DA5401 Assignment 7

Let's learn to deal with class-imbalance this time! We will consider the IDA2016 Challenge dataset for our experimentation. The dataset is a binary classification y = {'pos', 'neg'} problem with 170 features and 60,000 data points. The craziness here is that the class ratio is 1:59, that is, for every positive data point, there are 59 negative data points in the training data. The challenge dataset has a training file (aps_failure_training_set.csv) and a testing file (aps_failure_test_set.csv). We will consider only the training file for our experimentation.

Task 1 [20 points]

Split the data file (aps_failure_training_set.csv) into train and test partitions. Build baseline classifiers {SVC, LogReg and DecisionTree} by cross-validating the best hyper-parameters of the respective models. For SVC, the hyperparametes are {kernel, kernel-params}; for LogReg {regularization choice L1/L2, regularization params}; and for DT {depth, leaf size}.

Upon using GridSearchCV, the best parameters are to be found. Note that, GridSearchCV does 5-fold CV by default, which is sufficient for us. Once the parameters are fixed, you will learn the models on the train partition and report the performance metrics on the train and test partitions.

Task 2 [30 points]

Now, we want to address the class imbalance via multiple approaches. You are expected to apply the following in all the three families of classifiers.

a) Consider undersampling the majority class and/or oversampling the minority class.
b) Consider using class_weight which is inversely proportional to the class population.
c) Consider using sample_weights, where you may assign a penalty for misclassifying every data point depending on the class it falls in.
d) Consider any other creative ideas to address the class imbalance.

The goal here is the classification performance metric (macro average F_1) of the hacked classifiers should be better than the baseline classifiers.

Note: Preprocess the dataset to make in amenable for building classifiers.