Final Project CS2810 Fall 2021

Proposals Due December 7th 11:59pm

Project Due Tuesday December 14th at 11:59pm

Overview: For the final project you will implement and analyze a data mining algorithm that does clustering or classification. Algorithms can be chosen from the list or top 10 Data Mining Algorithms Paper or you may propose your own algorithm as long as it is commonly used in practice. Algorithms must be based on the linear algebra from class. To analyze the algorithm you should use publicly available datasets from locations such as Kaggle. You will also write up your findings in a short document.

Groups: Students may work in groups up to 2 members, but the work must correspond to a two person project. Details are outlined below. Group members should be designated in the proposal.

Algorithms: For the algorithms you should not use any supporting packages, but develop the algorithm on your own. You may use packages like numpy and the like, but not scipy. If you are unsure of what packages you can use please ask. You must develop at least one algorithm. If you wish to do an algorithm that is not classification or clustering you should document this in your proposal. If you are in a group of two people you should implement two different algorithms and compare them. They should both do the same task clustering or classification.

Datasets: You should use at least two or three different datasets for your algorithm. These datasets should be publicly available. Ideally, but not required, you would have an example dataset for which your algorithm performs well and another for which it does not perform well.

Analysis: You should use some metrics to determine the success of your algorithm. For classification or clustering you can use F-Score and Jaccard from class. You should implement these metrics as well.

Deliverables: You should submit your code and a pdf document. You will also submit a pdf proposal.

Proposal: Due December 7th. A pdf which should state which algorithm or algorithms you have chosen along with your group members if any. Each person should submit their own proposal. You should list possible datasets you would like to test. Please note that the datasets can change, but the algorithm chosen should not change without prior approval.

Final Written Document: This document should be a pdf and turned in with your code. It should explain what algorithm you implement. It should list what datasets were used and how to download and use them with your code. Your document should have a code section which explains the design and flow of your code. The document should also have a results section which details all results using statistic measures listed above. Finally, it should have a conclusion which is your own opinion or reflection on the algorithm and its use.

Code: All code should be written in python and clearly documented. You should make sure to break code into appropriate functions. Style is a factor in grading. Clear short functions that do one thing are best. Code should be reproducible, we should be able to run your code on your datasets and get the same results from your document. Your algorithm should be encapsulated in that we should be able to run your algorithm on a different dataset not specified and get accurate results.