CS771 : ASSIGNMENT 2

Group Members
Arabinda Karmakar (22111011)
Ayush Kothiyal (22111015)
Manish Kumar Ghildiyal (22111039)
Vamshikiran Morlawar (22111066)
Vinay Agrawal (22111068)

1 Question

Suggest a method that you would use to solve the problem. Describe your method in great detail including any processing you do on the features, what classifier(s) you use to obtain the final predictions, what hyperparameters you had to tune to get the best performance, how you tuned those hyperparameters (among what set did you search for the best hyperparameter and how) e.g. you may say that we tuned the depth of a decision tree and I tried 5 depths {2, 3, 4, 5} and found 3 to be the best using held out validation.

Solution:

We will use Logistic Regression method to solve the problem given in this assignment. Since, this is a multi-class classification problem as there are 50 classes of errors in which a particular error in code can belong to and logistic regression do not support classification with more than two classes, so will use One-vs-Rest approach for solving the problem. In One-vs-All approach we split the multi-class data set in multiple binary classification problems and then train a binary classifier on each one of them. So, model will predict a membership probability for all of the classes. Hence, we will get the probability of all data points (erroneous lines of code) falling in each of the 50 error classes and then we will return the top k classes in which the probability is maximum.

Also, we did processing on features since, none of the test classes fell into the error class 33, 36 and 38 we initialized it with a constant zero which remained zero even after finalizing the model.

Now, first classifier we used was 'Decision Tree Classifier'. The parameters in which were set as follows :-

(splitter = 'best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1 and max_features = None)

On setting parameter values as above the observed accuracy in metric prec@1 was observed to be 72.3%. But the major drawback of using this method was that the size of model was around 1 MB which is quite big.

So, then we used 'Logistic Regression' method to solve the given problem. Firstly, all the parameter values were set to default except the parameter 'Maximum Iterations'. So, initially with value of " max_iter = 1000 " the accuracy over metric prec@1 was found to be around 76.02%.

Now, we tried increasing the number of total iterations and when " max_iter = 7000" we observed increase in accuracy over the same metric i.e. prec@1 which was 80.4%. The reason of this increase in accuracy might be that the solver was not able to converge with 1000 number of iterations.

Plus, the benefit of using logistic regression method over decision tree was that the size of the model was very small as it was around 80 to 100 KB.

Also, since all the other parameters were set as default, all the above observations in accuracy were measured with parameter "C = 1.0". This parameter "C" denotes the "Inverse Regularization Strength" which must be a positive float. So, the smaller value of C denotes stronger regularization.

So, now we decided to increase the value of hyper-parameter "C" and the observations of accuracy in different metrics are mentioned in below tables:-

Parameter : 'C'	(Accuracy %) prec@1	(Accuracy %) prec@3	(Accuracy %) prec@5
1	80.4	93.5	96.8
5	82.4	94.7	97.2
10	82.7	94.8	97.3
20	83.0	94.9	97.4
25	83.1	95.0	97.4

Table 1: Accuracy with different values of hyper-parameter "C : "Inverse Regularization Strength".

Parameter : 'C'	(Accuracy %) mprec@1	(Accuracy %) mprec@3	(Accuracy %) mprec@5
1	50.8	78.56	88.14
5	61.65	85.22	91.63
10	63.53	85.68	92.57
20	65.95	86.57	92.92
25	67.09	87.05	92.77

Table 2: Accuracy with different values of hyper-parameter "C : "Inverse Regularization Strength".

Thus, on observing the above data, we found the value of C = 25 to be the most suitable for our model.

2 Question

Discuss at least two advantages of your method over some other approaches you may have considered. Discuss at least two disadvantages of your method compared to some other method (which either you tried or wanted to try but could not). Advantages and disadvantages may be presented in terms of prediction, macro precision, training time, prediction time, model size, ease of coding and deployment etc.

Solution:

Apart from using Logistic Regression for multi-class classification using built in one vs rest approach, the other approach that we considered for solving the given problem was by using Decision Trees.

Advantages :-

The advantages noted when using Logistic Regression over Decision trees were as follows :-

1. The model size was noted to be considerably small when we used logistic regression instead of decision tress.
2. The prediction gave comparatively better accuracy when we used Logistic Regression and the accuracy dropped a bit when we used Decision trees instead.
3. The time taken to predict the correct class was less when we used Logistic Regression. Method using decision tree was a bit slow while training the model as well the prediction time was more.
4. It was easier to code and deploy Logistic regression model.

Disadvantages :-

The disadvantages of using our method instead of Decision trees were :-

1. Macro precision was less by around 3% in the method used by us as compared to Decision Trees.
2. There is less scope of optimization in our method whereas the accuracy can even be further improved in Decision trees.

3 Question

Train your chosen method on the train data we have provided (using any validation technique you feel is good and also splitting the train data into validation split(s) in any way you wish). Store the model you obtained in the form of any number of binary/text/pickled/compressed files as is convenient and write a prediction method in Python in the file predict.py that can take a new data point and use the model files you have stored to make predictions on that data point. Include all the model files, the predict.py file, as well as any library files that may be required to run the prediction code, in your ZIP submission. You are allowed to have subdirectories within the archive.

Solution:

Code submitted in Zip File.