AI 511 – Machine Learning

Assignment 1

Only for students of Anurag P's TA group

---

1 Datasets

Please find 2 datasets in this folder.

- `bank_auth.csv` is the dataset for classification
- `garments_worker_productivity.csv` is the dataset for regression

1.1 Classification

The dataset `bank_auth.csv` contains data to classify whether a bank note is authentic or not. The dataset was created by extracting features from images of bank notes. The features extracted are listed below:

- `variance` of Wavelet Transformed image
- `skewness` of Wavelet Transformed image
- `curtosis` of Wavelet Transformed image
- `entropy` of image

Use accuracy as the metric for evaluation.

1.2 Regression

The dataset `garments_worker_productivity.csv` contains required data as features to predict the `actual_productivity` of the workers in the factory. Few of the features, which are not self-explanatory, that are used to predict the `actual_productivity` are listed below:

- `smv`: Standard Minute Value, allocated time for the task
- `wip`: Work in Progress, the number of unfinished items for products.
- `idle_men`: The number of workers who were idle due to production interruption.

Use Mean Squared Error and Mean Absolute Error as the metrics for evaluation

2 General Instructions

- Use basic preprocessing techniques taught in the last week.
- Code for training and testing models must be completely written from scratch using the Numpy library. **Do not use Scikit-Learn's models and functions for designing models, their training, or testing. `numpy.gradient` is not allowed either.**
- Include code for Univariate and Multivariate Linear Regression in closed form, Gradient Descent, Newton's method for optimization.
- Include code for Logistic Regression using Gradient Descent, Newton's method for optimization.
- Include code for Naive Bayes for Univariate Gaussian.
- Explore if you can take multiple columns and use Multivariate Gaussian for classification. Please mention the reason for choosing the columns.
- Use randomly generated data to check your implementations.
- Try to generalize your functions so that you can use it for any kind of dataset with a different number of features. In other words, code your functions in `np.array` so that the function runs irrespective of the length of the input.
- Leave segment of your data so that you can use it for testing your model. You can use `train_test_split` function.

3 Submission Details

- **Deadline : 3rd October 2021, 11:59 PM**
- Submit a **zip** file with the following contents:
  - Jupyter Notebooks for each of Regression and Classification
  - Code should be completely working and TAs should not need to modify to run the code.
  - A PDF report elaborating on your approach, assumptions, reasonings, observations, and conclusions.
- Try to include a variety of approaches. Just loading the data and running models is insufficient. More marks would be awarded to students who would have put in a lot of effort in trying out various ideas, visualizing data, etc.
- **This is an individual assignment. Although discussions with your classmates and friends are allowed, plagiarising each other's code is strictly prohibited. Plagiarism is a serious offense and infliction of harsh penalties is bound to happen.**

...