Big data and high dimensional data analysis, computer exercise 3

Xijia Liu

December 7, 2017

In this computer exercise, we practice on the knowledge what we have learned from lecture 4 to 12. Besides the studying points from computer exercises 2, kernel methods and ensemble methods will be covered in this exercises. There is only one exercises, that is solving the hand written digit recognition problem. There are around 9000 16×16 pixel images of hand written digit. I keep 2000 of them in my computer. You use the rest of them to build up and validate your model. The data is stored in a R workspace and you can download it from Cambro. Once the best model have been selected and trained, then you can submit your model and a report of your study to me. The final model should be stored in a R workspace. Some special requirements of submitting model see below. I try to run your model on the 2000 examples and reply the accuracy back to you. You have one opportunity to fix your model in case I can not run it on my computer.

The last day to hand in the report and your final model: 2017-12-15.

The grading on computer exercise 3 is as follows

- Grade 5: If the accuracy of your classifier on test data is higher than 97%.
- Grade 4: If the accuracy of your classifier on test data is higher than 95%.
- Grade 3: If the accuracy of your classifier on test data is higher than 92%.

To get the credits from this exercise, you also need to provide a complete report. The report should contain the following information:

1. Present all the models you have tried. If the number of models you compared is less than 2, then your report will be viewed as invalid. Different combinations of classifier and features can be viewed as different model.

2. For each model, you need to present all the details, e.g. about features, methods, validation methods, and so on. Each method should be presented in a simply word.

3. If you choose any other methods which are not discussed in our lectures, then you should present the methods in a good way.

Requirements of submitting model: Please write a R function for predicting by your model. The input should be the data matrix of features of new examples. It should have the same format as the training data. The output should be predicted labels of each images. You should keep this function and necessary objects in a R workspace. The unnecessary objects must be deleted.

The group which has the highest accuracy will receive a mysterious Christmas gift from Xijia. If there are several groups share the same highest accuracy, then it will be decided by lottery ☺