多變量分析—Multivariate analysis
Spring 2024

期末資料競賽, due on 2024/6/19

1. Materials:

   This assignment explores the complex realm of Chinese handwriting analysis through the use of machine learning techniques to identify and interpret handwritten characters. Your objective is to build an accurate model for **multiclass classification** of Chinese handwriting.

2. Dataset introduction:

   The dataset, sourced from AI-FREE-Team, includes a subset of 1000 digits from the training set. In the test set, some images have labels that do not appear in the training set, suggesting these classes were not present during training. These cases should receive a label of -1, indicating that the class of the image was not included in the training set.

3. Data download and submit format:

   The dataset can be downloaded from Kaggle platform, contains all datasets for this homework.

   The "**train.csv**" and "**test.csv**" files provided for this Kaggle competition contain handwritten Chinese character images along with their corresponding labels. The original images are stored in the folders named "**train**" and "**test**" respectively. Each row in the csv files represents an image sample, accompanied by its label denoting the class of the handwritten character.

   In order to submit predictions for the test dataset, participants are required to adhere to the submission format specified by the competition organizers. A sample submission file, typically named "**sample_submission.csv**", is provided to illustrate the expected format and serves as a template for participants to structure their predictions. It contains two columns: "**ID**" and "**label**". Participants are expected to populate the "**label**" column with their predicted class labels for each corresponding image identifier in the "**ID**" column.

4. Submissions:

This homework aims to develop predictive models to classify handwritten images. You are asked to:

a. Conduct a thorough analysis of the dataset and complete a comprehensive exploratory data analysis (EDA).
   Hint: Employing EDA and data preprocessing could enhance your ranking on the leaderboard.
b. Submit your submission to Kaggle platform.
c. Submit your report (pdf) and notebook (.ipynb, .rmd, .qmd, …) to E3 platform.

5. Scoring criteria:

The overall grade for the homework will be composed of contributions from the Kaggle submission (70%) and the E3 report submission (30%). The criteria for evaluating both submissions are detailed below:

Kaggle submission (70 pts):

a. Kaggle platform:
   The public leaderboard is calculated with approximately 50% of the test data. The final score will be based on the other 50% (private leaderboard), so the final scorings may be different.
   - Basic score (50 pts):
     Over baseline: 50 pts
   - Ranking score (20 pts):
     F1 score: $\left(1 - \frac{\text{rank}-1}{\text{num\_participated}}\right) \times 20$ pts

b. Display team name: <studentID>
   - Team name error: -5 pts
   - The scoring metric is macro F1 score.
   - You can submit at most 5 times each day.
   - You can only join one team, i.e., team merging is not permit.
   - You will get 0 pts if you join more than one team.

Report submission (30 pts):

c. Following chapters should be included in your report:
   - Introduction [3 pts]
   - Related work (reference or others' work) [3 pts]
   - Material (introduce your dataset and method) [9 pts]
   - Experimental (recording experimental) [9 pts]
   - Discussion (difficulty in the homework or future work) [6 pts]

d. Report format:
   - "Your Student ID".pdf
     You should compile your report file to pdf.
   - "Your Student ID".ipynb (.rmd, .qmd, …)
     TAs will reproduce your result, you should provide an executable file and also report your coding environment (version of package and interpreter).
   - Name typo error will be considered as not submitting the assignment.
   - Reproducing error: -10 pts

6. Rules:

The following rules should not be violated or you will receive 0 pts on this assignment.

a. The use of any external datasets, including pretrained models, is prohibited; all models must be developed from scratch.
b. Since the competition dataset is a subset of AI-Free-Team Chinese Handwritten Dataset, therefore, the use of AI-Free-Team Chinese Handwritten Dataset will not be allowed.
c. The use of OCR pretrained models is also forbidden.
d. Cheating or plagiarism.

Note:

e. Please reference when using others' work.
f. Your score will be deducted by 10 pts for each day of delay.