Final Project

This document provides you instructions about how to complete the final project. Please read it carefully! Final project is **30% of your grade**! ☺

Due Date: **Check the "Components of the Final Project" section for due dates.**

Goals

Your final for this class will be an ***original*** data mining project of your choice. The project will be a *theoretical* paper with a set of *empirical* applications. You will work on this project in a team of **~4** people. There are five goals that you will need to achieve:

(i) Introducing a new set of (at least four (4)) 'technically complex' methods of your choice from the world of data mining/machine learning that have not been covered in class syllabus

(ii) Explaining the philosophy and the mathematical ideas behind the set of methods that you have chosen in detail by providing formulas/pseudocodes/code examples

(iii) Showing how the methods can be applied by using at least four (4) different datasets (more information below)

(iv) Explaining how these methods could be useful by elaborating on their advantages and disadvantages

(v) Writing about your findings by discussing how important you think the methods you have chosen will be in the future

Guidelines – Before You Start

For all components of the final project, we will be using ***LaTeX***. You will be working with the following ***IEEE*** template: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk

Topic Choice and Data Sources

Please choose a theoretical topic of your interest (example: binary prediction) (more information below). I would suggest going through the class syllabus, the class textbook and other books in data science to make an informed decision. You are also welcome to review the syllabi for different classes.

There are several good sources to find good data for your projects. Some suggestions are provided below. You can ***Google*** them and very easily find them on Internet:

1. ***Kaggle.com***: Google company, has a lot of different types of datasets
2. ***Socrata Open Data***: Has cleaned data from government, business, and education
3. ***UCI Machine Learning Repository***: 440 datasets available for machine learning community
4. ***Data.gov***: Has data from many different kinds of government agencies in the US
5. ***Quandl***: A repository of economic and financial data

Deliverables and Due Dates

1) 1-page final project prospectus (one **.pdf** file) – **Due: Friday 11:59 PM, November 12**
2) Final project presentation (one **.pptx** or **.pdf** file) – **Due: Monday 11:59 PM, November 29**
3) Final project essay (min. 3 pages long) (one **.pdf** file), the datasets you have used (multiple files in **.csv** or **.xlsx** format), and the code file you have used to complete the project (one file in **.ipynb** format) – **Due: Tuesday 11:59 PM, December 7**

Please send your submissions on time. Late submissions will be subject to a 5% penalty for every day past the deadline. No submissions will be accepted two days after the deadline.

Components of The Final Project

The final project will be composed of several phases. Each phase will have a different deadline and a set of tasks to complete. Please read this section carefully.

Phase I: Planning and Preparation (5% of your total grade)

In the first phase of the final project, you are expected to write a 1-page prospectus (***IEEE*** format) that should summarize your project **goals** and **expectations** (ambitious and not so ambitious) and cover a summary of what you have done so far in the project. Also, you need to create a **table** and a **timeline** about what you are willing to accomplish and who will do it. The prospectus will help you start the project on time and also give your professor an opportunity to provide feedback. Before starting to work on the prospectus please read more about the project guidelines below and/or talk to your professor.

Deliverables: 1 page ***.pdf*** file in ***IEEE*** format and codes that you used to do the analysis in **`Python`** in ***.ipynb*** format. Files will be sent through *BlackBoard*.

Prospectus Due Date: Friday 11:59 PM, November 12

Phase 2: Presentation (10% of your total grade)

Using Google Docs, Power Point or LaTeX, prepare a presentation. Presentations should be roughly 20 minutes long (+5 minutes for Q&A). Each team member needs to speak around 3-7 minutes.

Deliverable: 1 presentation file in ***.pdf*** format that will be sent through *BlackBoard*.

Presentation Due Date: Monday 11:59 PM, November 29

Phase 3: Final Essay (15% of your total grade)

The final report needs to be a min. 4 page long report written in the ***IEEE*** format and needs to include ***Introduction***, ***Literature Review***, ***Methods***, ***Theoretical Topic***, ***Data Application***, ***Results, Division of Labor*** sections (for more information please read the section below).

Deliverables: Min. 4 page-long ***.pdf*** file in ***IEEE*** format and codes that you used to do the analysis in **`Python`** in ***.ipynb*** format. Files will be sent through *BlackBoard*.

Final Essay Due Date: Tuesday 11:59 PM, December 7

Project and Essay Guidelines

1) **Motivation:** Your final project should contain a motivational statement about why you chose your topic. Some potential questions you may want to answer are: *Why do you think this topic is interesting? What is the 'role' of this topic within the world of data mining? What do you think you will achieve by working on this project? …*

2) **Introduction**: This section should include your motivation and a brief summary of your project including your topic and your findings in a few sentences.

3) **Literature review**: A literature review on the subject (a rule of thumb is reading ten (10) different scientific articles/book chapters (or more) on your subject of choice – the articles/book chapters that you will read should be of *theoretical* and *substantial* nature; they should talk about the methods and empirical applications). This section should tell us what literature you are referencing from (examples: unsupervised learning, risk functions, outlier detection etc.) and should position the methodological tools you will be looking at within the literature (example: the clustering methods we are looking at in this paper have been invented in year … and contribute to the field in … and … ways. They have been developed to solve problems about … and …).

4) **Theoretical topic:** This section should introduce the methods you will focus on. You should introduce at least **four (4)** methods that we have not covered in class. You should introduce the methods (a brief history for each would be fine), talk about the mathematical philosophy behind them (explain in mathematical/computational terms how your methods work), provide a set of formulas (when applicable), pseudocodes (when applicable), talk about computational complexity of the methods (when applicable), and discuss in which scientific field(s) these methods are used. The methods that you choose should be theoretically deep and worthy of discussion (for instance, talking about linear regression is not a good idea). You are also welcome to share short excerpts from the code you used in your paper.

5) **Data application:** In this section, you should use at least 2 real-world datasets (many different examples can be found on data depositories such as *Kaggle*), and at least 2 artificial datasets (examples: a dataset with 10-normally distributed variables, and another dataset with 1000 normally distributed variables. You must introduce these datasets in some detail before you show your application. I would suggest using as many datasets as you can (**min. 4**), since the behavior of the methods you will be introducing will likely change according to the type of the dataset you are working with (distribution of the variables, number of observations, number of variables etc.). Based on this data application, you should write about what the advantages and disadvantages of using these methods in different data science applications could be (examples: computational complexity, challenges with regard to goodness of fit/optimization, usefulness in different applications, and others…)

6) **Results:** In this section, you will summarize your findings. You should highlight the most important set of your findings in the analysis, and talk about how you think the methods will be useful (or not useful) in the future.

7) **Division of Labor:** Lastly, you need to include a section on who does/did what (ideally, create a table and a timeline about what you are willing to accomplish and who will do it).

Structural Guidelines – Numbers to Keep in Mind

Your paper should have and cover ***at least***:

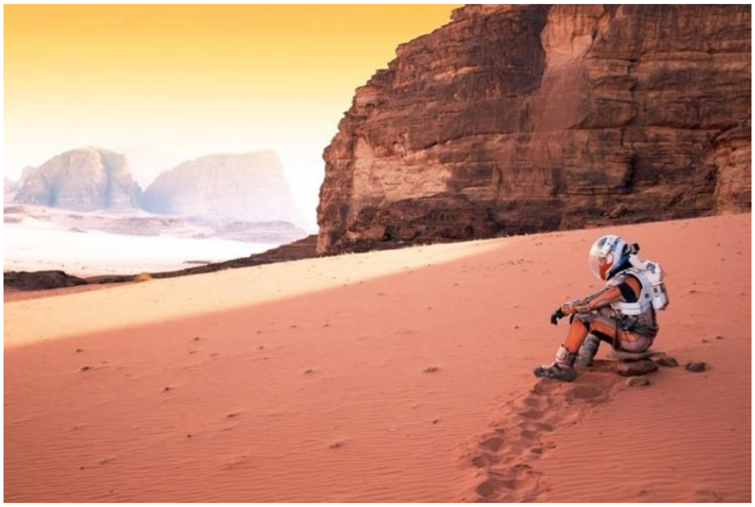
- **Four (4)** different technically complex data mining methods in detail
- **Ten (10)** different scholarly articles / book chapters
- **Four (4)** different datasets (at least 2 *real-world* datasets and 2 *artificial* datasets)
- **Ten (10)** high-quality visuals and/or tables
- **Four (4)** page-long lab report in IEEE format
- A good title for your essay and clearly-labeled ***Introduction***, ***Literature Review***, ***Methods***, ***Theoretical Topic***, ***Data Application***, ***Results, Division of Labor*** sections (actual sub-titles can be slightly different)
- Please use the ***IEEE*** journal template on ***overleaf.com***. Here is the link: https://www.overleaf.com/latex/templates/preparation-of-papers-for-ieee-sponsored-conferences-and-symposia/zfnqfzzzxghk

Choosing a Topic - Example

As stated above, you should choose a set of methods and datasets that have not been (and will not be) covered in the class syllabus. An example:

- **Title**: ***An Exploration of Density-Based Clustering Algorithms: Theoretical Insights About and Applications on Social Media Datasets***
- **Four methods:** OPTICS, FOPTICS, DiSH, HDBSCAN
- **Four datasets:** Tweets collected during BLM protests, YouTube comments obtained from live broadcasts sent by the White House, a collection of artificially created normally-distributed variables, a collection of artificially created uniformly-distributed variables
- **Data application:** The application of the four data mining methods to the datasets and evaluation of the results from theoretical and goodness-of-fit perspectives

After choosing your topics and datasets, you are welcome to schedule a short appointment with me to talk about your ideas. Make sure that you challenge yourself with a technically complex and interesting project! This will be an important part of your grade. ☺ Contact me or your TAs if you have any questions!



The long journey to the end of the semester is close to an end…