DS 420, Projects

There will be two projects in this course. In general, you will work on **the same dataset** in both projects. You may want to use other datasets for project #2, but be careful as you may not have enough time to work on both datasets. My suggestion is to find a good dataset for both.

Some useful courses

https://www.kaggle.com/learn/pandas

https://www.kaggle.com/learn/data-cleaning

https://www.kaggle.com/learn/feature-engineering

https://www.kaggle.com/learn/python

https://www.kaggle.com/learn/machine-learning-explainability

https://www.kaggle.com/learn/geospatial-analysis

Project Proposal

For this part, submit a pdf file with the requested information below filled in.

- Team members: You should find your teammate in the class. If you cannot find one, I will create a group for you
- Data set to analyze
    - Description/background, how big (#rows, #cols)?, what problem you will work with this dataset
    - Link:
    - Target – Can be qualitative or quantitative. If qualitative, list the categories
    - Predictor variables (features). Can be a mixture of quantitative and categorical.

Project #1

The purpose of Project 1 is to implement a machine learning project on a suitable data set that is of interest to your team. The preference is for you to work in a group of 2-3 members.

You need to find a dataset yourself. The data set should have at least 15,000 samples. There is no absolute maximum size. The label (response) variable can be either qualitative or quantitative. You will implement and compare 02 ML methods: choosing from linear/logistic regression, decision, support vector machine, …. You should use mini-batch GD to get good model performances and accelerate the training time.

You should include as many of the end-to-end steps (Chapter 2) as possible.

Requirements: **1 submission per group.**

1. (20 points) Exploratory Data Analysis (01 code file, name it “dataset_EDA”, for example “housing_EDA”). Your team needs to work together in this step. You do not need to exhaustedly explore/transform the data set in this step. Instead, you need to make data in numbers and scale them. You can come back to this step later when you run ML algorithms on the prepared data and you think you need to improve it. The point is that you should quicky make data ready for algorithms to run on their original version first. Then you may need to make transformations later on. Use assignment01 for your reference.
2. (20 points) Data preparation (01 code file, name it “dataset_preparation”). Make a pipeline to handle as many steps as possible. If this step has more than 10 code lines, make a .py code file for that, otherwise, you can include data preparation in the code file of the next step. If you make a code file for this step, you should know that you have to include it in other code files (import)
3. (20 points) Select algorithms (01 code file, name it “dataset_alg_investigation”). You need to investigate at least 05 algorithms for the data set you choose.
4. (20 points) Train algorithms. Each member trains one algorithm. A group of 2 members will have 02 code files, name them “dataset_alg1”, “dataset_alg2”, for example “housing_SGDRegressor”, “housing_DecisionTreeRegressor”): pick top 02 algorithms from the previous step and train them. Make one code file for each. Here, training algorithms means finding the best hyper parameters for them. Use grid search for hyper tuning. In this step, you may need to come back to #1 (EDA) to refine your data for training. If your group has 3 members, do the same but for 3 algorithms.
5. (20 points) Make comparisons for the algorithms you trained in the previous step (one notebook file: you may have markdown cells for explanation or some code to plot something). What metrics do you prioritize to evaluate your models? why?

Project #2

The purpose of Project #2 is to use ensemble learning and train deep neural networks (DNN) on the dataset used in Project #1. The problem is still as same as it is in project #1, but we will use ensemble learning and DNNs to solve it. In this stage, the expectation is that you use the advanced methods to improve the models you got from project #1 and you can interpret the models (a bit). It is not about knowing how to make algorithms work, but how to train good models.

Again, two approaches to get better results. [1] Data-centric: You still need to analyze the data set if needed (e.g., see how people do EDA for the data set on Kaggle discussion, notebooks, or maybe you have some ideas after trying some models in project #1, the point is you should deeply understand the data before moving on), and [2] algorithm-centric: here you use ensemble learning and DNNs.

If you cannot improve your models with advanced methods, you need to find the reason for that.

Requirements:

#1 (20 points) all members: State clearly the desired performance of the model for the data set. For example, state-of-the art models’ performances, human-level performance, the performances in project #1, … You need to refer to a source for that (e.g., a link). Your models’ performances need to be close to the desired performances. Put the desired performance in the comparison file (see #4)

#2 (40 points) Ensemble learning: try to maximize the diversity of the combination of the ensemble (data and algorithms). **Each member works on different algorithms (at least 2 algorithms per member)**. **Make one code file for each algorithm**. Put your full name on the notebook you work on.

#3 (20 points) DNNs: this step will focus on modeling. You need to review the slides we learn about error analysis for concepts and then apply them in this step. This way, you can interpret the models. **Each member works on different DNNs (at least one DNN per member)**. **One code file for each DNN**. Put your full name on the notebook you work on.

#4 (20 points) Conclusions: all members

Make a notebook of markdown cells.

Include the desired performance

You need to compare all algorithms (in project 1 and 2) learning on the dataset. Make a table having the following columns

+ Algorithm: sort algorithms in the order from the best to the worst.

+ Metrics (one or more columns): sort one of them from the best to the worst

+ Author: who does this algorithm.

Do your algorithms achieve the desired performance? If not, what reasons? What do you learn from doing the projects.

Here is the timetable

Time	Tasks you need to do	Documents you need to submit	Where to submit	Resources	Notes
Week 3, 4	Project proposal	Project Proposal	Katie, turn in pdf files	See links in chapter 2 of the textbook for datasets. Or any other data sources you can get a dataset from	You should submit your proposals as soon as you can. Once you submit your proposals, I will let you know if they are accepted
End of week 4					You get my approvals on your datasets via Katie
Week 5, 6, 7	Project #1	Code files	Katie, turn in one .zip file	Lectures, assignments, …	Clarify contributions of each member at the beginning of the files Make comments in code for each of steps so I can identify all the steps easily. For example: #prepare data, #train model #1….
Week 8 -> 13	Project #2	Code files	Katie, turn in one .zip file	Lectures, assignments, …	Clarify contributions of each member at the beginning of the files Make comments in code.