CS3339: Machine Learning
2024 Fall Project

November 28, 2024

1 Introduction

In this project, you are required to complete a classification task with **high-dimension sparse data**. The data come from an anonymous text classification dataset, in which each text is classified into one of the predefined 20 categories. We will provide pre-extracted features from the dataset, so there is no need to perform feature extraction from raw data.

There are totally 11314 texts for training, and we have another 7532 texts for testing. Each text is represented as a 10000 dim vector. You may simply load these features and their corresponding labels with `numpy.load` or `pickle.load` function. Note that the text features are quite sparse and are in high dimension. Properly handling such high-dim sparse data might be the key to satisfactory performance. You may split a validation dataset with a preferred ratio by yourself.

Your objective is to train machine learning models with data we provided, and achieve as high test accuracy as you can. Detailed descriptions are listed below.

2 Requirements

In this project, you should:

- try at least **three** different machine learning methods, and at least **two** of them are **NOT** deep learning. We recommand methods learnt in this course.
- use at least **two** model assessment and selection methods to choose the best model and enhance the robustness and generalization ability of your models.
- submit a report, your codes, and the results on test dataset as instructed in section 3.

In completing this task, you may want to perform:

- exploratory data analysis, including:
  - understand the feature space and the distribution of classes,
  - analyze the sparsity of the data and its implications;
- dimensionality reduction, including:
  - apply techniques like PCA or SVD which are specifically suited for sparse data,
  - discuss its impact on the performance and interpretability of models;
- model development, including:
  - employ various machine learning models (*e.g.*, Logistic Regression, Linear SVM),
  - experiment with different regularization techniques to handle overfitting;
- model evaluation: perform cross-validation to assess the generalization ability of the models.

Please be awear that although test accuracy (with its ranking) forms part of your score, it is not the only one. The detailed analysis, extensive experiments, a clear and well-written report, mannual

Preprint. Under review.

implementation of the algorithms based on basic libraries (*e.g.*, `numpy`), reasonable modifications to standard algorithms and other highlights all contributes to a good project with high score. You may not wish to be stuck in the minor improvements of performance and ranking.

3 Submissions

You should submit your report and code via Canvas, and submit your results on test dataset with Kaggle platform. The report and code should be packaged to a zipped file, named with your student ID and name, with structure like:

- ZhangSan-023XXXXXXXXX.zip
  - |_ Report.pdf
  - |_ Code
    - |_ Readme file
    - |_ Your code files here...

3.1 Report

Your report should describe how you complete this project, and it may include:

- the data processing techniques, and the reasons you choose them;
- a brief introduction of used models;
- the used model assessment and model selection methods, and how do you deal with the sparsity and high dimension of the data;
- how do you evaluate of your models;
- the conclusion;
- any other things you want to report in this project.

The report should be submitted as **a PDF file**, and recommend to have **six** or more pages. We suggest to use the LaTeX templates provided by top conferences and journals.

3.2 Code

Your code will be used for checking reproducibility. Therefore, your code should contain a `Readme` file, describing how to reproduce the results you submitted in kaggle and mentioned in the report. It coule be a `txt` file for simple cases and recommended to be a `PDF` file if its complex.

3.3 Test Results

Kaggle Link: https://www.kaggle.com/t/c15006c8e13c4e319da3c92d939ee7bb

You need to submit your test results in this kaggle competition. You should register **before the registration deadline** and submit your results **before the submission deadline**. You are also required to include the screen shoots on the accuracy of methods used in your report.

4 Deadlines

Kaggle Registration Deadline: December 8, 2024

Kaggle & Canvas Submission Deadline: January 5, 2025.