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

Assignment-4 Rubrics (56 points)

Release: March 29, 2023 (Wednesday) Submission: April 10, 2023 (Tuesday)

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 *A4_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 *A4_RollNo_Report.pdf* or *A4_RollNo_Report.ipynb*.
- **File Submission**: Submit a *.zip* named A4_RollNo.zip (e.g., *A4_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. Kindly justify your answers mathematically in theory questions and via commented text in the programming questions appropriately. Questions without justification will get zero marks.

---

1. **[P || CO2 & CO3] Clustering** **(8 points)**
Load the 'Human Activity Recognition Using Smartphones Dataset' [1] and visualize it through UMAP plot. Implement the K-Means and spectral clustering algorithms. Using the elbow curve, find the optimal value of 'K' for each algorithm. Evaluate your results using an appropriate evaluation metric and present the test scores for each algorithm.

2. **[P || CO3 & CO4] Random Forest** **(10 points)**
Use the wine dataset[2] to build one multi-class classifier each using the following techniques.

   (a) Random forest

   (b) Decision Tree with Bootstrap Aggregation

   (c) Decision Tree with AdaBoost

   Which of the above techniques worked best? Use an appropriate evaluation metric. Explain the inferences of each technique.

3. **[P || CO3 & CO4] Ensemble of Methods** **(12 points)**
Use the 'Vehicle dataset from Car Dekho' dataset [3] to prepare a pipeline that performs an ensemble of the below models (built from scratch) to produce the final result. You are free to choose any ensemble approach.

   - Lasso regression
   - Ridge regression
   - Locally weighted linear regression
   - Regression decision trees
   - KNN regressor

4. **[P || CO2] Decision Trees** **(26 points)**
Apply decision tree classifier (allowed to use the in-built library functions) on the bank marketing dataset[4]. Use four different decision tree algorithms: ID3, C4.5, C5.0, and CART. You are free to choose any library. For each algorithm, take the complexity parameter as a hyperparameter, and perform a grid search (from scratch) for at least 10 different values of the complexity parameter to find its optimal value. Prepare a table representing train accuracy and testing accuracy for each value of the complexity parameter. Now, replicate the above table using the sklearn Decision Tree Classifier's 'cost_complexity_pruning_path' function. Draw inferences on the results obtained with Comments on the (1) different algorithms used and (2) the deviations between the results from your implementation and the inbuilt function.

[1] https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html#sklearn.datasets.load_wine
[3] https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho
[4] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing