COMP4434 Big Data Analytics

Project

PolyU, Hong Kong

Objectives

You have learned logistic regression classifiers, support vector classifiers, and multi-layer perceptron classifiers. In labs, we learned that their implementations are available in Python package "sklearn". Actually, this package contains many other types of classifiers [1]. In this project, we aim to explore more types of classifiers and apply them to handle a real-world problem. In particular, you need to perform four tasks as follows.

1. Selecting an appropriate real-world dataset from kaggle [2], UCI ML Repository [3], or other sources.
2. Analyzing and preprocessing the data.
3. Applying different classifiers to your dataset and comparing their performance.
4. Summarizing your contributions, observations, and conclusions as a report with at least five pages.

Requirements

1. Use the "Groups" on Blackboard to sign up for a group with your classmates. Each group could at most have **three** members. Each group has an assigned time slot for proposal presentation.
2. The given websites kaggle [2], UCI ML Repository [3], MSR [4], and Tianchi [5] contain many data sets, but not all of which are appropriate for this project. You are also welcome to identify data from other sources, such as papers. In your datasets, the number of instances should be **at least** $5{,}000$, and the dimension of features should be **at least** 250. You should provide the practical meaning of features. After identifying an appropriate dataset, you should post its name in "Discussions" on Blackboard, with original source included, e.g., links and paper references. You would get a low score in terms of novelty if your dataset has already been selected by others, according to the post time.
3. When analyzing the data, you should explain the benefits of having accurate classification predictions in your specific scenario.

[1] https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

[2] https://www.kaggle.com/datasets?search=classification

[3] http://archive.ics.uci.edu/ml/datasets.php

[4] https://msropendata.com

[5] https://tianchi.aliyun.com/dataset

4. The three types of classifiers learned in this course would serve as baseline algorithms. You should include at least $2\times$ **number of group members** other types of classifiers, e.g., if you have 3 members, then you should include at least 6 types of classifiers other than logistic regression, support vector machine, and multi-layer perceptron. You would get a high score in terms of novelty if you employ one advanced classifier or machine learning algorithm (e.g., dimensionality reduction algorithm) that is not included in "sklearn", but in a recently published paper.

5. You should design a scientific way to performance evaluation, such as applying k-fold cross-validation. Remember, you should never use the test set to fine-tune parameters.

6. You will give a four-minute presentation to explain your proposed problem and solution. The presentation time is shown in your group name.

7. Your report should use single space and 11pt in Times Romans. We expect the report to be thorough, yet concise. Broadly, we will be looking for content as follows.

   (a) Good motivation for the project and an explanation of the problem statement.

   (b) A description of the data, e.g., what is in the data, and what preprocessing was done to make it amenable for solving your problem.

   (c) Any hyperparameter and architecture choices that were explored. E.g., parameter settings, or decisions to ignore some features. Describe your reasoning behind the choices.

   (d) Presentation and analysis of results.

   (e) Any insights and discussions relevant to the project.

   (f) References.

Grading

The final report will be judged based on the clarity of the report, the novelty of the problem, and the technical quality and significance of the work. Three good examples have been provided on Blackboard. More examples could be found here [6]. The deadline of final report is April 20, 23:59 PM.

[6] `https://cs230.stanford.edu/past-projects/`