CSE343/CSE543/ECE363/ECE563: Machine Learning (PG)
Winter 2023

Assignment-2 Rubrics (UG: 30 + 10 (bonus) points || PG: 40 points)

Release: Feb 10, 2023 (Friday) Submission: Feb 19, 2023 (Sunday)

Instructions

- **Institute Plagiarism Policy Applicable.** Both programming and theoretical problems will be subjected to strict plagiarism checks.

- This assignment should be attempted individually. All questions are compulsory.

- **Theory [T]**: For theory questions, only hand-written solutions are acceptable. Attempt each question on a different sheet & staple them together (for ease of checking). Do not start a new question at the back of the previous one. Do not forget to mention the page number (bottom center) and your credentials (bottom right) on each sheet. It must be submitted in *Assignment submission box 4* kept outside B-609, R&D block. Scanned PDFs are not acceptable.

- **Programming [P]**: For programming questions, the use of any one programming language throughout this assignment is acceptable (Python/R/MATLAB). For python, you must submit a single *.py* file named as *A2_RollNo.py*. 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. Use random seeds wherever applicable to retain reproducibility. Further, save & submit (in the zip) the trained ML-model using either pickle or joblib.

- **Report.pdf**: Create a *.pdf* report of programming questions that contain your applied approach, pre-processing, assumptions, analysis, visualizations, etc.. Anything not in the report will not be evaluated. Alternatively, a well-documented *.ipynb* file (in addition to a single *.py* file mentioned in the previous bullet) with answers to all the questions may be submitted as a report. The report must be named as *A2_RollNo_Report.pdf* or *A2_RollNo_Report.ipynb*.

- **File Submission**: Submit a *.zip* named A2_RollNo.zip (e.g., *A2_PhD22100.zip*) file containing the report and code files.

- **Submission Policy**: Turn-in your submission as early as possible to avoid late submissions. In case of multiple submissions, the latest submission will be evaluated. Expect **No Extensions**. Besides, submission within 24 hours of the passing of the deadline will incur a penalty of 1 mark out of the total 6 marks allocated to this assignment. Submission, between 24 and 48 hours of the passing of the deadline, will incur a penalty of 2.5 marks out of the total 6 marks allocated to this assignment. Beyond this, late submissions will not be evaluated and hence will be awarded zero marks.

- **Clarifications**: Symbols have their usual meaning. Assume the missing information & mention it in the report. You are allowed to use any machine learning library until exclusively mentioned in the question that it is supposed to be done from scratch. You can always use basic python libraries such as numpy, pandas, and matplotlib, unless specified 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. Start your assignment early. No queries will be answered in Google Classroom comments 12 hours before the submission deadline.

- There could be multiple ways to approach a question. Please justify your answers mathematically in theory questions and via commented text in the programming questions appropriately. Questions without justification will get zero marks.

1. **[T ‖ CO1] Logistic Regression** **(4 points)**
   Given n number of training samples belonging to three classes where the feature vector is m-dimensional, derive the expression of the loss function, gradient, and Hessian of the loss function for a multiclass logistic regression. Write the expressions in the matrix notations.

2. **[P ‖ CO3] OVO and OVR** **(10 points)**

   (a) Download the Fashion-MNSIT dataset (https://github.com/zalandoresearch/fashion-mnist) and perform EDA (at least 4 techniques) on it. (2 points)

   (b) Use the binary logistic regression approach to perform one-versus-one (OVO) classification of the data from scratch. (3 points)

   (c) Use the binary logistic regression approach to perform one-versus-rest (OVR) classification of the data from scratch. (3 points)

   (d) Use the sklearn modules for OVO and OVR and generate the corresponding results. Do the results match? Discuss. (2 points)

3. **[P ‖ CO3] Multil-layer perceptron (MLP)** **(16 points)**

   - **Dataset**: Cell Image Disease Classification[1]
   - **Dataset description**: The dataset contains 27,558 cell images with equal instances of parasitized and uninfected cells.
   - **Task**: Develop a neural network to classify whether a cell is infected or not using the following architecture. In case of any ambiguity, please make relevant assumptions. Your assumptions should be clearly mentioned in your report.

     (a) Analyze the dataset, including but not limited to the following: (3 marks)

       i. visualization of two images of each class

       ii. check for class imbalance

       iii. Image size variability (& its remedy if necessary)

     (b) Divide the dataset into a ratio of 75:15:10 in a stratified manner, corresponding to train, validation, & test, respectively. (2 points)

     (c) Implement a MLP architecture with custom hyperparameters to classify the images into parasitized and uninfected cell images. (5 points)

     (d) During training, track the loss versus epoch plot for the train and validation set. Attach this plot in your report and give the analysis on the goodness of training. (3 points)

     (e) Print train, validation, and test F1 scores. Try to secure the best scores possible. (3 points)

4. **[T ‖ CO2, CO3] Cross Validation - Compulsory for PG students (Bonus for UG students) (10 points)**
   Download the standard 'wine dataset' from sklearn.datasets and perform EDA (at least 4 EDA techniques) over it, followed by classification (using any suitable algorithm) over the data for each of the following cross-validation (CV) techniques. You are allowed to use any library(ies).

   (a) Hold Out Cross-Validation

   (b) 4-fold cross-validation

   (c) Stratified 3-fold Cross-Validation

   (d) Monte Carlo Cross-Validation

   (e) Leave P Out Cross-Validation

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

[1] https://drive.google.com/drive/folders/1-tjbaqsCIxgd4tLJlp-B0xNjfCWXg7DZ?usp=sharing