Final Project

TA email: `html_ta@csie.ntu.edu.tw`

RELEASE DATE: 04/20/2023

COMPETITION END DATE: **05/31/2023 NOON**

ENGLISH-WRITTEN REPORT DUE DATE: **06/08/2023 13:00**

Unless granted by the instructor in advance, no late submissions will be allowed. That is, you will not be allowed to submit your report after the deadline and will get zero point for the final project. The gold medals cannot be used for the final project.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

You need to write your report in English with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

Introduction

In this final project, you will have the opportunity to participate in an exciting machine learning competition. It is well-known that music can set the mood at a party. However, if the music played fails to inspire joy and encourage dancing among the crowd, it can actually ruin the party atmosphere. That is where "danceability" comes into play—an index that describes how suitable a track is for dancing. Your objective will be to predict the danceability of various tracks, which will be quantized into 10 scales.

In particular, imagine you are the DJ for the final party of the HTML course. Your friend, *Nijika*, the manager of *STARRY* LiveHouse, requires your help to spice up the party by picking more danceable tracks from some library. You need to study different approaches for danceability prediction and then recommend *Nijkia* some of them in order to accomplish this task. The accuracy of the approaches that you study will be reflected through your competition on some internal scoreboard. Then, you need to submit a comprehensive report that describes not only your recommended approaches to *Nijika*, but also the reasoning behind your recommendations. Oh, *Nijika* only speaks English, and hence you need to write your report *in English*. Well, let's get started!

Data Set

The data sets are processed from the Spotify and Youtube data on Kaggle.[1] To maximize the level of fairness, you are not allowed to download or check the original data at any time.

The problem is formalized as an ordinal ranking (i.e. ordinal regression or ordinal classification problem), where the goal is to predict the danceability "ground-truth" (a discrete value) accurately. Both the predicted value $\tilde{y}$ and the ground-truth value y are assumed to be within $\{0, 1, 2, ..., 9\}$. Similar to any real-world data, some of the examples that you get may contain missing values.

[1] `https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube`

Evaluation

For the evaluation, we calculate the Mean Absolute Error (MAE) to measure the average absolute difference between the predicted and ground-truth values. For the introduction and definition of the MAE, please refer to

`https://en.wikipedia.org/wiki/Mean_absolute_error`

Survey Report

You are asked by the LiveHouse manager to study at least THREE machine learning approaches using the data. Then, you should make a comparison of those approaches according to some different perspectives, such as (but not limited to) efficiency, scalability, and interpretability. Then, you need to recommend THE BEST ONE of those approaches as your final recommendation and provide the "cons and pros" of the choice.

The survey report should be no more than SIX A4 pages with readable font sizes. The most important criterion for evaluating your report is reproducibility. Thus, in addition to the outlines above, you should also describe how you pre-process your data, such as the features you build; introduce the approaches you tried and provide specific references, especially for those approaches that we didn't cover in class; list your experimental settings and the parameters you used (or chose) clearly. Other criteria for evaluating your survey report would include, but are not limited to, clarity, strength of your reasoning, "correctness" in using machine learning techniques, the work loads of team members, and properness of citations.

Our sincere suggestion: *Think of your TAs as your boss who wants to be convinced by your report.*

For grading purposes, a minor but required part in your survey report for a two- or three-people team (see the rules below) is how you balance your work loads.

Competition

The submission site will be based on Kaggle and will be announced later. Please simply form a team on the site and then participate in the competition. Use your submissions wisely—*you do not want to leave the TAs with a bad impression that you just want to "query" or "overfit" the test examples.* After submitting, there will be a score board showing the test error on a random half of the data set. The "hidden" test error on the other half will eventually be used to evaluate your performance. The competition ends at noon on 05/31/2023. The competition site will continue to be open until the due day of the report.

Misc Rules

Report: We ask you to write your report *in English*. Please upload one report per team electronically on Gradescope. You do not need to submit a hard-copy. The report is due at 13:00 on 06/08/2023.

Teams: By default, you are asked to work as a team of size THREE. A one-person or two-people team is allowed only if you are willing to be as good as a three-people team. It is expected that all team members share balanced work loads. Any form of unfairness, such as the intention to cover other members' work, is considered a violation of the honesty policy and will cause some or all members to receive zero or negative score.

Algorithms: You can use any algorithms, regardless of whether they were taught in class.

Packages: You can use any software packages for the purpose of experiments, but please provide proper references in your report for reproducibility.

Source Code: You do not need to upload your source code for the final project. Nevertheless, please keep your source code until 06/30/2023 for the graders' possible inspections.

Grade: The final project is worth 1000 points. That is, it is equivalent to 2 usual homework sets. At least 900 of them would be reserved for the report. The other 100 may depend on some minor criteria such as your competition results, your discussions on the boards, your work loads, etc..

Collaboration: The general collaboration policy applies. In addition to the competitions, we still encourage collaborations and discussions between different teams.

Data Usage: You can use only the data sets provided in class for your experiments, and you should use the data sets properly. Getting other forms of the data sets is strictly prohibited and is considered a serious violation of the honesty policy. Using any tricks to query the labels of the test set is also strictly prohibited.