College of Liberal Arts and Sciences
Department of Computer Science and Physics
CSC 410 Data Engineering

Group Project: 20 points

Purpose: This is a team project with maximum two students in each team. The project should involve analytics on real-world data of significant size and must be approved by the instructor. The goal is to prepare you to apply machine learning algorithms/models to real-world tasks and to qualify you for AI or machine learning research or related higher study.

There are few key deliverables which are described below. It is expected that each group member will contribute equally in the project. Only one group member needs to submit assignments, but both students' names must be included in submissions. Please do not make separate submissions.

1. Task 01- Project Proposal:

Proposal Submission Due: 02/16/2020
Proposal Presentation Date: 02/12/2020
Proposal Submission Points: 3 points for submission, 1 point for presentation, total of 4 points

Your first task is to come up with a project topic.

- Pick an application project that interests you and explore ways that machine learning algorithms can be applied to solve it.
- You can also pick an algorithmic/theoretical project that develops a new algorithm or a novel variant of an existing algorithm or proves some interesting properties of an existing algorithm. You can also combine elements of both theory and application.
- You can replicate others' experimental results to learn. But for your project, you should use the techniques and datasets on other applications or do some analysis of how the components of the model contribute to the resulting performance.

Generally, in an application project, you will have steps as follows: data collection, data visualization and preprocessing/normalization, feature extraction, model creation (training and testing occurs here), and model evaluation. For machine learning algorithms, you may choose any combination of: support vector machines, neural networks, random forests, logistic regression, k-nearest neighborhood tests, Naïve Bayes models, Bayesian Gaussian mixture models, Gaussian mixture models, decision trees, isolation forest, k-Means tests, Bagging, AdaBoost, hidden Markov models, deep learning, etc.

You should pay attention to the final project due date when you finalize your project topic. Please visit my office if you want to discuss about your topic before submission or if you need help finding a topic. Once you have identified a topic for your project, search for existing research on relevant topics by searching related keywords in Google Scholar, or other scientific research sources.

Your project proposal should include the following information:

1. The final choice of team members.
2. The motivation of your project? What problem are you going to deal with? Is it an application/theoretical/algorithmic project?
3. What particular data repository you are going to use and why? What is the type of data? (i.e. structured, semi-structured, quasi-structured, or unstructured) Perform some exploratory analysis of the data. Show some example data, show some plots or other visualization of the data. If possible, discuss your hypothesis about the data and what you hope to learn from the data. Describe the analytical tools you might use to explore the data.
4. In regard to your proposed method, what machine learning techniques are you planning to apply?
5. What is your plan for running your experiments? How will you evaluate the performance of your model/algorithm?

Your project proposal should address the questions discussed above. There is no specific format in which to submit your proposal, but it is recommended that the first page in your submission (word/pdf file) should be a cover page with the project name and group members' names. You do not need to submit your project proposal presentation slides.

Grading Rubric: The project proposal is mainly intended to make sure you decide on a project topic and get feedback from your instructor. As long as your proposal answers questions 1-5 as listed above and the project seems to have been thought out with a reasonable plan, you should get maximum points. Each group member should participate in the presentation to receive the proposal presentation point.

2. Task 02- Mid Project Submission:

Mid Project Submission Due: 03/16/2020
Mid Project Presentation Date: 03/11/2020
Mid Project Submission Points: 4 points for submission, 1 point for presentation, total of 5 points

Your mid project submission should include the following information:

Introduction (half page): what problem you are going address, what is the setting you are considering, and why it is important. Discuss the motivation for pursing this specific problem. Include some background if necessary. Discuss your input data, which algorithm(s)/model(s) you have tried and why, and present output from your preliminary experiments (if any). Try to be very explicit and also reference the source of your dataset (if any).

Related Work (max one page): You should find some existing related works, discuss their approaches, strengths, weakness, how they are similar to and differ from your work. Give your own opinion. (do you like or dislike any of their approaches? If so, why?). If you are using any of their datasets, explain what issues you aim to resolve or aspects you plan to improve on. Make sure to note if previous researchers have tried to solve the same problem, and, if so, also note which dataset and algorithms/model were used. You should have at least 5-7 scientific references in the related work. You can use Google Scholar to find and cite references. Use IEEE citation format.

Dataset and Features (max one page): You should describe your dataset: Is it publicly available? (include citations) How have you treated the data? How was it split into training and testing samples? Have you done any preprocessing, normalization, missing data handling or data augmentation? Does your dataset have any special characteristics? If possible, create some visualizations from your dataset. Next, discuss what features you have extracted and which of those features you have used in your experiment. Also note what techniques you have used for feature extraction. Include examples of your features if you have enough space.

If you have done some preliminary experiments, describe the experiment that you have run, the outcomes, and the next steps that you are considering. You should have tried at least one baseline experiment at this point.

References (No page limit): You should include citations for any papers you have mentioned in the related work section and for the algorithms and libraries (such scikit-learn or Matlab toolboxes) you have used in your work. There is no limitation on the number of references, but your instructor expects that you include at least 5-8 scientific research paper.

Your mid project submission should address the questions discussed above. You should use IEEE manuscript template (US letter, Times New Roman, double column, font size 10) to write your mid project submission report. You have to submit your report in Microsoft word format. You do not need to submit your presentation slides.

Grading Rubric: The mid project submission is mainly intended to make sure you are on track and get feedback from your instructor. You should write it as if you are writing the first few pages of your final project report, so that you can reuse most it in your final report. As long as your mid project submission follows the instructions above and you are on proper track to complete your project, you should get maximum points. Each group member should participate in the presentation to receive the presentation point.

3. Task 03-Advanced Analytics-Technology and Tools: Presentation

Presentation Date: 04/22/2020 or 04/27/2020
Presentation Points: 3 points

Each group should complete an independent study on an advanced analytics-technology and tools such as Hadoop, MapReduce, NoSQL, etc., conduct a lecture on the selected technology and tool, and demonstrate a working example. Each group will be allotted 30 minutes for the lecture and demo.

Grading Rubric: You should create and submit presentation slides of your lecture. The grade will be based on the quality, clarity, and technical content of your slides and lecture. Demonstrating a working example can help you to get maximum points. Each group member should participate in the presentation to receive the presentation points.

4. Task 04-Final Project Submission:

Final project submission Due: 05/06/2020
Final project presentation Date: 05/07/2020
Final project submission: 6 points for submission, 2 points for presentation, total of 8 points

Your final project submission should include the following information:

Abstract (one paragraph): The abstract should consist of one paragraph describing of the motivation for your paper, a high-level explanation of the methodology you have used, and the results of your experiment.

Introduction (half page): Explain the problem that you are going to address in your project report and why it is important. Discuss the motivation for pursing this specific problem, including some background if necessary. Discuss your input data, which algorithm/model you have used, and the output of your experiment. Try to be very explicit and also use the reference for your dataset (if any).

Related Work (max one page): You should find some existing related works, discuss their approaches, strengths, weakness, how they are similar to and differ from your work. Give your own opinion. (do you like or dislike any of their approaches? If so, why?) If you are using any of their datasets, explain what issues you aim to resolve or aspects you plan to improve on. Make sure to note if previous researchers have tried to solve the same problem, and, if so, also note which dataset and algorithms/model were used. You should have at least 5-7 scientific references in the related work. You can use Google Scholar to find and cite references. Use IEEE citation format.

Dataset and Features (max one page): You should describe your dataset: Is it publicly available? (include citations) How have you treated the data? How was it split into training and testing samples? Have you done any preprocessing, normalization, missing data handling or data augmentation? Does your dataset have any special characteristics? If possible, create some visualizations from your dataset. Next, discuss what features you have extracted and which of those features you have used in your experiment. Also note what techniques you have used for feature extraction. Include examples of your features if you have enough space.

Methods (max 1.5 pages including figure if any): In this section, discuss your machine learning algorithms/model, your proposed algorithm or mathematical proof (if any). Briefly discuss how the algorithm works so that it may be understood by others. Use proper citations when discussing existing algorithms/models. Include your novel approach (if any), making sure to note how the model works (with proper model architecture).

Experimental Results (max 2 pages including figure if any): In this section, discuss your parameters, training sets, and testing sets of your dataset. Include your validation techniques: If

there is cross-validation, how many folds? Note the accuracy of your experimental results, confusion matrix or AUC curves, and the equal error rate. You can briefly discuss your primary metrics such as accuracy, precision, AUC, EER before you explain your results. You should have a performance comparison table comparing your works/models against other existing works/models.

Conclusion and Future Work (one or two paragraphs): Summarize your works and recap key points. Why did you get better/worse performance than others' works? What are the reasons that some algorithms work better than others? What are the limitations of your work? How can you extend your work in the future to solve other problems?

References (No page limit): You should include citations for any papers you have mentioned in the related work section and for the algorithms and libraries (such scikit-learn or Matlab toolboxes) you have used in your work. There is no limitation on the number of references, but your instructor expects that you include at least 5-8 scientific research paper.

Appendices (optional): Include your code and/or additional proofs/algorithms that are important but were not included in methods section.

Grading rubric: keep in mind that, the above guidelines are not a grading rubric. This will help you to structure your report and guide you as you finish up your projects. You will not get a certain grade completing all the sections discussed above. Your final report/work will be judged based on clarity of the report and the technical quality and significance of your work.

Deliverables: 1) Submit your final report in both word and PDF format. You should use IEEE manuscript template (US letter, Times New Roman, double column, font size 10) to write your report. 2) PowerPoint presentation slides 3) Source code, dataset, and instructions to run the experiment in a zip file.

These guidelines are modified from Stanford's CS229: Machine Learning course project report guidelines.