**Data Mining &Machine Learning**

### This practical work concerns the following datasets :

|  |  |
| --- | --- |
| Data set | Description |
| Auto | Gas mileage, horsepower, and other information for cars. |
| Carseats | Sales of child car seat |
| Trees of Paris city | https://opendata.paris.fr/explore/dataset/les-arbres/analyze |

### Delivery : you must send your file (.R extension) + images on one file (zip) to [bruno@seznec.net](mailto:bruno@seznec.net) and for archiving on moodle platform. The images the main plot of the exercices and screenshot of relevant parts.

### Quizz

### 1°) Gives two Use-Case of PCA algorithm

### 2°) Gives 2/3 main characteristics and advantages of the Spark framework ?

### 3°) Numerical application :

Using the model of logistic regression (as in the book) with intercept = β0 = 10.6513,

β1 =0.0055 which means that knowing the Balance we can predict a default Rate.

Give the value of the Balance that will give a default rate of 60%. (0.6)

### Linear Regression (Chap3 of the book)

This question involves the use of simple linear regression on the Auto data set. Use the command read.table() to import the dataset *Auto.txt*

1. Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results.
2. Is there a relationship between the predictor and the response?
3. How strong is the relationship between the predictor and the response?
4. Is the relationship between the predictor and the response positive or negative?
5. What is the predicted mpg associated with a horsepower of 98?

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

### Multiple Linear Regression

This question involves the use of multiple linear regression on the Auto data set.

1. Produce a scatterplot matrix which includes all of the variables in the data set.
2. Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.
3. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results.
4. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the year variable suggest?

### Logistic regression (Chap 4 )

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

1. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.
2. Explore the data graphically in order to investigate the associ- ation between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scat- terplots and boxplots may be useful tools to answer this ques- tion. Describe your findings.
3. Perform logistic regression on the training data in order to pre- dict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?
4. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

### Tree based methods (Chap 8)

In this exercise, you will use the Carseats data set in order to run a regression tree and related approaches to predict the Sales variable.

1. Split the data set into a training set and a test set.
2. Fit a regression tree to the training set. Plot the tree, and inter- pret the results. What test error rate do you obtain?
3. Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test error rate?
4. Use the bagging approach in order to analyze this data. What test error rate do you obtain? Use the importance() function to determine which variables are most important. (e) Use random forests to analyze this data. What test error rate do you obtain? Use the importance() function to determine which variables are most important. Describe the effect of m, the num- ber of variables considered at each split, on the error rate obtained.

### Unsupervised learning on Open Data

### On the tree of Paris city dataset, do some EDA, Exploratory Data Analysis, already available on the Paris Open Data platform, and realize a clustering analysis that seems understandable.

### https://opendata.paris.fr/explore/dataset/les-arbres/analyze