Project Guidelines

DS540 Data Mining: Spring 2023

Project deliverables

The complete grade of the project is divided among the following deliverables:

1. Report (40%)
2. Slide and In-class presentation (40%)
3. Code (20%)

Team

For this course project you are required to work in teams of 2 to 4 students. You have the flexibility of forming your own teams. If you are unable to find team members please let me know.

Problem/Dataset

Data Mining is considered to be one of the most important aspect of data science. For this course project, the specific problem you solve should be ideally based on your own interest. Basically, you have 3 options

1. **You bring your own problem**: In case you are already working on some project and you already have access to some dataset, then I would be more than happy to consider that dataset for this project and we can formulate a specific statistical learning problem which is based on that dataset.
2. **Get datasets online**: There are multiple websites where researchers share datasets publicly which can be used by the community for testing algorithms. Some examples:
   - UCI
   - Kaggle
3. **Talk to me**: My current research interest is broadly based on understanding the mathemtical foundations of deep learning algorithms. Besides that,

I also collaborate with other engineers, scientists and doctors on a variety of data science related problems. You can look at some of my recent papers to learn more about my research interests:

- Paper 1
- Paper 2

Ideally, if we are able to get good results, this project would lead to a conference/journal publication.

Methods/Algorithms

For this course project, I want you all to explore any one of the following standard algorithms in the field of data mining (choose either Supervised or Unsupervised learning and then explore as many algorithms as you want):

1. **Supervised Learning:** Here you have the option to explore various well known approaches such as:
   (a) Linear Regression
   (b) Ridge Regression
   (c) Lasso Regression
   (d) Logistic Regression
   (e) Decision Trees/Random Forests
   (f) Bagging
   (g) Boosting
   (h) K- Nearest Neighbor Classification
   (i) Support Vector Machines
   (j) Neural Networks
2. **Unsupervised Learning**: For this class of problems you have options such as
   (a) Apriori algorithm
   (b) K-means algorithm
   (c) Anomaly detection
   (d) Unsupervised Deep learning

Please note that the above mentioned algorithms are just suggestions. Please feel free to look into other supervised/unsupervised learning based algorithms as well.

Structure of the report

You are expected to use Python programming language for this project. Your report will have at max 20 pages and will contain the following 6 sections:

1. **Abstract:** A very short description of your project (5 to 10 lines)
2. **Introduction:** A detailed description of the project with related literature published online. Give a description of the problem you are solving.
3. **Methodology:** Give a detailed explanation of the method you propose to solve the problem under consideration.
4. **Results:** Using the method described in the Methodology section, present the obtained results (plots,tables etc) here.
5. **Discussion and Conclusion**: Summarize what you learnt from the results you presented in the previous section. Discuss the implications of these results (how will they be useful to people). Additionally mention any possible future work that might be able to improve upon the results you presented in this report.
6. **References**: Include the references to all research papers, websites or any other source that helped you complete the project. Try to be as exhaustive as possible here.

Structure of the presentation

The contents of your presentation should be compatible with your project report. You will be given 15 minutes to talk about your project in class. Your presentation needs to have following sections:

- **Title**: 1 slide
- **A short introduction of the team**: 1 slide
- **Introduction of the project**: 1-3 slides
- **Methodology**: 3-5 slide
- **Results**: 2-4 slides
- **Conclusion**: 1 slide
- **References**: 1 slide