Homework 9

Introduction to Machine Learning

October 2018

1 Introduction

In file *countries.csv* we have a dataset with 227 countries. For each of them there is a country name, region and 17 numerical features. The problem is based on numerical features predict, is the country located in Europe or not.

Our task is to train AdaBoost algorithm for this problem, find outliers based on the model parameters and retrain without outliers.

2 Outliers detection using AdaBoost

We will use AdaBoost both for classification and outliers detection. It generates a set of base algorithms and their weights.

Procedure 1 AdaBoost

Input: X^l, Y^l – training set. T - maximal number of base algorithms.
Output: Base algorithms b_t and their weights α_t

1: Initialize all samples weights $w_i = 1/l$, $i = 1, \ldots, l$;
2: **for** t=1,...,T **do**
3: $b_t = arg\min_b N(b, W^l)$;
4: $\alpha_t = \frac{1}{2} \ln \frac{1-N(b_t,W^l)}{N(b_t,W^l)}$;
5: Recalculate weights $w_i = w_i * exp(-\alpha_t y_i b_t(x_i))$; $i = 1, \ldots, l$
6: Normalize weights $w_0 = \sum_{i=1}^{l} w_i$; $w_i = w_i/w_0$,$i = 1, \ldots, l$
7: **end for**

Here $N(b, W^l) = \sum_{i=1}^{l} x_i[b(x_i) = -y_i]$ - a total sum of negative classifications.

Base algorithms weights are used for prediction. But it each iteration the algorithm recalculate sample weights. For hard samples it increases the weight, for others decreases. We can expect, that outliers are the hardest samples on the training set, so after the final iteration they will have the highest weights.

We used AdaBoostClassifier from sklearn library. It calculates sample weights inside during the training phase, but don't provide them using public methods and don't even store them in private fields. Nevertheless as we can see from the lines 1, 5, 6 in the Procedure 1, weights depend only on the dataset, base algorithms and their weights. The algorithm is shown below.

3 Implementation

The first problem for the implementation is that some values are missed. The simplest strategy is just to skip the lines with a missed data. 179 countries left.
If training set is small, we have too few positive samples in the training set, quite often exactly 0,

Procedure 2 Samples weights

Input: X^l, Y^l – training set. Base algorithms and their weights b_t, α_t, $t = 1, \ldots, T$
Output: Samples weights w_i, $i = 1, \ldots, l$

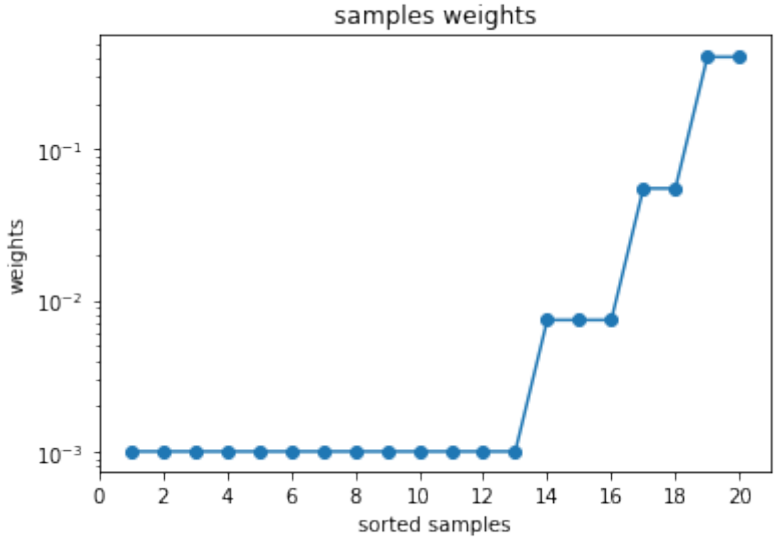
1: Initialize all samples weights $w_i = 1/l$, $i = 1, \ldots, l$;
2: **for** t=1,...,T **do**
3: Recalculate weights $w_i = w_i * exp(-\alpha_t y_i b_t(x_i))$; $i = 1, \ldots, l$
4: Normalize weights $w_0 = \sum_{i=1}^{l} w_i$; $w_i = w_i / w_0$,$i = 1, \ldots, l$
5: **end for**

which is not enough for training. If test set is small, we don't have a statistical significance in the result. It is impossible to get statistical significant result here, we decidet to put 30% sampel to the test data.

Training the Adaboost classifier on all data gave us 0.9259 accuracy.

Without statistical significance we can not make a valid experiment which base algorithm is the best. We decided to take logistic regression as a base classifier in the algorithm and 10 of them in the ensemble.

As a formal criteria for decite which weight is an outlier, we use 3 sigma rule - all values inside 3 sigma interval from the median value considered as an inliers. But we can see, that in this case calculating median and average values is very nonrobust - outliers increase variance too much. So we decided to estimate those parameters all samples except highest 20 weights. This estimation is biased too (now it underestimates values), but not so significant. We got 7 outliers. Below you can see value of the 20 highest weights in logarithmic scale. It is obvious that our criteria identified them right.



Trainng AdaBoost without outliers increases accuracy from 0.9259 to 0.9629.

4 Conclusion

We removed outliers using AdaBoost. It increased the accuracy for a result classifier. Nevertheless changing the seed we can see, that the results are not repeat all the time. The reason is that the test size is not sufficient for statistical significant result. It also prevented us to experiment with AdaBoost parameters such us base classifier and the number of them. A problem of this dataset that it has not enough data to train and test the model. At the beginning we removed missed values. Instead of that we could change the missed value by the average in the column. It could improve the situation, but will not solve the problem completely. In this work we didn't proved that the experiment really work, because the result is not significant, but achieved some learning outcomes such as understand how

Adaboost works and how to apply it to outliers detection.

5 Feedback

Based on submitted solutions we want to give a small feedback for you. Many of you did a really good job. It included lean approach for data (in this report for simplicity we just skipped samples), providing many experiments to find parameters, creative justifications. All this effort deserve a good grade. Nevertheless there is a room to improve in writing the report in average. Read here some thoughts how to writhe a report. It is not a complete guide, but could be helpful. For example, common mistake with a graphics is described there.

I want to focus on the report structure. In the same document you can read what should be in the report. The introduction is missed for a half of submissions. We didn't expect a background part since it this case you used only things that were covered on the labs. It means that you don't have to repeat AdaBoost algorithm, for example. In this report it is included just to show how to apply it to outlier detection. This part may also be discussed on labs, but it is safer to explain it in the report anyway. Later when somebody else will read this work and will be able the idea behind. The body of your report also should be structured. It should be clean which steps you are making. Your decisions should be explained here. It is not always justification. In practice you are not always experimenting on all hyperparamethers of your method. In this case just tell about your decisions. Don't forget to make a resume for your work. Which results are achieved? What could be done different? Any other thoughts about the task. We found that for ones who used ipython for the report is was harder to follow the good structure.

In addition, I would recommend you to make a report in LaTeX. It is not mandatory, it is possible to produce the document with the same quality in Markdown of Office application. But works in ipython format were mostly the code with some comments instead of report included some code. All this tools allow to produce a nice formatting.

You hadn't a precise instructions and made some mistakes. University is a safe environment to make mistakes. And it it the best way to learn. You could have a template and follow it. After repeating it multiple times you may be able to repeat it. But it will not understand why this format is good and what could be improved, because you hadn't thought about it. Moreover, thinking will make you able to make it better than we can.

6 References

www.MachineLearning.ru lectures (rus)
Non-Technical Guide to Writing a Report