DSA5105 Project Instructions

Due: 22st November 2024 (End of Reading Week)

1 Preface

The goal of the course project is to familiarize yourself with

1. Basic python data manipulation and visualization tools, such as numpy, pandas, matplotlib and seaborn
2. Popular machine learning libraries such as scikit-learn and keras
3. Scientific methodologies in data analysis and machine learning

2 Instructions

2.1 Project Submission Format

In this project you will perform a comprehensive data analysis on a dataset of your choice. Your project should be the form of a **juptyer notebook**, taking note of the following:

1. Your submitted notebook should be runnable as is. If some parts require long training, then please attach a saved model that can be loaded into the notebook and comment out the training steps (but still include them as comments!) in your submission.
2. Include clear comments (use the markdown capabilities of jupyter!) to explain what you are doing at each step, and also your findings and how they may be interesting. Treat it as an "essay" on your dataset!

2.2 Mandatory Components

The following components should be included as graded components:

1. An introduction of your dataset, with visualizations
2. A clear outline of two problems of analysis you post on your dataset:
   (a) a supervised learning task (e.g. regression, classification, segmentation, etc)
   (b) an unsupervised learning task (e.g. dimensional reduction, generative models, etc)
3. For the supervised learning task, employ at least 3 different methods (linear models, decision trees, ensembles of decision trees, kernel methods, neural networks, etc) and compare their performance
4. For the unsupervised learning task, employ at least 2 different methods (autoencoders, PCA, kernel PCA, etc) and compare their performance
5. Cross validation should be used for model selection

3 Grading Criteria

The grade breakdown is based on the following

1. Completion of the aforementioned tasks
2. Scientific correctness of your methodologies
3. Correct use of libraries

4. Clear documentation of code and findings
5. Creativity and style

The total number of points is 15 and is distributed as follows:

- 2pt for introduction of the dataset + visualization
- 4pts for supervised learning tasks (1pt for each method)
- 3pts for unsupervised learning tasks (1pt for each method)
- 2pt for the correct use of cross validation
- 4pts are reserved for clear documentation of the code, the methodology, the findings, and the presentation style.

4 Datasets

You are encouraged to find a dataset with an application area that interests you. You may browse the following resources:

1. UCI Machine Learning Repository [`https://archive.ics.uci.edu/ml/index.php`]
2. Kaggle [`https://www.kaggle.com/datasets`]
3. GovTech [`https://data.gov.sg/dataset?organization=govtech`]

In case you have difficulty finding a dataset, I have included some possible choices below:

1. Heart Disease [`https://archive.ics.uci.edu/ml/datasets/Heart+Disease`]
2. Retail Data Analytics [`https://www.kaggle.com/manjeetsingh/retaildataset`]
3. Atomization Energies and DFT [`http://www.quantum-machine.org/datasets/`]
4. New York Stock Exchange [`https://www.kaggle.com/dgawlik/nyse`]
5. Traffic Signs [`https://www.kaggle.com/valentynsichkar/traffic-signs-preprocessed`]
6. COVID-19 [`https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`]