CS771 Introduction to Machine Learning
Assignment 2

Gunj Mehul Hundiwala
22111024
gunjmehul22@iitk.ac.in

Kartick Verma
22111029
kartickv22@iitk.ac.in

Kush Shah
22111033
kushshah22@iitk.ac.in

Raj Kumar
22111050
rajkumar22@iitk.ac.in

Saqeeb
22111053
saqeeb22@iitk.ac.in

Solution 1

We have used logistic regression classifier for this problem as we conducted our experiments using Decision Tree, KNN, Naive Bayes and found out logistic Regression to be best performing.
We used OVA technique with logistic Regression. OVA Stands for One Vs All.In OVA we train C(no. of classes) models one for each class and predict the most probable class . The Following are the steps for OVA:
1. Here we create 50 data sets one for each class . In each data set we consider one class at a time and set its label to 1 and others to -1
2. Once we have 50 data sets we apply logistic Regression to find weight vector for each class
3. Once we have the weight vector we can create PMF for each data point using sigmoid function and fetch top k values to found the Prec and mprec at k values

Preprocessing:
we have converted the input data X form sparse matrix to dense matrix using toarray() function in python .If we print the frequency of each classes we found out that classes 1,2,3,4 have high very frequency(1428,2558,1509,1475) and the rest classes have less frequency (less than 300) also some classes like 27,40,6,14 are very rare i.e their frequency(3,9,9,12) is extremely low. If we apply logistic Regression on this type of data set the mprec will be less since logistic Regression does not perform well on rare classes due to lack of data set for rare classes. The accuracy on this data set is as follows:

K	prec	mprec
K=1	0.78	0.553
K=3	0.923	0.80
K=5	0.97	0.85

Table 1 : prec mprec values for 10,000 data points

To Tackle this problem we can use can use the concept of sampling i.e increasing the size of data set based on the given dataset. We used randomOversampler() function in the imblearn library to

perform sampling of the data. internally it uses the concept of Generative models as taught in the class .after applying sampling the frequency of each classes become equal .The accuracy of with this kind of data set is as follows:

K	prec	mprec
K=1	0.678	0.663
K=3	0.903	0.835
K=5	0.949	0.894

Table 2 : prec mprec values for 10,000 data points

We can see that our mprec is increased in Table 2.

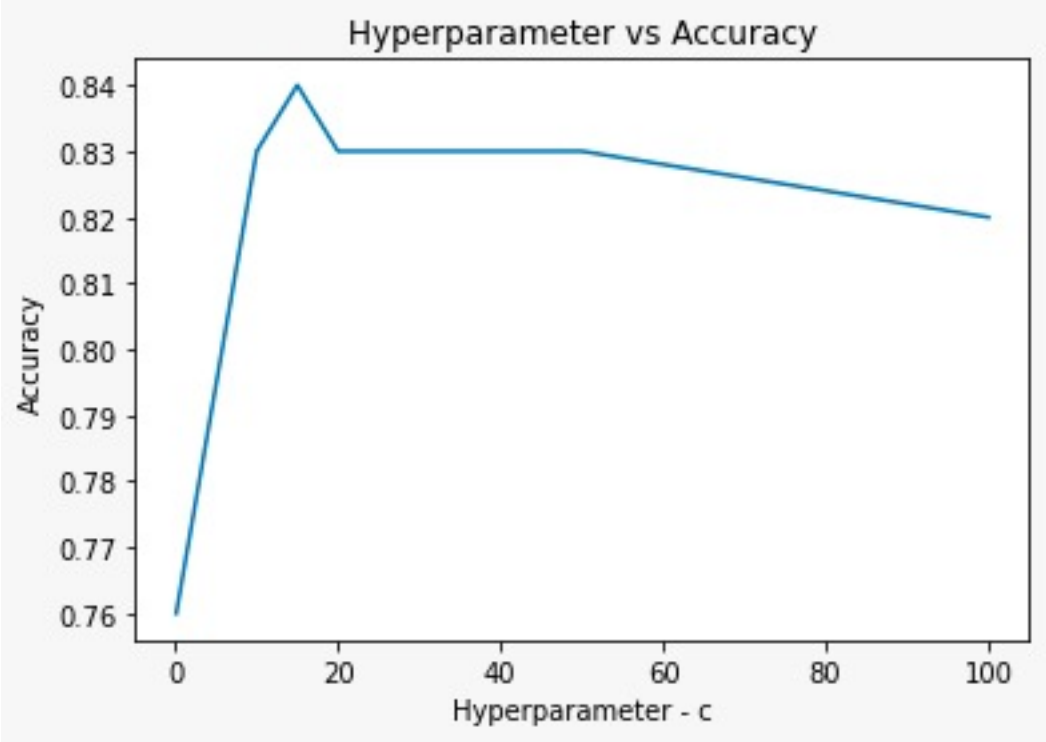
Finally we tried out various tests on sampling of data and found out that if we leave top 4 classes frequency as it is and sample the other classed and make their frequency equal and then the accuracy came out to be .

K	prec	mprec
K=1	0.823	0.847
K=3	0.952	0.971
K=5	0.975	0.984

Table 3 : prec mprec values for 10,000 data points

The reason why this data set gives higher prec and mprec is that in the previous data we were giving equal importance to all classes ,which is not a good approach because some classes have very high frequency than others.

HyperParameters Tuning : There is a regularizer parameter C in Logistic Regression which we have to tune in order to find out best results. To tune the parameter we use grid search in order to find out the best values of C. we took different values of C and plot a graph of the of the accuracy with various combination of other parameters such as penalty and max_iter as well. The best result came out when the parameter penalty of 'l2' is used and max_iter is equal t0 2000 and the value of regularizer parameter C=15.


Hyperparameter vs Accuracy
Accuracy
0.84
0.83
0.82
0.81
0.80
0.79
0.78
0.77
0.76
0
20
40
60
80
100
Hyperparameter - c

Solution 2

Advantages:-

1. In Logistic Regression, the size of model is less than Decision Tree. In Decision Tree size of model is more as it has to store the information of each node in the Decision Tree.

Logistic Regression Model Size	**Decision Tree Model Size**
85KB	850KB

2. Logistic Regression is easy and simple to implement and provides good Accuracy
3. Logistic Regression is not Prone to Over fitting
4. less Train Time than Decision Tree

Disadvantages:-

1. Logistic Regression has slower prediction time than Decision Tree.

Logistic Regression Prediction Time (for 10,000 data points)	**Decision Tree Prediction Time** (for 10,000 data points)
0.08 sec	0.05 sec

2. If the number of observations is lesser than the number of features, Logistic Regression does not perform well
3. Requires a large enough training set to get good accuracy
4. While performing multi-class classification using logistic regression, We come across a problem of less accuracy when some of the classes have less number of training data compared to the other classes.