy and matplotlib.

Submission Guidelines :

1. Submit .py python files for all the questions. If only .ipynb (notebook) files are submitted, we will not evaluate your submission.
2. Strictly submit a single report (.pdf) for all the questions. No .doc, .docx file will be accepted
3. If you are using colab, then attach your colab link in the report (preferred)
4. Submit a single zip file containing all python files and report.
5. The name of the zip file should be M21AIEABC.zip, python files should have the name M21AIEABC_qu1.py or M21AIEABC_qu2.py, etc. The report should have the name M21AIEABC.pdf (last three digits of your roll no in place of ABC). If the naming convention is not followed, we will award zero marks.

---

30 marks

Question 1

There are two models, M1 and M2, used to predict the scores for some input data. Suppose M1 predicts the score for input data as score1.npy and M2 predicts the score for the same data as score2.npy. Actual labels for a given score is label.npy
(use np.load to load .npy files)

1. Plot ROC curve (from scratch) for both the models in a single plot. **(10 marks)**
2. Explain which model performs better on this data and why? **(5 marks)**
3. Compute AUC for both the ROC curves. **(5 marks)**
4. Calculate true positive rate for both models when false acceptance rate is 10% **(5 marks)**
5. Draw your analysis on (3) and (4) **(5 marks)**

Note: Scores here represent the distance between two samples using two different models. 0 in the label represents similar samples and 1 represents different samples.

15 marks

Question 2

Dataset link: Link

Consider a fingerprint recognition dataset, having 600 images in the gallery and 9854 images in the probe. A model is used to classify probe images into 600 classes. The probabilities predicted by the model for all 600 gallery images are given in score.npy. The correct labels are given in label.npy.

1. Plot CMC curve up to rank 10. **(10 marks)**
2. Comment on the results **(5 marks)**

Question 3 **15 marks**

You are requested to solve the fruit classification problem based on the features in the given dataset using decision trees. Load this dataset for your decision tree classification problem. The dataset has 3 features and one target variable. The target variable takes either Papaya (0) or Banana (1). The features are "Size" in cm, "Weight" in kg, and "SkinColor" (100-green, 200-yellow, and 300 orange).

- Load (Train-Test Split) and prepare required packages and shuffle the dataset. **(2 marks)**
- Build and Train a DecisionTree classifier. **(5 marks)**
- Don't stick to a single configuration for your model. Try different hyperparameters. (At least 5) **(3 marks)**
- Test the model for each configuration **(5 marks)**
- Visualize the tree, evaluate it based on the metrics given in previous questions. **(3 marks)**
- Report the confusion matrix for your best model (don't use inbuilt function) **(2 marks)**

If the hyperlink doesn't work copy-paste the URL below -

https://drive.google.com/file/d/1O-Txgca54gFnocTszrYq3n7OIKnz502m/view?usp=sharing

If you have any doubts regarding the assignment, post on Google classroom.