## ASSESSMENT 01: ENSEMBLE METHODS AND GENETIC ALGORITHMS

|  |  |  |  |
| --- | --- | --- | --- |
| **Author/s** | | | |
| **Group:** |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name:** |  |  | **Surname:** |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name:** |  |  | **Surname:** |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name:** |  |  | **Surname:** |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name:** |  |  | **Surname:** |  |

|  |
| --- |
| **General guidelines** |
| * **This practice must be done using R.** * **Use English (o español, como queráis)** * **Do not change the format of this report.** * **Add your R code at the end of this report.** * **Submit your work via Moodle:**   + **This report (pdf format)**   + **Your scripts (.m)** |

|  |
| --- |
| **Comments** |
| *Add here relevant comments related to the course of the assessment, if any* |

|  |  |
| --- | --- |
| **TASK 1** | UNDERSTANDING BOOTSTRAP Use the script AssML2\_T1.A1.R |
| *Use bootstrap in order to estimate the variance of the complexity (“number of leaves”) and accuracy of a classification tree fitted using rpart. It is obvious that these variances are very sensitive to the sample size, the larger the dataset, the lower the variance should be. There are other sources of variability such as the number of variables used as inputs, the selected complexity of the model and the inherent complexity of the problem at hand.*  *In the script AssML2\_T1.A1.R you can find the basis for completing this task. Note that using the “boot” function it is also possible to get an empirical idea of the distribution of “accuracy” or “number of leaves” by using the histogram.*  *In particular, you should analyze the influence of the size of the training set, the number the variables and the cp parameter in the variability of the number of leaves and the accuracy of the classification trees. Try with two different problems:*   * *SimDataRECTANGLEwithNoise.dat* * *SimDataRECTANGLEwithoutNoise.dat*   *Both problems have 5000 observations with 6 input variables and one output Y, with values I (inside the rectangle) or O (outside the rectangle)* | |
| **EXPLORATORY ANALYSIS**  **TASK 1** | **Include here a basic exploratory analysis of the problem.**  **Justify your answers, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 1.1** | ***Consider the cases with and without noise, using the same small cp (e.g. 0.001) and a subset of 1000 observations with all the candidate inputs.***  **In which case the variability of the trees is larger? Why?**  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 1.2** | ***Consider using a subset of 1000 observations of the case with noise (all candidate inputs). Compute the variability of trees when cp is small (0.001) and large (e.g. 0.3).***  **In which case the variability of the trees is larger? Why?**  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 1.3** | ***Using the dataset SimDataRECTANGLEwithNoise and setting a small cp. Compare the results with 500, 1000 and 5000 observations.***  **In which case the variability of the trees is larger? Is the result you expected?**  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 1.4** | ***Using the dataset SimDataRECTANGLEwithoutNoise, with small cp and a subset of 1000 observations, consider two cases:***   * ***Case A: Build the trees only with candidate variables X1 and X2*** * ***Case B: Build the trees only with candidate variables X4 and X5***   **In which case the variability of the trees is larger? Why?**  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **CONCLUSIONS**  **TASK 1** | **Summarize here your conclusions resulting from the previous experiments. In particular, discuss the impact of the sample size, the *cp* parameter and the number of input variables in the accuracy and complexity of the model.** |

*(Add here your main conclusions)*

|  |  |
| --- | --- |
| **TASK 2** | USING BASIC TREES, BAGGED TREES, RANDOM FORESTS AND GRADIENT BOOSTING Use the script AssML2\_T2.A1.R |
| *Fit your best set of models for the problem SimDataCIRCLE, using the dataset in SimDataCIRCLETR.dat. After that, check the “generalization error” of all those models using the dataset SimDataCIRCLETS.dat, avoiding refitting models as result of this final comparison.*  *This problem has 6 input variables and one output Y, with values I (inside the circle), or O (outside the circle). The TR dataset has 500 observations whereas the TS has 10000.*  *Note that Y has been obtained from the inputs in a pure deterministic way, i.e. given the inputs, the output is perfectly determined.* | |
| **EXPLORATORY ANALYSIS**  **TASK 2** | **Include here a basic exploratory analysis of the problem.**  **Justify your answers, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 2.1** | ***Summarize the accuracy of the different models in the TR and TS datasets. Include some measure of complexity of the resultant model.***  ***What is the best model? Consider both complexity and accuracy***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 2.2** | ***What are the two more important variables according to the different models? Is X3 more important than X6?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 2.3** | ***Create some new variables derived from the raw ones. The idea is to build some variable/s that simplify the classification process using trees.***  ***Create a simple tree and compare with the previous one. Is it better? Why?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **TASK 3** | USING GENETIC ALGORITHMS Use the script AssML2\_T3.A1.R |
| *Use a genetic algorithm to estimate the two parameters defining a straight line (i.e. a simple linear regression model).*  *In the script AssML2\_T3.A1.R you can find the basis for completing this task. Note that you can use the R function “lm” in order to fit a regression line to the data by ordinary least squared (OLS), see: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm*  *In particular, you should analyze the obtained results using GAs, comparing with the theoretical target line and the line obtained by OLS. Impact of main hyperparameters of the GA in the results are also very interesting.* | |
| **EXPLORATORY ANALYSIS**  **TASK 3** | **Include here the parameters defining the true underlying straight line of the problem (slope and intercept), as well as the estimated parameters using OLS regression with the R function lm.**  **Justify your answers, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.1** | ***Fit the regression line using a GA. Set the hyperparameters to obtain a good result in a reasonable number of iterations. Include in the answer your set of GA’s parameters and the best linear model obtained using GA (summary(GA)).***  ***Show a plot with the individuals in a generation where the population is very different and in another generation where it is more stable.***  ***Show the fitness evolution. Discuss why your set of GA params is reasonable.***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your answer)*

|  |  |
| --- | --- |
| **QUESTION 3.2** | ***Compare the OLS model and the GA model.***  ***What is the best model? Why?***  ***Quantify in only one number the quality of each solution and, according to such as quantity, explain why one model is better than the other is.***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.3** | ***Analyze the importance of the mutation probability (pmutation) in this problem. Compare three cases: no mutation, pmutation = 0.1 and pmutation = 0.8***  ***Describe the impact of this probability in the result. Discuss the effect of using pmutation = 0 and pmutation = 0.8***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.4** | ***Change your GA implementation to fit the linear model using, instead the MSE, the MAE.***  ***Is this new GA model better than the previous GA model? Why?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.5** | ***Using your GA implementation of question 3.1, modify now the lower and upper limits in the GA from*** “***lower = c(-3, -3), upper = c(8,8)” to “lower = c(-3, -3), upper = c(3,0)”***  ***Is this new GA model better than the previous GA model of question 3.1? Why?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.6** | ***Using your GA implementation of question 3.1, modify now the lower and upper limits in the GA from*** “***lower = c(-3, -3), upper = c(8,8)” to “lower = c(-3, -3), upper = c(0,8)”***  ***Is this new GA model better than the previous GA model of question 3.1? Why?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **QUESTION 3.7** | ***Improve (if possible) your solution of question 3.1 using Hybrid GA (set the optional argument optim = TRUE)***  ***What is the best model? Why?***  **Justify your answer, including relevant information that supports your reasoning (e.g. graphs or tables from your data analysis with R)** |

*(Add here your analysis)*

|  |  |
| --- | --- |
| **R CODE** | ***Include here your R code for each task***  ***Justify the text for easy reading*** |