Final Project Requirements

ECE 498 DS, Spring 2020

Final projects enable graduate students in the course to delve deeper into the process of data analytics by addressing challenging real-world problems. The important aspects in solving any data science problem are as follows:

1. Defining the problem and problem statement
2. Getting access to data
3. Pre-processing and cleaning the data
4. Summarizing the data; feature-engineering
5. Building a model to solve the problem
6. Evaluating the model
   - Includes comparing with existing methods and solutions
7. Discussing results and implications
   - Includes identifying limitations and directions for future work

For this project, students are expected to perform each of the above steps while working in a group of three (usually in the same group that the course mini-projects are done with). They are encouraged to work closely and brainstorm with Professor Iyer and the teaching staff to ensure timely progress in the right direction. The project amounts to 1 credit hour, so we expect commensurate effort. For projects that reach a certain level of quality, we can work with you to get them published.

Problem statement:

Before you can begin working on a project, you need to define a problem that you wish to address. We recommend three different avenues for finding relevant project topics:

1. Self-motivated project

   Do you have a problem in mind that you are interested in solving? For example, you could choose something that is close to your research. Or something that you want to try out just for the fun of it. Feel free to propose such an idea for the project. However, keep in mind the following:

   1. The project must be new (and **not** something that you are already working on)
   2. You must get access to the data yourself

   To find datasets, UCI Machine Learning Repository is an excellent source.

2. Mini Project 1 and 2

   You can propose a meaningful and reasonable extension to either Mini Project 1 or Mini Project 2. Having completed MP1 and having started MP2, you are familiar with the datasets and the questions that are answered (or not answered…) with them. What are some other questions that can be answered with this dataset? What about if we augment it with data from other sources?

3. Online Competitions (Kaggle)

According to its website, Kaggle is "the place to do data science projects". Organizations can host data science challenges/competitions on Kaggle by sharing (i) a problem statement of interest and (ii) data with which to solve the problem. Groups of people from around the world can compete in the challenge to solve the problem within a given time period. There are many stimulating challenges presented on the website. Below we have shortlisted a few that we thought might be interesting and relevant.

   a. Challenge 1 – Accelerometer Biometric Competition
   b. Challenge 2 – Eye Movement Verification
   c. Challenge 3 – Bike Sharing Demand
   d. Challenge 4 – How much did it rain?
   e. Challenge 5 – Home Credit Default Risk

You are expected to choose the data from the challenge of your choice. Note that these competitions are closed, which means that existing solutions to the challenges are publicly available. Study these solutions and then pose your own problem statement (different from the original one of the challenge) that can be answered with the data.

You are not allowed to pull code directly from Kaggle – you must present your own novel problem statement and build and test your own model as part of your final project. The Kaggle project can be a guide to finding a problem, but it should not define your final project…

Regardless of how you choose your project, **we would strongly recommend that you choose a dataset that you can get started with quickly** so that you don't need to spend a significant amount of time pre-processing or understanding it. The focus of your project should be on creating a good model and interpreting and evaluating its performance and implications.

Here are some titles from the final projects in ECE 498 DS last year. Note that groups last year only had ~4 weeks to complete a final project, whereas this year students will have ~9 weeks.
- *Book Cover and Title Classification*
- *Histopathologic Cancer Detection*
- *Boston Crime Data Analysis*
- *Eye Movement Verification and Identification Competition*
- *Africa Soil Property Prediction and Clustering*
- *Human Activity Recognition using Hidden Markov Model*
- *Predicting Time Signature in Bach Chorales*
- *Data Analysis of Microsoft Malware Datasets*
- *Multiple Models on Autonomous Vehicle Disengagements in Integrated and Dynamic View*
- *Tomato Price Prediction with Hidden Markov Model*
- *Recommendation System Based on Customer Review*

Timeline and Deliverables:

<u>Timeline</u>

Feb 26 – Propose two ideas for the project (one paragraph for each project)

Feb 26-28 – Meet with course staff to discuss about project ideas and finalize one project (signup info will be released on Piazza)

Mar 6 – Submit project proposal (not the same as proposing two ideas; see below for details)

Mar 27 – Submit project progress report 1

Apr 17 – Submit project progress report 2

May 8-10 – Final project presentation

May 12 – Final project submission (includes report, presentation and code)

<u>Project proposal</u>

2-page report that includes background and motivation for the problem statement, description of the data, related work, and solution plan.

<u>Project progress report 1</u>

2-page report describing (i) the methods, (ii) the results (if any), (iii) the current challenges with the project, and (iv) the plan for the upcoming weeks. Any deviations with the dataset and/or original problem from the project proposal need to be specified here as well.

<u>Project progress report 2</u>

3-page report describing (i) all updates since the first progress report, (ii) the current model that has been constructed for the dataset, (iii) the current performance of the model, (iv) the current challenges with the work, and (v) the plan for the upcoming weeks until the final report and presentation.

<u>Final report</u>

Up to 8-page IEEE or ACM conference style report that includes (i) introduction, (ii) related work, (iii) methods, (iv) results and (v) conclusions. Results should include a thorough evaluation of the proposed model and interpretation of results. Also include details about individual contributions in the group and a section on challenges faced in the course of the project. This report should be completed using Latex.

<u>Final presentation</u>

15 min presentation (10 minutes slides + 5 minutes QnA).

We will give a template for the slides as we get closer to the date of the final presentation.

Grading:

- Project proposal – 15 points
- Project progress report 1 – 10 points
- Project progress report 2 – 15 points
- Final presentation – 30 points
- Final submission (report, presentation, code) – 30 points

Miscellaneous:

1. All submissions are due by 11:59 PM on the given date. Late submission policy applies to all deadlines.
2. All deliverables/reports are to be typed. Unless otherwise specified, all reports should be completed in Microsoft Word (or equivalent) with single-spaced line spacing and size 12 Times New Roman font.
3. Submissions are to be made on Compass2G.