

Assignment - 2

Classification of Peptides

Group: 58

Anuneet Anand (2018022)

Pankil Kalra (2018061)

Akanksha Arora (PhD20208)

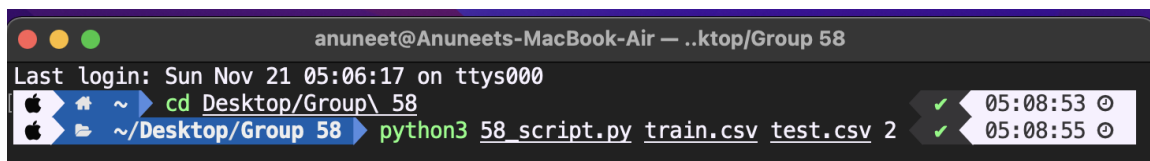
Files

- 58_script.py: Python code file with all the models.
- train.csv: Provided training dataset.
- test.csv: Provided test dataset.
- Submission_x.csv: 2 Files corresponding to 2 models with probabilities.
- Predictions_x.csv: 2 Files corresponding to 2 models with binary values.

Usage

- Requirements: scikit_learn, xgboost (Google Colab version), numpy, pandas
- The given script takes the training dataset, test dataset and model number of predefined models as arguments and accordingly generates the predictions as CSV files in the same directory.
- The train_data and test_data files should be present in the same directory.
- Valid values for the model are 1, 2 (corresponding to our 2 best models) and 0 to run all the models.
- We generate Submission_x.csv files for predicted probabilities (used on Kaggle) and also Prediction_x.csv files for predicted 0/1 values.
- We had to drop the last row of the training dataset as it had some unknown characters that couldn't be parsed.
- Note that Model 2 might generate different output based on XGBoost Version. Hence it should be run on Google Colab to reproduce the same file.

Usage: `python3 58_script.py <train_data> <test_data> <model>`



```
anuneet@Anuneets-MacBook-Air ~ % cd Desktop/Group\ 58
Last login: Sun Nov 21 05:06:17 on ttys000
anuneet@Anuneets-MacBook-Air ~ % python3 58_script.py train.csv test.csv 2
```

The screenshot shows a macOS terminal window with the title bar 'anuneet@Anuneets-MacBook-Air — ..ktop/Group 58'. The prompt is 'anuneet@Anuneets-MacBook-Air ~ %'. The first command executed is 'cd Desktop/Group\ 58', which changes the directory to the project folder. The second command is 'python3 58_script.py train.csv test.csv 2', which runs the script with the training dataset, test dataset, and model number 2. The terminal shows the execution of the script with a green checkmark and a completion time of 05:08:53. The prompt then changes to 'anuneet@Anuneets-MacBook-Air ~ %'.

Feature Generation

- Amino Acid Composition was calculated by our own function.
- Dipeptide Composition was calculated by our own function.
- Mass and Charge was calculated using the pyteomics library.
- Isoelectric Point was calculated using the Bio.SeqUtils library.

Models

- **Model 1:** We used Amino Acid Composition, Dipeptide Composition, Mass, Charge and Isoelectric point of the given peptide sequences as features. The model consists of a BaggingClassifier (100 estimators) with RandomForest as a base estimator (100 trees). The predicted probabilities are stored in Submission_1.csv which has the same name on Kaggle.

Public Score: 0.78620

Private Score: 0.77318

- **Model 2:** We used Amino Acid Composition, Mass, Charge and Isoelectric point of the given peptide sequences as features. The model consists of a BaggingClassifier with a tuned XGBClassifier as a base estimator. There are 23 base estimators and they each have 100 estimators. The predicted probabilities are stored in Submission_2.csv which has the name submission11.csv on Kaggle.

Public Score: 0.77427

Private Score: 0.75699