CMPE 239 HW4 Report

Shiyu Mou
011551358
Rank and Score: NO.1 with 0.80000 F1 score

Homework goal

Drug Activity Prediction on imbalanced daug compound data to determine a given particular compound whether it is active (1) or not (0).

Approach

Data Processing
First I read the dataset and transfered them into sparse matrixes in order to speed up the processing. The traning set is 800*100001 and test set are 350, 100001

Feature selection and Dimensionality reduction

For feature selection and dimension reduction, I tried two algorithms and two different ways to do it.

For algorithms I chose TruncatedSVD and chi2 based on Sklearn library. PCA in sklearn is not suitable since it only deals with dense data. However *TruncatedSVD* and *chi2* can be applied to sparse matrices.

The main goal of TruncatedSVD is to reduce dimensionality and maintain the varience between samples. Here I tried two different ways. The first one is build the TruncatedSVD model based only on positive data (data that are active (1)). I tried this because I thought the positive are rare compares to negative data. If we do select features according to positive data, means we pay more attention on those rare data.
The second way is just build model on the full training set

Data Validation

The Data vaildation is mean to find the best parameters for Feature selection as well as training algorithm.

The vaildation mehod is Stratified K Fold. I separate the traing set into 10 parts and do vaildation on every 1-9 pairs.

Because the dataset is imbalanced, the common way to compute accuracy makes no sense to us. So here I used F1-score to compute the accuracy.

Training Algorithms

I tried 3 different algorithms: KNN, SVM, Neural network. They are all from sklearn.

Parameters finding and model selecting

The first training algoritm I tried was SVM, the first Feature selection method was TruncatedSVD. In TruncatedSVD, when I applied it on full training data, it will maintain 100% varience on 800 features. The validtion f1-score is about 0.84, however about 0.2 on test set(CLP). When I applied it on positive data and reduce to 22 features. I got about 0.97 validtion f1-score and 0.56 on CLP.

Then I tried to use KNN and chi2, to be honest, after several experiments, they didn't make big changes on result. So here I won't address much details.

Then I tried to use sklearn.neural_network.MLPClassifier. Here's what's really confused me. I can only get about 0.66~0.68 on validtion f1-score, however I got 0.8000 easily in CLP. The amount of hidden layers is 2, with 19 nodes for the first hidden layer and 5 for the second. And also, if build model on positive set, it will give a high validtion f1-score(0.9~) but really bad on CLP. So at last I chose to build SVD model on fullset with 800 features and use Neural network.