WIH2001 Semester 2, Session 2021/2022
Exercise for Week 7: Lab Exercise
Submit in Spectrum: 6 May 2022

Tasks concerning a machine learning classification problem in data science.

The tasks presented below will aid you to work through a binary classification problem using R programming. These tasks concern the Wisconsin Breast Cancer Dataset found in the "mlbench" R-package. Each record in the dataset represents one breast cancer tissue sample. The data was collected from University of Wisconsin Hospitals. Below is a summary of the attributes:

i. Sample code number id number.
ii. Clump Thickness.
iii. Uniformity of Cell Size.
iv. Uniformity of Cell Shape.
v. Marginal Adhesion.
vi. Single Epithelial Cell Size.
vii. Bare Nuclei.
viii. Bland Chromatin.
ix. Normal Nucleoli.
x. Mitoses.
xi. Class.

Task 1:
The dataset is called "BreastCancer" available in the "mlbench" package. Write R-code that loads the "mlbench" and "caret" packages into the R environment and load the "BreastCancer" dataset.
Hint: *you may need to install the packages if they are not already installed.*

Task 2:
Write R-code that displays the dimension of the "BreastCancer" dataset.

Task 3:
Write R-code that displays the first 20 recods of the "BreastCancer" dataset.

Task 4:
Write R-code that displays the data types for the attributes in the "BreastCancer" dataset. It should be clear from the result that only the "Id" attribute has the integer data type while the rest of the attributes have the "factor" data type.

Task 5:
Write R-code that removes the "Id" attribute from the "BreastCancer" dataset and also convert the data types of the remaining attributes to "numeric".

Task 6:
Write R-code that displays the summary information of the data in the "BreastCancer" dataset. Comment on the summary information of the data.

Task 7:
Write R-code that displays the distribution of classes in the "BreastCancer" dataset. Comment on the class distribution.

Task 8:
Write R-code that displays the correlation of attributes in the "BreastCancer" dataset.

Task 9:
Write R-code that displays the distribution of individual attributes in the "BreastCancer" dataset.
Hint: *Plot the histogram and density plots and comment on these graphs.*

WIH2001 Semester 2, Session 2021/2022
Exercise for Week 8: Lab Exercise
Submit in Spectrum: 15 May 2022

Task 10:

There are several Linear and Non-Linear machine learning algorithms that could be used for predictive tasks. The Linear Algorithms include Logistic Regression (LG), Linear Discriminate Analysis (LDA) and Regularized Logistic Regression (GLMNET). The Non-Linear Algorithms include k-Nearest Neighbors (KNN), Classification and Regression Trees (CART) and Naive Bayes (NB).

1. For each of the above outlined algorithms, develop classification machine learning models for the breast cancer data set and identify the best performing model.
   Hint:
   The 10-fold cross validation could be applied to this scenario The starter R-code for the 10-fold cross validation as well as the code for the logistic regression model has been provided for you below and your task is to build the rest of the models:

   ```
   # 10-fold cross validation with 3 repeats
   trainControl <- trainControl(method="repeatedcv", number=10, repeats=3)
   metric <- "Accuracy"
   # LG
   set.seed(7)
   fit.glm <- train(Class~., data=dataset, method="glm", metric=metric,
   trControl=trainControl)
   ```

2. Explain how the performance of these models could be improved.
   Hint: *You should answer this question while referring largely to the tasks 7 to 9.*