Machine learning for Real-world Datasets

1. Overall Description

In this project, you will team up with two other students to complete the tasks described in the following.

In this assignment, your team is required to choose a real-world dataset and build a model based on the knowledge we learned in the machine learning part. In this process, you need to preprocess the raw data on your own and try different models for better practical applications. After these, your team needs to submit the codes, write a report, and prepare a 10-minute presentation to show the problem you want to solve and how you solve this problem properly. The final score of this project will be given based on the report and the presentation.

2. Details of Experiments

a) In the first stage, your team needs to choose a problem that you want to solve. Currently, there are countless open-source datasets on the Internet. For example, in the UCI Machine Learning Repository, you can find 1) the census income dataset that is used to predict whether a person's income exceeds 50k per year, and 2) the thyroid dataset that is used to predict whether a person has a diseased thyroid.

b) In the second stage, your team needs to preprocess the data. There are many issues in this part. For example, the null values, useless features, standardization, category features, and so on. The python package Pandas is a powerful open-source data analysis and manipulation tool, and it is highly recommended for your project.

c) In the third stage, your team needs to choose a model and apply techniques to handle issues, for example, avoiding overfitting, tunning parameters, and so on. You should figure out why this model is chosen and why the other models are not as good as the chosen model for this problem.

3. Codes and Report

In the report, you need to first introduce the problem you want to solve and the dataset you used for this problem. Then, present the methodology about how you're going to tackle this problem. A summary of your main results with different methods needs to be shown. After that, a thorough analysis and discussions of your results are required: what did you find in your result? Which method is better for this problem? Which technique is used to solve which problem, and whether it works? Which method does not work and why?

You also need to submit codes containing all experiments and results mentioned in the report.

Grades of this part will be mainly based on a clear description of your problem, a clear justification of approaches used, and a critical interpretation of your results. If you cannot get satisfactory results for your problem, do not worry about that. Try to discuss why these methods do not work for your problem.

4. Presentation

Every team has 10 minutes for presentation and 5 minutes for Q&A. Present your work with confidence!

5. Key Dates:

a) Submission DDL for codes and report: Aug.2nd , 23:55, 2022

b) Presentation: Aug.3rd, 16:20~ 18:20, 2022