BLG 454E - Learning from Data
Fall – 2024
-Term Project-

Problem: Prediction of insurance costs.

Deadline: 29 December 2024, 11:59 PM

The term project has 3 deliverables: The Kaggle competition results, the source code and a report explaining the designed framework and presenting your results. You will be provided with **templates to use** for the report.

1. Kaggle Competition (20 points)

Description

We have created a private class competition on Kaggle. Please click the following link for the term project competition:

https://www.kaggle.com/t/f9c20bb8f4934c5191589253805db3a5

Dataset

A dataset containing images for 10 classes was used for the term project. Features were extracted using four different models—CLIP, DINO, ResNet, and ViT—and provided as tabular data. Four separate feature vectors are provided, corresponding to different feature extraction methods:

1. **CLIP (Contrastive Language-Image Pretraining):** CLIP capture semantic and visual information, aligning image features with text features from a pre-trained multimodal model
2. **DINOv2 (Self-Supervised Vision Transformer):** DINO focus on structural and object-level patterns, providing robust representations by using self-supervised learning
3. **ResNet (Residual Networks):** ResNet capture hierarchical information and detailed patterns, drawing on the depth and structure of a deep convolutional neural network
4. **ViT (Vision Transformer):** ViT represents images by processing smaller patches, capturing spatial and contextual relationships within the image

We generate feature vectors $x_i \in \mathbb{R}^{1 \times d_i}$ representing a single sample where $i \in \{CLIP, DINOv2, ResNet, ViT\}$. By stacking the samples vectors vertically across N=40000 images, we construct the data matrices $D_i \in \mathbb{R}^{N \times d_i}$.

Goal

Participants are given multiple data matrices and tasked with applying machine learning tools to classify each sample into one of the ten classes. There is no restriction on the use of the data matrices we provide. You may choose to concatenate all features and apply any pre-processing and feature selection methods or focus on a single matrix.

Submission Process

To see the performance of your model on test data, submit your predictions of test data to Kaggle in the defined format. Kaggle will calculate and rank the submission scores using the public test data throughout the competition. These scores are publicly visible on public leaderboard. After the competition end, a *private* test data is used to calculate final model performance. Private leaderboard is not released to users until the competition has been closed. Public leaderboard is calculated with 50% of the test data. The final results will be based on the other 50%, so the final standings may be different. Therefore, train your model as general as possible to avoid overfitting on train and public part of the test data.

Scoring Metric

In Kaggle, your submission is evaluated by the F1-Score.

Submission File Format

You should submit a csv file with exactly [20000] x [2] entries plus a header row. The file should have exactly 2 columns:

1. ID: [0,...,19999]
2. Predicted class

Submission CSVs must have a header row consisting of ID and Predicted as in the sample submission. Using different column names causes a fail in submission process. ID column must include all ID values between [0, 19999].

PS: Your submission will raise an error in cases: you have extra columns (beyond ID and Predicted), extra rows, ID column doesn't consist of integers between [0,19999], Predicted column includes values other than real-values.

```
ID,Predicted
0,9
1,8
2,7
3,6
...
19998,1
19999,0
```

You can download the sample submission file (sampleSubmission.csv) to have a better idea.

Rules

- Every student has to create a Kaggle account
- Form a team of **3 to 4 students (The "team" tab on the competition)**
- Individual submissions are **not allowed**.
- Team members **must** be students **officially registered** to the LFD class
- Your team name in Kaggle should match the name you provided in the list.
- Submission format is explained and a sampleSubmission file (sampleSubmission.csv) is given in the competition webpage.
- You are allowed to use only **Python** programming languages (with jupyter) for the implementation.
- You are only allowed 10 submissions per **DAY**. Start early so you can submit more submissions.
- Academic dishonesty including cheating, plagiarism, and direct copying is unacceptable. Note that your codes and reports will be checked using plagiarism tools!

2. Report (40 points)

Prepare a report in Latex using provided IEEE Conference Paper template. Your report must **not exceed** 2 pages (**one extra page** can be allowed for the **main Figure** illustrating the learning pipeline)!

The report should consist of the following sections:

1. (**6 points) Introduction:** Mention about what and why you did in this project briefly. Give your final score and rank in the competition with your team and Kaggle name.

2. **(6 points) Datasets:** Explain your methods for data preprocessing in detail.

3. **(20 points) Methods: The how?** Describe each component of your classifier. Include a **main figure** illustrating the key steps of the proposed solution (learning pipeline). Explain how you train and test your model in general. **The why?** Explain **why** you have made selected such components. Give all details about the methods like the algorithms used, parameter tuning, etc.

4. **(6 points) Results and Conclusions: First**, report your **5-fold cross-validation** results. Explain your results. **Second**, give your Kaggle score and ranking.

5. **(2 points) References:** The list of references cited in the report. Don't forget the citation to the related reference in the report.

3. Code with 5-fold CV (30 points)

The version of your code that you will upload should have 5-fold cross-validation implemented. The code should take the train data as input, and perform 5-fold cross-validation for training and validating the designed framework model. The code will have two outputs: (1) the predicted samples saved in a predictions.csv file (you can use the same Kaggle format to same them), and (2) the average F1 score.

Important note 1: the code will take in 40000 samples provided to you and perform 5-fold CV on this set. At this stage, you don't need to use the extra test set that's used in evaluation for Kaggle competition

Tidy up your code as to

- run simply,
- get all necessary inputs as function parameters (train and test data, model parameters),
- produce output, i.e. the submission file (test predictions)
- have explanatory comments

Important note 2: Use the following random anchorization seed when applying 5-fold CV:

- **import** random **as** r
- r.seed**(**1**)**

Important note 3: For computing the F1 score, once you complete your 5-fold CV, you will end up with predicted classes and ground truth (actual) classes. You can compute the F1 score as follows:

- **from** sklearn.metrics **import** f1_score **as** f1s
- f1s**(**actual,predicted,average='macro'**)**#returns f1 score result for two melted matrices

Important note 4: You can use the pandas module to read in your CSV files

(https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)

- **import** pandas
- pandas.read_csv**(**[ARGS]**)**# Read a csv files into DataFrames.

4. Project Overall Evaluation

For the project, you will provide a final report in IEEE conference paper format (that is given to you in Latex format). Total score of your project will be calculated as follows:

- The Kaggle competition (50 points)
  - o 30 points: 5-fold CV
  - o 20 points: Your Kaggle rank
- Report: 40 points
- Code: 10 points
  - o Your code should be clean and readable. Implement your code as powerful as possible befitting for a 4th grade student. Weak coding might cause losing 5 to 10 points.

Bonus Marks

Top five teams will be rewarded with bonus marks, 20pts, 16pts, 12pts, 8pts and 4pts. respectively, according to the average of the public and private leaderboard scores.

Ninova Submission Policy

- Submit your PDF report, and code in a zip file through Ninova on time.
  - **Unnecessary uploadings (files, pictures, etc.) will be penalized!**
  - **Only put things in your zip file that you are asked to.**
- No late submissions will be accepted.

References

To learn more about Kaggle Competitions, https://www.kaggle.com/docs/competitions

---

Res. Asst. Erdi SARITAŞ, saritas21@itu.edu.tr

Res. Asst. Ömür Fatmanur ERZURUMLUOĞLU, erzurumluoglu18@itu.edu.tr