Data Challenge: Reality AI
A Report

By: Sravya Pamula

Objective:
Build a multi-class human activity classifier based on 3-axis accelerometer readings.

Data:
Data consists of 6 csv files each corresponding to each activity and each sensor location. The activities are ‘standing’, ‘walking’, and ‘jumping’. The sensor locations include ‘chest’ and ‘thigh’.

Methodology:

1. Loading the data: The different datasets provided are loaded using Pandas. Activity and sensor columns are added to these dataframes. These datasets are merged to form a single dataset for modeling.
2. Data- preprocessing: After merging the datasets to get a single dataset, the different features and the target variable is studied. The features include ‘attr_time’, ‘attr_x’, ‘attr_y’, ‘attr_z’ and ‘sensor’. As the data is time series data, to remove the dependency between observations, I perform feature extraction using the windowing method. I create statistical features such as mean, standard deviation etc and also spectral features using fast fourier transform function in Python.
3. Before doing the windowing process, I do the train-test split as we don’t want the test data to mix up with the train data which can happen if we do the windowing with the entire data.
4. Also, it is noted that the ‘jumping class’ is very less in count. To handle this imbalance, I used the SMOTE method.
5. After feature extraction is done, I move on to do modeling. I chose an SVM and a simple logistic regression model. I also perform a GridSearchCV to pick the best hyperparameters for the chosen models. The confusion matrix and the classification report is used to evaluate and compare the models.

Results:

As accuracy scores are not a great option to work with imbalance target distribution, I use precision, recall and f1 scores to evaluate the models. However, since I have also handled class imbalance, I expect that accuracy should still be a good metric to evaluate the models.

Based on the accuracy scores, both the models perform the same. They show an accuracy score of 0.77. However, the logistic regression model performs better than the SVM model in terms of the f1 scores. The f1 score for the ‘standing’ class is 0.71 for the logistic regression model and it is 0.66 for the SVM model. Both the models could classify the ‘standing’ and ‘walking’ classes better than the ‘jumping’ class. The “jumping” class shows perfect precision and recall scores. This could be an artifact of the oversampling that we performed on the dataset to generate synthetic data of this class.

The code is available in the form of a jupyter notebook. The bonus questions are also answered in the same notebook.