COMP4434 Big Data Analytics

Project

PolyU, Hong Kong

Objectives

We have learned several machine learning algorithms, such as support vector machine (SVM), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), as well as their implementations. In this project, we aim to apply them to handle a real-world problem. In particular, you need to perform six tasks as follows.

1. Find a paper in data mining conferences, including NeurIPS, ICLR, ICML, theWebConf, KDD, IJCAI, AAAI, WSDM, SIGIR, from 2020 to 2025.
2. Read and analyze the paper, including the research problem, proposed framework, datasets, and experimental settings.
3. Reproduce the deep learning framework proposed in the paper, by using the codes and dataset released by the authors (please select a paper that has released its codes and datasets).
4. Apply two traditional machine learning algorithms, such as SVM and random forest, to the same research problem and dataset in the paper. There are many machine learning algorithms available in scikit-learn [1]. You can implement them by a few lines. You should use k-fold cross-validation, unless the training set and test set are fixed, i.e., the split is provided by the paper.
5. Apply one basic neural network, such as multi-layer perceptron, CNNs, and RNNs to the same dataset. Summarize the performance of traditional machine learning algorithms, the basic neural network, and the proposed framework, in a table.
6. Implement more complicated neural network architectures by using PyTorch. Design and implement a new neural network architecture, and apply it to the same dataset. It cannot be simple models (e.g., traditional machine learning algorithms, multilayer perceptron, CNNs, RNNs) or the frameworks proposed by the paper.
7. Summarize 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 must have 4 members.

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

2. The paper must be a formal publication in NeurIPS, ICLR, ICML, theWebConf, KDD, IJCAI, AAAI, WSDM, SIGIR, from 2020 to 2025. It **cannot** be a workshop paper, a short paper, or a demo paper. You only need to select one dataset. Its number of instances m and the dimension of features d must be large enough $m * d \geq 1{,}000{,}000$. After identifying an appropriate paper, you should post its name in "Discussions" on Blackboard, with original source included, e.g., links and paper references. You will get an extremely **low score** if your selected paper has already been selected by others, according to the post time. Your dataset **cannot** be MNIST, ImageNet, or their variations.

3. When analyzing the research problem and data, you should explain the motivation and challenges in your own words.

4. You should apply k-fold cross-validation to scientifically evaluate all models. Remember, you should never use the test set to fine-tune parameters.

5. Each group needs to give a 7-minute presentation to explain the selected paper and designed model. **All members should show up**.

6. 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 explanation of the research problem, motivation, challenges, and experimental settings.

   (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) A description of your proposed deep learning model. It is expected that your model cannot outperform the proposed one in the paper. However, you should explain the reason, and attempts that you have made to improve the performance of your proposed model. 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 the experimental results.

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

   (f) References.

   (g) You should also submit your codes, but not the dataset.

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 can be found here [2]. The deadline for the final report is **16 Dec**, 11:55 PM.

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