Develop/Adapt and Validate ML Techniques for Large-scale Classification

Introduction

Classification is one of the most fundamental and important tasks in Machine Learning. In the conventional classification task, the objective is to assign one category (or label/class) to a given sample from a given (fixed) vocabulary of classes. This can be thought of as a binning task, where we can assume one bin corresponding to each class, and a sample (object) can belong to one of the bins.

A generalization of the conventional single-class or single-label classification problem is "multi-label classification" where a sample needs to be assigned multiple labels from a given vocabulary. This can be thought of as a labeling task, where we can assume that we are sticking one or more labels to a given sample (object).

Large-scale single-/multi-label classification involves dealing with a large number of classes (e.g.; of order 10^5 or more), and is often termed as extreme classification or extreme multi-label learning (XML).

The objective of this project is to work on the XML problem. For benchmarking purposes, we will use the publicly available XML repository.

Challenges

As part of this project, at least one of the following challenges need to be addressed for both training and testing:

- Peak memory requirement (e.g.; reducing it by a factor of >=2)
- Compute resources required (e.g.; reducing the CPU and/or GPU hours by a factor of >=2)

Approach

The above challenges need to be addressed by developing new (or adapting existing) ML algorithms with focus on one or more of the following aspects (this list is non-exhaustive):

- Classification approach
- Feature clustering
- Label clustering
- Feature embedding (dimensionality reduction)
- Label embedding
- Feature sparsity
- Scalable deep-learning
- Improving prediction accuracy for rare labels

Grading (Deliverables are highlighted in blue)

The project will be graded based on the following criteria:

- Each team needs to update the project topic in this sheet by 02 Feb, 11:59 PM, and submit a one-page write-up (PDF) summarizing the broad idea/steps by 05 Feb, 11:59 PM.
- Novelty of the work.

- Thoroughness of experimental analyses and validation of claims using multiple datasets, with at least one dataset containing >10^3 labels. Don't hesitate to work with even bigger datasets if your algorithm/system can afford.
- Verbosity will be penalized.
- Final report (PDF) using the ACM template with the following setting:
\documentclass[sigconf, review]{acmart})
- A standalone package/folder containing complete and executable codes, along with a detailed readme file and scripts for reproducing results. Put a link in the report to download this.
- A poster (PDF) summarizing your work in the provided template.
- A video of not more than 3 minutes explaining the poster. Put a link in the report to download this.
- Each report must cite all the relevant references, and include a section acknowledging all sorts of technical help received from friends/classmates/others.
- In case of two-member teams, the contribution of each member needs to be explained in detail towards the end of the report. There may be a viva after the submission deadline, and individual team members may receive different scores.

General Information

- Each team can have one or two members.
- All the deliverables will be collected through a google-form.
- It is not allowed to change your project topic/direction at a later stage.
- There is no restriction on the programming language/platform, and one can use any publicly available library/code with appropriate citation.
- More than one team can work in the same/similar direction. However, it is the responsibility of the respective teams to ensure that the outcomes of their projects are entirely different. Otherwise, this will be considered as a case of plagiarism for all such teams.
- Teams are encouraged to discuss their approaches, share ideas, and even compete with each other unofficially. However, any form of plagiarism will lead to a zero in the project and additionally a reduction in the grade of all its members up to 'F'.
- There will be no relaxation in any deadline.
- If you face any query related to the project, post it as a public comment in google classroom under this project.

Roadmap to submit your work for a potential publication

There will be an option to submit an intermediate status report of your project by 20 March. Those who submit this and demonstrate promising progress will be optionally offered mentoring from the course-instructor for an extended period (~ 4 weeks after the course is over) and access to a dedicated workstation to conduct experiments on larger datasets, so that they can work towards submission to a suitable conference/journal. This provision will be available to not more than the top-2 teams at that stage, and no team if none found suitable. If a team comprising only UG students is one such team, its members would have an option to earn design project credits for the upcoming summer term based on their extended work.

Note that this provision is optional, and your score in the project will be based on your final submission by the due date and not what you do after that. Also, not submitting the intermediate status report will not affect your score.