Assignment 4

Machine Learning and Big Data for Economics and Finance

Exercise 1. (60 points)

For the following questions, write `R` functions that complete the required task and execute each function on some random inputs. Each function needs to match the input and output arguments indicated in the question. Each set of random inputs must correspond to an appropriate scenario for testing each function and should be thoroughly justified. The code in each function must be explained mathematically and fully justified as well.

1. **Task**: Creation of a basis of step functions for nonparametric supervised learning.
   Inputs: Data vector $\boldsymbol{x}$ with n observations; number of cut-off points K.
   Output: Matrix of basis functions.
2. **Task**: First iteration in a hierarchical clustering algorithm.
   Inputs: Matrix $\boldsymbol{X}$ of n observations in p variables.
   Output: List where each element contains the indices of the $n-1$ clusters.
3. **Task**: Linear aggregation of M classifiers.
   Inputs: Logical vector of size M containing the binary predictions from the M classifiers; Numeric vector of size M containing the weights.
   Output: Logical variable containing the prediction from the aggregate classifier.
4. **Task**: Computation of posterior probabilities for quadratic discriminant analysis in a classification problem with one input variable.
   Inputs: A point x corresponding to the value taken by the input variable; vector $\boldsymbol{\pi}$ of size K containing the prior probabilities; vector $\boldsymbol{\mu}$ of size K containing the sample averages for each class; vector $\boldsymbol{\sigma}$ of size K containing the sample standard deviations for each class.
   Output: Vector of posterior probabilities.

Exercise 2. (30 points) Describe mathematically what the following code does. Add comments to each line describing what the line does. Design a scenario where this code could be used and write that scenario in `R`.

```
c_k = function(d,k){
  n = dim(d)[1]
  X = d[,1]
  Y = d[,2]
  kf = 10
  ck = rep(0,kf)
  for (i in 1:kf){
    ii = ceiling(1 + n*(i-1)/kf)
    ii2 = ceiling(n*i/kf)
    tt = ii:ii2
    tr = setdiff(1:n,tt)
    bh = sum(X[tr]*Y[tr])/sum(X[tr]^2)
    yh = X[tt]*bh
    ck[i] = mean( ( Y[tt] - yh )^2 )
  }
  return(mean(ck))
}
```

Exercise 3. (10 points)

We are interested in predicting a variable Y given another variable X. Write an `R` function that returns the optimal tree stump for that supervised learning exercise. Apply that function to the data generated by the model $Y_i = X_i - Z_i$ for $i = 1, ..., 10$ where X_i and Z_i are independent uniform random variables.

Note: No additional `R` packages that require loading using `library()` may be used in this assignment.