Name: Gaurav Misra

ID: 011449815

Rank and F1 score: 17, 0.8108

Approach:

- Read the train and test data.
- Generate sparse matrices out of train and test data.
- Store the classes of train data in a separate list.
- Separate class 0 and class 1 records into separate numpy arrays.
- In order to balance the train data, create 9 buckets out of the train data. Each bucket contains 78 actives and 78 inactives.
- For each bucket of balanced training dataset, use PCA object for dimensionality reduction with number of components for PCA = 300 or 350 (both gave same results).
- For the processed dataset, use SVC with linear kernel as a classifier and classify the test dataset.
- For each of the nine buckets, store the clasification results obtained for the test dataset in a matrix.
- Loop over the generated matrix, and mark a record as active when any one of the classifier provided class 1 to test record.
- Write the final results to the file format.dat

Methodology:

For classification I made use of Support Vector Machine due to the following reasons:

- No missing values in train data.
- Over fitting is handled.
- The linear kernel was used because it tends to perform very well when the number of features is large.

- Linear kernel is much faster than rbf to train, and can give you the same accuracy as rbf kernel.

For dimensionality reduction I used PCA as not all features were good for prediction. PCA allowed me to get rid of the unnecessary features, but at the same time, keep the features which capture the variability in the data, thus not losing the important characteristics in the dataset.

Balancing unbalanced dataset:

I had initially tried to oversample(using SMOTE) the actives, then tried to give more weight to the actives(while using knn), however the bucketing technique explained above worked better for me. Balancing the dataset this way allowed me to go consider all the inactives(after data reduction), without facing the issues related to unbalanced datasets, which lead to better accuracy.

References:

- http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC
- https://stats.stackexchange.com/questions/73032/linear-kernel-and-non-linear-kernel-for-support-vector-machine
- https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
- http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
- Professor’s lecture slides