CSE543/ECE563: Machine Learning (PG)
Monsoon 2022

Assignment-2 (60 points)

Release Time: September 23, 2022; 5:00 pm | Submission Time: October 10, 2022; 9:00 am

---

Instructions

- This assignment should be attempted individually. All questions are compulsory.
- **Theory.pdf**: For conceptual questions, either a typed or hand-written *.pdf* file of solutions is acceptable.
- **Code Files**: For programming questions, the use of any one programming language throughout this assignment is acceptable. For python, either *.ipynb* or *.py* file is acceptable. For other programming languages, submit the files accordingly. Make sure the submission is self-complete & replicable i.e., you are able to reproduce your results with the submitted files only.
- **Regarding Coding Exercises**: You can use modules from sklearn or statsmodels or any similar library for writing the code. Use random seed wherever applicable to retain reproducability.
- **Report.pdf**: Create a *.pdf* report of programming questions that contains your applied approach, preprocessing, assumptions, analysis, visualizations, etc.. Anything not in the report will not be evaluated. Alternatively, a well-documented *.ipynb* file with answers to all the questions may be submitted as a part of both code file and report.
- **File Submission**: Submit a *.zip* named A1_RollNo.zip (e.g., *A1_PhD22100.zip*) file containing *Theory.pdf*, *Report.pdf*, and Code files.
- **Submission Policy**: Turn-in your submission as early as possible to avoid late submissions. Expect **No Extensions**. Besides, submission within 10 min of the passing of the deadline will incur 20% penalty in the total marks of this assignment. Beyond this, late submissions will not be evaluated and hence will be awarded zero marks.
- **Resource Constraints**: In any question, if their is a resource constraint in terms of computational capabilities at your end, you are allowed to sub-sample the data (must be stratified). Make sure to exclusively mention the same in the report with proper details about the platform that didn't work for you.
- **Clarifications**: Symbols have their usual meaning. Assume the missing information. You are free to use any libraries and need not do anything from scratch unless specifically stated otherwise. Use Google Classroom for any queries. In order to keep it fair for all, no email queries will be entertained. You may attend office/TA hours for personal resolutions. No queries will be answered in Google Classroom comments when 12 hours or less are left for the submission deadline.
- **Compliance**: The questions in this assignment are structured to meet the Course Outcomes CO2, CO3, and CO4, as described in the course directory.
- **Institute Plagiarism Policy Applicable.** Both programming and theoretical questions will be subjected to strict plagiarism check.
- There could be multiple ways to approach a question. Please explain your approach briefly in the report.

---

1. **Data's Objective Identification** **(8 points)**

   (a) Exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Open Government Data (OGD) Platform India (https://data.gov.in/) is a recent initiative by the Government of India to release open source real statistics of national importance for data scientists and researchers to analyse.

Select any one tabular data from the website and perform EDA over it. You must describe the data in-brief and explain the possible use case of such a data, if analysed. Give a short (max. 5 bullets) description in your report. (3 points)

(b) Many ML-algorithms expect data to be scaled consistently. Perform normalization of the entire raw data and display its 10th row from the top before and after normalization. (1 point)

(c) Perform standardization over the entire raw data and display its 10th row before and after standardization. (1 point)

(d) In not more than 3 sentences each, describe which type of data is suitable for normalization and which type of data is suitable for standardization *(Hint: Think about their underlying assumptions.)*. Also, explain which of the two is suitable for your current data and why? (3 points)

2. **Data Augmentation** **(6 points)**

Generally, ML-Models need a lot of data to achieve generality and become robust. Data augmentation are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from the existing data. Perform the following image augmentations using the image data set by the name 'Images (BMP, 1.27GB)' available here https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52756988. Pick a set of 3 different random images from the data set for each augmentation task *[In case storing entire 1.27 GB folder is a problem, make a folder of first 10 images and then randomly select 3 images each time]*. You are free to select the heuristics and use any modules.

(a) Resize

(b) Pad

(c) Crop

(d) Gray Scale

(e) Contrast

(f) Saturation

3. **Logistic Regression** **(7 points)**

(a) For N-class instances, how many binary classifier models are required to be generated for (2 points)

- One-vs.-All multi-class classification using logistic regression
- One-vs.-One multi-class classification using logistic regression

(b) Download the MNIST Dataset (http://yann.lecun.com/exdb/mnist/). Visualize 2 images of each class. Use binary logistic regression to perform OVO and OVR over data set. Finally, print class-wise accuracy. You are free to choose other information such as evaluation metric and ratio of train-test split. (5 points)

4. **Generalized Linear Model: GLM** **(5 points)**

(a) Prove that the Gamma Distribution belongs to the same family of curves as Poisson Distribution. Be informed that both Gaussian and Bernoulli distributions also belong to this family. (5 points)

5. **Unsupervised Learning** **(9 points)**

(a) Nilesh Karnik is a student of data science at the University of California. He recently attended a lecture on clustering. In order to boost his thinking abilities, he often critiques the algorithms taught to him. On a publically available data, he applied KMeans using sklearn.cluster module. Thereafter, he partitioned the instances and calculated covariance matrix (S_i) separately for each group. When shared this approach with his professor, the professor pointed out two reasons for this algorithm to be not a good idea. Can you guess what those two reasons could be? (3 points)

(b) Use Fashion MNIST (https://paperswithcode.com/dataset/fashion-mnist dataset to apply a clustering algorithm each of the following types

i. Hierarchical Clustering

ii. Density-based Clustering

Use one appropriate evaluation metric for each of the above two and produce a train, val and test score. The split ratio will be 70:10:20. Which of the two algorithms performed better? Explain. (6 points)

6. **Classification Metrics** **(8 points)**

Annealing is a metallurgical process of heat treatment of a metal to improve its physical properties by allowing a slow cool down. Heat causes random rearrange of atoms in the metal, removing the weak connections and residual stress within. This idea was transferred to ML wherein annealing of different combinations of subsets of features is simulated. The set of features producing best metrics outputs is then chosen in the final model.

(a) "But this ship can't sink!" -Bruce Ismay
Use the idea of combinations as explained above to select the best subset of columns from the Titanic Dataset (https://github.com/datasciencedojo/datasets/blob/master/titanic.csv) to predict the possibility of survival of the passengers on the ship. You are free to use any classifier. Use ROC-AUC score as the performance criteria. (6 points)

(b) Plot the ROC-AUC curve for your best model above. (1 point)

(c) Did you find this approach greedy? Explain. (1 point)

7. **Thinking beyond what is written** **(7 points)**

Varun is a curious student. He thinks of designing a new evaluation metric 'F5-score'.

(a) Derive the formula of F5-score in terms of Precision and Recall in concurrence with the F1 score. What will be the corresponding value of alpha in that case? (2.5 points)

(b) What is your inference about the emphases on each (precision and recall) i.e., which of these is more emphasised now? Show. (2 points)

(c) Given the following data, use code to calculate precision, recall, F-0.5 score, F-1 score, and F-5 score. (2.5 points)
$y_{true} = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0]$
$y_{pred} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$

8. **Cross Validation** **(10 points)**

Download the standard 'iris dataset' from *sklearn.datasets* and perform classification (using any algorithm) over the data for each of the following cross-validation (CV) techniques.

- Monte Carlo Cross-Validation
- Leave P Out Cross-Validation
- Stratified 3-fold Cross-Validation
- Hold Out Cross-Validation

Compile the CV-score of all the four into a table and explain according to you which one is best suited for this dataset and why.