MSc Data Science and Advanced Computing
CSMDM21 - Data Analytics and Mining

Practical 7 & 8: Estimation of classification accuracy in R

The task of this exercise is to predict the class of iris flowers from the four attributes. Any classification algorithm can be adopted for the task. For simplicity, it is suggested to use a decision tree classifier. The goal of the exercise is to understand the concept of "estimation"of classification accuracy and to practice with the implementation of four estimation methods in R. The estimation of the accuracy should be reliable and ideally with low variance (or standard deviation). The standard methods do not provide a way to compute the standard deviation of the accuracy. A solution is the adoption of multiple random repetitions, where allmodels are trained with the same amount of training data for a fair comparison.

Write R code following the pseudocode for the following methods as provided below.

Practical 7:

- Resubstitution method,
- Hold-out method,

Practical 8:

- 10-fold Cross-Validation (xVal) method and
- Leave-One-Out Cross-Validation (LOOCV) method.

Pseudocode:

```
input data <- read the data set
N <- number of records in the input data
numTrials <- 20
```

 # Resubstitution and Hold-out methods (training on 90% of input data)

```
For each trial t in [1..numTrials]
      test set <- randomly sample 10% of input data
      training set <- input data \ test set
      dt <- build a predictive model from the training set
      resub_accuracy[t] <- apply the model dt to the training set
      hold_out_accuracy[t] <- apply the model dt to the test set
end For
compute average and standard deviation of resub_accuracy[]
compute average and standard deviation of hold_out_accuracy[]
```

 # 10-fold xVal method on all input data (training on 90% of input data)

```
For each trial t in [1..numTrials]
      shuffle the input data
      correct_predictions <- 0
      create 10 partitions of the input data
      For each fold f in [1..10]
            dt <- build a predictive model from all partitions but f
            correct_predictions += apply the model dt to the partition f
      end For
      xVal_accuracy[t] <- correct_predictions / N
end For
compute average and standard deviation of xVal_accuracy[]
```

 # LOOCV method on a sample of the data (training on 90% of input data)

```
sampleSize <- N * 0.9 + 1
For each trial t in [1..numTrials]
      data <- randomly sample sampleSize records from input data
      correct_predictions <- 0
      For each i in [1..sampleSize]
            dt <- build a predictive model from all records n data but i
            correct_predictions += apply the model dt to the record i
      end For
      loocv_accuracy[t] <- correct_predictions / sampleSize
end For
compute average and standard deviation of loocv_accuracy[]
```