COSC 74/274: Machine Learning and Statistical Data Analysis

Spring 2022 Class Project

OBJECTIVE

The goal of the course project is to implement machine learning models and concepts covered in this course for a real-world dataset. The project will utilize the Amazon product review dataset and focus on binary classification, multi-class classification, and clustering approaches to analyze and categorize product reviews. All code must be implemented in Python and all models must use the Scikit Learn toolkit - https://scikit- learn.org/stable/index.html. Additionally, you are not allowed to use transformer network architectures for your project results.

Projects will be individual: each student will work on their own project. Students can discuss with each other for clarification, but they should make sure not to share codes. Any collaboration or sharing of ideas should be acknowledged in the project reports with sufficient details.

TASKS

1. **Binary classification**: In this task, you have to develop binary classifiers to classify product reviews as good or bad. The cutoff of 'goodness' will be an input, i.e., you have to develop classifiers with the following cutoffs of product rating: 1,2,3,4. Note: The cutoff is *not an input to the model, but to the experiment*. For example, when cutoff=3, all samples with a rating <= 3 will have label 0, and all samples with a rating > 3 have label 1. You are expected to report the performance of at least three different classifiers for each of the four cutoffs. You need to perform cross-validation for hyperparameter tuning. Your report should describe why certain model parameters help or hurt the model performance. For each classifier, you should report in your report the confusion matrix, ROC, AUC, macro F1 score, and accuracy for the best combination of hyperparameters using 5-fold cross-validation. We will share a baseline macro F1 score for classification and at least one of your classification models must achieve at least the baseline score for full credit.
2. **Multiclass classification**: Turn the above classifier into a multiclass classifier where the target classes are 1,2,3,4,5. In other words, you want to classify the product rating on a five-class scale. For each classifier, you should report the confusion matrix, ROC, AUC, macro F1 score, and accuracy for the best combination of hyperparameters using 5-fold cross-validation. You are expected to report the performance of at least three different classifiers. We will share a baseline macro F1 score for classification, and at least one of your classification models must achieve at least the baseline score for full credit.

3. **Clustering**: In this task, you will cluster the product reviews in the **test** dataset. You will need to create word features from the data and use that for k-means clustering. Clustering will be done by product types, i.e., in this case, the labels will be product categories. You will use the Silhouette score and Rand index to analyze the quality of clustering. We will share a baseline silhouette score for clustering, and your model must achieve at least the baseline score for full credit.

KAGGLE COMPETITIONS

To receive instantaneous feedback on the performance of your models on the test data as you are building them, we have provided five Kaggle competitions. There is one Kaggle competition for each of the five classification tasks (four binary classification tasks with cutoffs [1,2,3,4] and the one multiclass classification task). Links to join these competitions are given below. These competitions will allow you to obtain the macro F1 score of your model and compare it to the macro F1 score of the baseline for that classification task. Your model results in these competitions are not going to be directly correlated to your final score - instead, your model results in these competitions will allow you to compare your model with the baseline macro F1 score you must beat. You are allowed to submit five times a day to each of the five competitions.

The public leaderboard for each of these competitions will be given by your models' performance on a random subset of 40% of the test dataset. The other 60% of the test dataset will be reserved for a private leaderboard. A video explaining how to submit your model results to Kaggle is linked below.

For the multiclass classification Kaggle competition, students whose models score in the top 5% of the leaderboard will be eligible for extra credit. The extra credit will be **at most 5%** of the final project grade.

Kaggle Competition Links

How to submit to these contests on Kaggle (VIDEO)

Binary Classification

- Cutoff 1 Competition
- Cutoff 2 Competition
- Cutoff 3 Competition
- Cutoff 4 Competition

Multiclass Classification

- Multiclass Competition

DATA

Link to dataset: https://tinyurl.com/22yau9r8

You will be given two files: Training.csv and Test.csv.

Training.csv: This file is a CSV file consisting of review-related information with the following fields:

- **overall**: This is the product's rating on a scale of (1-5)
- verified: A boolean variable denoting if the review has been verified by Amazon
- reviewTime: time of review
- reviewerID: The unique ID of the Amazon reviewer (some have left multiple reviews)
- asin: Product ID. One product will have many different reviews
- reviewerName: Encoding of the Amazon reviewer's username
- **reviewText**: The Amazon review
- unixReviewTime: unix time of review
- vote: How many people voted this review as being helpful
- image: If there is an image, link to the image
- style: If there is style information (e.g., size of shirt, color of phone), it is embedded in a dictionary here. Only available for some samples
- **Category**: The Amazon product category of the product.

Test.csv

- This file contains all the same features as Training.csv, but the **overall** variable is withheld. You will submit your predictions of the **overall** for each product using this file and we will compare them with the true labels.

PROJECT GRADING.

The deliverables of this project are 1) code and 2) project report. 80% of the grade will be based on the model performance and clustering quality. The performance of your models over each of the five classification tasks (four binary classification tasks with cutoffs [1,2,3,4] and the one multiclass classification task) will be graded as measured by your 5-fold cross-validation scores over the training data for each model. The remaining 20% grade will be based on the quality of the report. **The deadline for project submission is 6/7/2022.** You must submit the project (a link to your project code and report as a zipped file) on Canvas. Your project code **must** be submitted as a Google Colab or Jupyter notebook, with relevant graphs and results saved as cell output.

Your project will be graded on several factors.
i) Readability and executability of your code. As is customary with programming projects, your code must be well organized and well documented. **You will lose points if your code is not submitted as a Jupyter notebook or if the graders need to run your code to observe graphs and results**.
ii) Innovative aspects of any features you might define.
iii) Performance of your predictive models measured via different metrics such as Precision, Recall, Macro F-Score, and AUC_ROC score.
iv) The due diligence exhibited in your work in terms of hyper-parameter optimization when building your predictive model.

BASELINE SCORES

For your reference, these are sone baseline scores for your reference. You do NOT need to beat these baselines. The binary and multiclass classification tasks should be tested on the Test.csv dataset on Kaggle, as described above, and highlighted in your report. The clustering task should be tested on the Test.csv dataset and highlighted in your report.

Task	Metric	Score
Binary Classification Cutoff 1	Macro F1	0.66
Binary Classification Cutoff 2	Macro F1	0.78
Binary Classification Cutoff 3	Macro F1	0.80
Binary Classification Cutoff 4	Macro F1	0.70
Multiclass Classification	Macro F1	0.47
Clustering	Silhouette	0.59