Assignment 9

Name: Jingyuan Liu;
AndrewId: jingyual

Description:

The data set you will be working with for this assignment is the CPU dataset that comes with Weka. The purpose of this assignment is to practice doing optimization. You can use the same train/test pairs included with the assignment.

(1) What is your baseline performance for each algorithm using default settings?
Solution:

For the first algorithm, LWL the baseline performance using default settings would get a correlation coefficient score with average 0,85.

For the second algorithm, KStar the baseline performance using the default settings would get a correlation coefficient score with average 0.97

(2) For each algorithm, describe your optimization procedure and results using the prompts below as a guide.

Algorithm 1:
Stage 1: What setting did you determine you would use to build your model and why?

Solution:

I first test with the LWL algorithm. I first use the locally weighted learning with linear regression as the base classifier. I choose the KNN value with -1 as the baseline for the model. Basically, this setting means using no near neighbors. Then I could tune the parameters of KNN value to add more neighbors and see its influence on the classifier performances.

Then I change the KNN value with 2, and 5, increasingly. I first train on the training dataset to see its performance on each datasets. Then I would decide which is the optimal setting. After I got the optimal setting, I would use on the test set to see its performance

Stage 3: What do you estimate will be that model's performance on a new set of data?

Train fold for KNN value -1	Train fold for KNN value 2	Train fold for KNN value 5	Optimal Setting	Test set performance
0.87	0.94	0.97	KNN 5	0.91
0.83	0.76	0.97	KNN 5	0.82
0.83	0.79	0.97	KNN 5	0.95
0.81	0.71	0.97	KNN 5	0.92
0.88	0.66	0.98	KNN 5	0.91

Average = 0.97 (Using total training datasets with best setting 10-fold cross validation)

Algorithm 2:
Stage 1: What setting did you determine you would use to build your model and why?

Solution:

I would use the KStar Algorithm with the globalBlend value varying from 0, 10, to 20. I could use the globalBlend value with 0 as the baseline. Then by increasing the global value, I would see the classifier performance changes.

Then I change the gloabalBlend value with 10, and 20, increasingly. I first train on the training dataset to see its performance on each datasets. Then I would decide which is the optimal setting. After I got the optimal setting, I would use on the test set to see its performance.

Stage 3: What do you estimate will be that model's performance on a new set of data?

globalBlend value = 0	globalBlend value = 10	globalBlend value = 20	Optimal Setting	Test set performance
0.97	0.97	0.97	globalBlend 0	0.87
0.97	0.97	0.97	globalBlend 0	0.95
0.98	0.97	0.96	globalBlend 0	0.98
0.98	0.97	0.97	globalBlend 0	0.98
0.98	0.97	0.97	globalBlend 0	0.96

Average = 0.97 (Using total training datasets with best setting 10-fold cross validation)

(3) Based on your results, was it worth it to do the optimization? Or should you just use default settings? How did you make that determination?

Solution:

To do the optimization or not, that is a question. Basically, the performances changes with different settings would be different given different algorithms.

Some algorithms are sensitive to the parameters, like the KNN value, the neighbor number for LWL models. Changing the parameters and doing the optimization would help us achieve better performances.

On the other hand, some algorithms, like KStar, are not sensitive to the parameter tuning. For these models, we would not need to do the optimization, just use the default setting is okay.

We need to make the determination based on our prior knowledge to the algorithms or we could just do some experiments to justify our thinking.