COMPSCI 589 Fall 2023 Course Project

Contact TA	Yunda Liu
Released	Tuesday, Nov. 21
Submission deadline	Monday, Dec. 11
Total Marks	100

Task Descriptions

The objective of this project is to apply machine learning algorithms to gain insights from **the provided kaggle dataset.** The project expects you to practice what you have learnt in the class to classification/regression tasks and gain in-depth understanding of general machine learning knowledge. Students are required to **compare at least three different learning approaches** on the dataset and report the performance metrics.

Your machine learning models should cover all three categories of feature learning approaches we covered in the class :

1. Fix-shape universal approximators (kernel methods) from Chapter 12,
2. Neural network based universal approximators from Chapter 13,
3. Tree-based approaches from Chapter 14.

You are free to use whatever libraries you are comfortable with to complete the project. For example, Scikit-learn provides APIs for common kernel methods, NN, and tree-based approaches. It should be sufficient for you to implement the three distinct types of machine learning models in the project and thus is strongly recommended. For advanced students who'd like to learn more state-of-the-art libraries, Tensorflow and Pytorch are also popular Machine Learning libraries to use if you are already familiar with their APIs.

Regardless of the libraries you use, your code should be executable on a CPU machine, without the requirement of a GPU.

You are also free to use any code snippets (for pre-processing, etc) from Internet/Github/Kaggle, etc. You are supposed to understand the code snippets you plan to reuse and give credit to the code source (as footnote or citations in your report).

Dataset

Titanic: https://www.kaggle.com/competitions/titanic/overview
This dataset contains information about passengers, including class, gender, age, etc., and the objective is to predict whether they survived or not. You should follow the dataset page on Kaggle to download the dataset.

Your submitted code doesn't need to include the original dataset.
However, your submitted code should explicitly point out under which relative directory the graders need to put the dataset in order to execute your code.

Requirements

1) **Data Preprocessing**
   - Load data from csv files and organize data into proper format suitable for machine learning models.
2) **Model Implementation**
   - Implement at least three distinct machine learning models to achieve the goal.
   - The models should be appropriately trained, validated, and tested on the preprocessed dataset.
3) **Performance Metrics**
   - Use appropriate metrics for model performance evaluation and comparison.
4) **Code Submission**
   - Submit well-commented code on Gradescope that can be used to reproduce the results reported in the project report.
   - Ensure code readability and documentation, focusing on explaining key components and decisions.
5) **Project Report**
   - Provided a detailed project report with the following sections:
     - Introduction
       - Briefly describe the **problem**, **dataset**, and **objectives** of the project.
     - Methodology
       - Explain the data preprocessing steps (such as encoding, normalization and how you handle the columns with missing values, etc.) and rationale behind the chosen methods. -> You should clearly describe how you preprocess the data from the three aspects that we provided, including what approaches you use and why you use them.
       - Detail the implementation of each learning approach, including but not limited to model initialization, hyperparameters tuning, feature engineering, and cross validation. -> You should briefly talk about what approaches you use and which category (i.e. fix-shape universal approximator, neural network, tree-based methods) they belong to. For each

method, talk about what parameters you tune and why these are important to your model. If possible, report the best parameters you find.

- Results
    - Justify the design decisions made during the project, such as the choice of models and hyperparameters.
    - Present the optimal parameters for each model.
    - Describe metrics used to evaluate the model performance. -> For training and validation set, what metrics do you use to evaluate the model performance? For the test set, you need to submit your prediction to Kaggle in order to get the accuracy.
    - Use visualizations (e.g. graphs or tables) for better clarity and comparison. -> Use at least one visualization method (i.e. graphs or tables) to compare your model performances and results. This will also be used as evidence to support your conclusions.
- Conclusion
    - Summarize key findings, insights, and potential areas for improvement.

Deliverables

1) **Project Report**

- A clear presentation of methodology and comparison results.
- Documented findings and insights from the project.
- Properly formatted and well-organized.

2) Code Submission

- Code should be executable to reproduce reported results.
- Grading based on code comments and reproducibility.
- Comment on the hyperparameters tuning part in your code so that we can execute your code using the optimal parameters you have identified.
- Your code should require less than 15 min to set up and run. If the graders cannot set up your code/execute your code in 15 min, it would be considered a failure and cause point deduction.