IE506 Programming Challenge Assignment

Ashish Kumar Uchadiya (**23m1521**) IEOR M. tech Date: 10 May 2024

Kaggle details:

- Name: Aashish
- username: aashish31476
- Roll No.: 23m1521

Environment Required

Scikit learn, joblib, tqdm, scipy.

Data Preprocessing

1. Open *_1.2_data_final_sparse.py* file and change the path variables that I have created accordingly.

```
69  if __name__ == '__main__':
70
71      dataset_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset"
72      train_txt_path  = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_train.txt"
73      test_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_test.txt"
74      sample_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/sample.txt"
75      sample2_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/sample2.txt"
76      sample_test_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/sample_test.txt"
77
```

2. After setting the correct paths, run this code, it will preprocess the train and test dataset.
3. It will save all classes (csr_label), features (csr_features, csr_features_sub), unique names of classes (cols_C) and features (cols_F, cols_sub) in .npz and joblib files in the dataset_path initialised above.

Training Model

1. Open *_9.1_CCmultioutput_L2LRtrain.py* file and change the path variables that I

```
56  dataset_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset"
57  train_txt_path  = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_train.txt"
58  test_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_test.txt"
59  load_features_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_feature.npz"
60  load_labels_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_label.npz"
61  load_cols_F_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/cols_F.joblib"
62  load_cols_C_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/cols_C.joblib"
63  load_features_sub_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_feature_sub.npz"
64  model_save_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/models"
```

have created accordingly.
2. After setting the correct paths, run this code, it will train a L2 penalty Logistic Regression model on the full train features (I have already test this method using train-test split) and all train 41 classes, after finding the right hypermeters using GridSearchCV, using ClassifierChain with the order of specified in *Order* list.
3. After train the model will be saved in the *model_dave_path* initialised above as a joblib file.

IE506 Programming Challenge Assignment

Ashish Kumar Uchadiya (**23m1521**) IEOR M. tech Date: 10 May 2024

INFERENCE ON TEST DATA AND MAKING SUBMISSION CSV

Open `_8.2_submission_file.ipynb` file and change the path variables that I have created accordingly.

```
dataset_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset"
train_txt_path  = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_train.txt"
test_txt_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/IE506_2024_progchallenge_test.txt"
load_features_path = "/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_feature.npz"
load_labels_path = "/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_label.npz"
load_cols_F_path = "/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/cols_F.joblib"
load_cols_C_path = "/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/cols_C.joblib"
load_features_sub_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/dataset/csr_feature_sub.npz"
model_save_path = r"/home/23m1521/ashish/Kaggle/_3_IE506_2024_Programming_Challenge/models"
```

In this notebook

1. The model, *features_arr, labels_arr, cols_F, cols_C,* and *X_sub* are loaded.
2. The *Doing Perdition* cell will do inference on train data (200000 samples) and test data (150000 samples).
3. The *Calculating Accuracy* cell will calculate the accuracy on train data.
4. The *Making Submission File* cell will make the submission file using *make_submission_csv* function which takes the *savename* input for naming the csv file, to test this function it also takes *break_* Boolean input which breaks this function at 4^{th} sample.
5. And the later cell is just for uploading this csv file using Kaggle api and subprocess libraries.