CENG499 HW2 Report

Yekta Demirci

29/04/2018

1 Decision Trees

In decision tree part. I only could able to write an alghorithm which gives gain information, gain ratios and ginis. Alghorithm calculates them well. I tested according to decisiontree pdf given in the lecture and the results are the same. According to information gain and gain ratio, the attribute with the highest value should be picked as root and new gains should be found to find the next root until all leafs become pure. For gini index lowest value is selected as root. It is said the code should not print anything however in order to show Information Gain,Gain Ratio and Gini values are correct I printed the resulted values for the first root. Getting decision tree without any M.L library was beyond my experience and knowledge. However I hope I can get some partial credit from coding the gain,ratio and gini which are supposed to be learnt from CENG499 course.

2 SVM

2.1 Seismic Bumps Dataset Experiments

1. I obviously preprocessed the data. I did by both reasoning and experiences from lectures. I upsampled the minority samples,created cross validation data set and feature scaling. Because some samples are between 0-1 whereas some others are more than 10.000. If feature scaling is skipped some features would be ignored

2. I did class balancing because there are 1609 0s where as 113 1s. I upsampled 1s up to 1609 according to 1s we already have. If I wouldn't make upscaling I would get good accuracy since 1s would have been neglected. Which is bad.

3. I created a validation set. After upsampling I had many train data so I split it also into cross validation set. It is necessary because there 4 different alghorithms and which one performs better is unknown. Therefore it is wise to do cross-validation checks. I only did 1 cross-check validation though. Test set is NOT shuffled.

4. There are 4 kernels available. Linear, poly, rbf, sigmoid. Sigmoid seems like the worst one. Linear is slightly worse than poly and rbf. rbf is the best however poly accuracy is also very close to rbf according to cross validation results. For different iterations rbf or poly may work better. However rbf is more likely to work better. Dataset has a distribution shape similar to gaussian.

5. I did not change the hyperparameters much. Because there are only a few ones important. One of them is the kernel and it is changed obviously.

6. Confusion matrix is obtained to find accuracy for each different experiment. I only changed kernels because they are the most important factor. If others are set something optimal, changing them does not bring a high success. 0.801 accuracy for linear kernel from cross validation. 0.821 accuracy for poly kernel from cross validation. 0.837 accuracy for rbf kernel from cross validation. 0.704 accuracy for sigmoid kernel from cross validation. As a result rbf kernel is used for test prediction and test accuracy is 0.885. Since upsampling is taken into consideration this result is quite well.If I ignored minority data,accuracy would seem better but actually machine would work worse. Also, test result is better than cross validation because in cross-validation, some artifical upsampled data is used,however test data is more real and similar to train data.

2.2 Website Phishing Dataset Experiments

1. I used a very similar alghorithm for this dataset as well. I explained what I have done longly in Seismic Bumps Dataset part so I briefly explain here. I preprocessed the data. This time, there are 3 class option and there are 68 0s,365 1s and 468 -1s. Upsampling is used. Cross-validation set is obtained. HOWEVER in this part there is no need for feature scaling. All features are around 1 already. So feature scaling part is deleted from the code.

2. As I told in 1. 0s are minority so I upsampled them to 468. As a result there is 468 0s and-1s, 3651s which is okay since 365 is not so low than 468.

3. Likewise Seismic part we use different kernels so a cross-validation set would be good. Therefore I create a cross-validation set from upsampled training sets. Test set is NOT shuffled with this cross-validation. Likewise I only used 1 cross-validation.

4. Likewise Seismic case rbf is the best,poly kernel is very likely to it, linear is so so and sigmoid is the worst. Dataset has a distribution shape similar to gaussian.

5. Also in this part I only changed kernels not changed others because of similar reasons like in Seismic part.

6. Similar approach is fallowed like Seismic pump code. 0.756 accuracy for linear kernel from cross-validation. 0.830 accuracy for poly kernel from cross-validation. 0.837 accuracy for rbf kernel from cross-validation. 0.648 accuracy from sigmoid kernel from cross validation. As a result rbf kernel is used for test and accuracy is 0.897.

Regards