CS771: Introduction to Machine Learning
Assignment 2

Avadhi Jindal, Kartik Jhanwar, Khyathi Vagolu, Nikhil Mehta, Sarthak Gothalyan

Problem 2.1

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

OvA (One vs All) is our suggested method for solving this classification problem. In OvA method we will pitch each class against all other classes and see if the data point is a part of the class which we have picked or if it exists within some other class. We will repeat the above procedure for all classes and in the end we will get which classifiers believe for a data point to be of their class.

For above method we will have to create multiple binary classifiers with each classifier corresponding to one class. A classifier will classify the data points whether they exist within that classifier's class or not. We will have to create seperate datasets for each class and use them to train binary classifiers for that class.

We normalized the features as some token was more frequent than others using l2 norm. Logistic regression was used with one vs rest classifier. As logistic regression gives probability of an event belonging to a particular class and thus number of binary classifier trained is equal to number of classes. We tried various methods for the optimization problem for loss function and found Newton Conjugate Gradient to give the best accuracy when we used cross validation and size of test data to be 20% of entire data-set. We used 'multinomial' in which the loss minimised is the multinomial loss fit across the entire probability distribution.

The value of hyper-parameter **C** (Inverse of Regularization Strength) was varied from 0.01 to 100 and it was best performing near values of 10 to 20, value chosen was 11 using various cross validations.

We also tried using SVM with a similar approach but Logistic Regression was giving slightly better result and thus was used.
These are some of the scores when using the above parameters and cross validation on test size of 20%:

$$prec@1 : 0.838 \quad prec@3 : 0.959 \quad prec@5 : 0.981 \tag{1}$$

$$mprec@1 : 0.6781 \quad mprec@3 : 0.8926 mprec@5 : 0.9488 \tag{2}$$

Similarly when we used SVM with fine tuning the parameters results were this:

$$prec@1 : 0.767 \quad prec@3 : 0.879 \quad prec@5 : 0.931 \tag{3}$$

$$mprec@1 : 0.4233 \quad mprec@3 : .57511 \quad mprec@5 : 0.7006 \tag{4}$$

Problem 2.2

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

In our approach, we used the One vs All method where we modified Logistic Regression by splitting the mutli-class classification problem into multiple binary classification problems and fitting the standard logistic regression model on each subproblem and predicting using the model that is most confident.
We have also tried to apply SVM instead of logistic regression but found that logistic regression was better for the following reasons:
Firstly, in terms of computational complexity, logistic regression is better than SVM as in the case of SVM, all the training data must be stored while in logistic regression, all the data is condensed into a single D-dimensional vector.
Secondly, although the prec score when using SVM on some features was comparable, the mprec score was worse than the score we got using Logistic Regression.

Also, when compared with a basic decision tree approach, the mprec, prec scores were worse in spite of using up more space than the logistic regression approach.

As an alternative to the One vs All method, we could have used a method wherein for each error class, we could have separate feature vectors and train them to get different scores. Then we could pick the top five errors by picking those that gave the highest scores.

Problem 1.3

Code submitted through the google form.
Link: https://home.iitk.ac.in/ nikhilme/files/CS771$_A SS2.zip$
Password: 2i3AquDQOs