Course ID: CSE-506
Course Name: Data Mining (DMG)

Assignment 3

September 7, 2022

Instructions

1. The assignment is to be attempted in groups, ranging from one to five students.
2. Programming Language: Python.
3. For Plagiarism, institute policy will be followed.
4. You need to submit the report.pdf, readme.pdf, source code files, images, model files and **PPT**.
5. Report, models and code in .py format should be submitted in the Google Classroom in a zip folder with the name 'A3_Rollnumbers separated by underscore.zip'.
6. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
7. One member should submit on Google Classroom while other members can mark turned in without the attachment.
8. In case of doubts, please comment on Google Classroom.
9. You need to **make a presentation** of up to 20 slides in which the first slide should **clearly mention the contribution of each group member**. Penalty will be imposed on slides' limit violation.
10. **You should be well aware of theory of the algorithms mentioned and chosen by you. You will end up losing marks in case you haven't prepared well**.

Extension and Penalty clause: You can submit the assignment till t+3 day with penalty. Submitting within t+2 day will attract 10% penalty and submissions after t+2 and before t+3 will attract 20% penalty. Any submissions after t+3 will be not be considered. Even a 1 minute late submission on google classroom will be considered as late. Please turn-in your submissions atleast 5 minutes before the deadline.

Total Marks: 80 (will be scaled later)
Deadline: November 20 11:59:59 pm
t + 2 days: October 21 (10% penalty)
t + 3 days: October 22 (20% penalty)
t + 4th day onwards: October 23+ (0 marks)

In this assignment, you are required to perform anomaly detection on a dataset. You will need to show that if you detect and remove anomalies from your data, you can get better results for downstream tasks.

1. Datasets for Anomaly Detection are present in the following GitHub repository.
2. Download **one** of the **Categorical Datasets** present from the repository with **at least** 50 dimensions.
3. On the chosen dataset, do the following -
   (a) **(5 points)** Perform an analysis on the data to present its characteristics. Show **at least** 5 different statistical measures on the data and present them in an infographic.
   (b) **(5 points)** Train a machine learning model of your choice to establish baselines on the data.
   (c) **(20 points) Dimensionality Reduction** transforms the data into a lower dimension retaining the most important properties. Choose 3 algorithms for Dimensionality Reduction and implement them on the data. On each of the algorithms, find and remove the anomalies from the actual data and retrain the machine learning model from part (b). Create an infographic with the results and analysis for the results. There are 5 points for each algorithm implementation and another 5 for the analysis. You may choose algorithms taught in class, or other popular algorithms, or from recent papers.

(d) **(20 points) Clustering** is a technique to represent similar data close to each other. Choose 3 algorithms for Clustering and implement them on the data. On each of the algorithms, find and remove the anomalies from the actual data and retrain the machine learning model from part (b). Create an infographic with the results and analysis for the results. There are 5 points for each algorithm implementation and another 5 for the analysis. You may choose algorithms taught in class, or other popular algorithms, or from recent papers.

(e) **(20 points)** Very similar to Clustering, **Classification** also tries to bring similar data closer, but with data for which we know the actual labels. Choose 3 algorithms for Classification and implement them on the data. On each of the algorithms, find and remove the anomalies from the actual data and retrain the machine learning model from part (b). Create an infographic with the results and analysis for the results. There are 5 points for each algorithm implementation and another 5 for the analysis. You may choose algorithms taught in class, or other popular algorithms, or from recent papers.

(f) **(10 points)** Create a report with all your findings, difference between algorithms, their shortcomings, hypothesis to mitigate them, final analysis and your learnings from this assignment in a pdf format.

N.B.

1. No algorithm can be repeated. That it, you need to 5 statistical measures, and 3 algorithms for each, or 9 of the anomaly detection techniques, making 14 algorithms in total.
2. You may use any form of infographic that you feel represents your results and analysis best.
3. Detailed explanation of assumptions made for solving the mentioned problems.
4. Provide the various parameters and hyperparameters used.
5. Your zip file should contain a folder “visualizations”, “code”, “models”, report.pdf, a readme.pdf file and a ppt.
6. The folder “visualizations” should contain all the images and folder “code” should contain all the codes including notebooks if used.
7. Make a section “Learning’s” in the report.pdf and describe your learning’s from this assignment.
8. In the readme.pdf file, you will have to mention steps to recreate all your experiments and results.
9. Please provide references to all the sources including but not limited to libraries, GitHub Repositories, Research Articles, Blog Posts used in completing this assignment.