Kursus Bigdata F2017

Problem Week 7

Prepare an R script file documenting the following:

Probblem 7.1

In the following is focused on the randomForest() classifier using the classification example in 6_R_Intro_Classification.R close to line No. 250 in Section 6.5: "Random Forest". The problem is solved through the following three steps.

Step. 1. Study the details in the manual for randomForest() using ?randomForest() in the RStudio console window. Study the manual for no less than two of the most important parameters of randomForest() ntree and mtry.

Step. 2. For each of the two parameters, determine the recommended values from the literature. E.g. mtry is typically recommended to be equal to $\sqrt{p}$ where p is the number of variables in the training set used. Now, select for the parameter ntree two additional values such that ntree can assume the values 20, 100 or 500 (default). Furthermore select for the parameter mtry a value which is much lower than the recommended and a value which is much higher than the recommended. E.g. mtry could then assume the values 2, $\sqrt{p}$, *p-2.*

Step 3. Finally run randomForest() training and test experiments for all 9 combinations of values of ntree and mtry and prepare a 3 x 3 table with the results. Comments on the results.

Problem 7.2

In the following is focused on the function svm() from package "1071" near Line 300 in 6_R_Intro_Classification.R . The problem is solved through the following three steps.

Step. 1. Study the details of ?svm() manual for no less than these parameters: kernel="radial" and the parameter gamma.

Step. 2. Determine the gamma value recommended from the literature. Then select one value much less than the recommended and one larger than the recommended. Notice that gamma determine the influence range of a given data sample.

Step. 3. Finally run svm() training and test experiments for the three parameters, and compare the results to the results obtained in Problem 7. 1 using the randomForest().

The resulting R codes from Problem 7.1 and 7.2 are inserted into a Problem7_xxx.R script file where xxx are characters chosen from the persons name. Each participant keeps the script for later submission.

Course material

[James, 2013] Gareth James, Daniela Witten, Trevor Hastie
"An Introduction to Statistical Learning withn applications in R", Springer, 2013.
http://www-bcf.usc.edu/~gareth/ISL/code.html

[Kabacoff, 2015] Robert I. Kabacoff, "R in Action", 2'Ed, Manning Publications, 2015.

2017.03.16/jaas