CS4622 - Machine Learning
Lab 01 - Feature Engineering

Dr.R.T.Uthayasanker

August 20, 2023

****Updated version****
This is an individual assignment!
Due Date: 22nd August 2023 by 11.59 PM

Data-set Description

1. For this lab, 2 CSV files have been provided.
   - **train.csv :** Training data set with 28,520 rows and columns with 256 features and 4 target labels
   - **valid.csv :** Validation data set with 750 rows and columns with 256 features and 4 target labels
2. Both CSV files are generated using the dataset **AudioMNIST**.
3. The first 256 columns are 256 values of the speaker embedding vector of each audio file in the data set AudioMNIST created using **wav2vec-base**. The last 4 columns are speaker-related labels corresponding to each speaker embedding vector.
   - Label 1 - Speaker ID
   - Label 2 - Speaker age
   - Label 3 - Speaker gender
   - Label 4 - Speaker accent
4. Both the train and validation data sets can be downloaded from the link given below
   - train.csv
   - valid.csv

Assignment Tasks

- Your task is to apply all that you learned about feature selection & engineering for each target label.
  1. Feature selection/removal: Eg. using data cleaning/feature scoring techniques (SHAP values)
  2. Feature engineering
  3. Feature crossing

4. Any other advanced feature engineering techniques
5. Dimensionality Reduction
6. Etc. . .

- Finally, you should give the reduced set of features enough to predict each target label.

Note : There are some **missing values** in the label 2 column and the label 4 column is not equally distributed. Consider these things when you are applying feature engineering techniques.

Evaluation

- We have another CSV file called test.csv with 750 rows. That will be given on **22nd August 2023** at 10:15 am through Moodle.
- You should transform the 256 features given in the test.csv using your developed feature engineering and data preprocessing techniques
- You should be able to upload the CSV files for every 4 labels with your final set of features for the 750 rows for classifying each label in each submission link. [**4 sets of transformed features**]. Note: Submission links are provided under section Lecture 04
- The submitting csv files should be in the following format and named the files with your index number (e.g. **190001X_label_1.csv** and similarly for all the four labels).
- The expected csv file for eacg label should have the following columns in the right order. (Note: sample csv file is provided in the submission links)

  1. Predicted labels before feature engineering
  2. Predicted labels after feature engineering
  3. No. of new features (total features in the final set after feature engineering and transformation)
  4. New feature 1
  5. New feature 2
  6. etc.

- In addition, you should submit a **report** stating the feature engine techniques and other processing techniques you used and your **python notebook** or link for the notebook, comprising the code for data preprocessing and feature engineering. (Submission link is provided, 'Lab 1 report submission' under Lecture 04 and the file should be named as **190001X_report.pdf**)

References

1. Feature selection techniques in machine learning
2. Machine Learning Explainability - A kaggle short course