Project PJE
(*B*) (2)

Septembre 2020

Chaabane Djeraba (Prof.)

chabane.djeraba@univ-lille.fr

The project is composed of three parts. Part 1 concerns iris flower recognition, based on decision trees. Part 2 completes courses, with new concepts. The third part deals with real scale problem. You implement the three parts, mandatory. In part 2, you select at least one concept (one concept per 2 students). In part 3, you select only one use case. It is possible to work with groups of 1, 2, 3, 4 students. That said, it is recommended to work with groups of 4 students.

Part 1 (6 points) – IRIS recognition

We consider IRIS database. The objective (Task) is to recognize IRIS flower specimen, based on the visual characteristics of the flower (length and width of Sepal and Petal), and to qualify the learning through the 5-Fold cross validation

- Considering the 5-fold cross validation, what is the size of the test sample and the training sample?
- Install the "party" and "Mlmetrics" packages. What are they?
- Realize the function "scoring" of parameter N. N corresponds to N-Fold. The function returns:
    - the decision trees (visualized graphically and displayed in the console)
    - the quality of the decision trees and their average. The quality of the knowledge is calculated by the method MAE (Mean Absolute Error)
- What do you think about the quality of models?

(Indications) The "scoring" function takes the following steps:

- sampling in train.data and test.data,
- display and visualization of the decision tree on the basis of train.data,
- visualization of the prediction table on the basis of the train.data,
- visualization of the prediction table based on test.data

Part 2 (6 points) – Concepts

Concept 1: Clustering

Chapter 6 of *book - reference - R -*

Concept 2: Outlier Detection

Chapter 7 of *book - reference - R -*

Concept 3: Time Series Analysis and Mining

Chapiter 8 of *book - reference - R -*

Concept 4: Association rules

Chapitre 9 of *book - reference - R -*

Concept 5: Text mining

Chapitre 10 of *book - reference - R -*

Concept 6: Social Network Analysis

Chapitre 11 of *book - reference - R -*

Part 3 (8 points)

Case Study I: Analysis and Forecasting of House Price Indices

This case study is on analyzing and forecasting of House Price Indices (HPI). It demonstrates data import from a .CSV file (House-index-canberra, whippet, directory Project – B - 2019), descriptive analysis of HPI time series data, and decomposition and forecasting of the data.

Keywords: Time series, decomposition, forecasting, seasonal component

Main steps:

- Importing HPI Data
- Exploration of HPI Data
- Trend and Seasonal Components of HPI
- HPI Forecasting
- The Estimated Price of a Property

- Discussion

Case Study II: Customer Response Prediction and Profit Optimization

This case study is on using decision trees to predict customer response and optimize profit. To improve customer contact process and maximize the amount of profit, decision trees were built with R to model customer contact history and predict the response of customers. And then the customers can be prioritized to contact based on the prediction, so that profit can be maximized, given a limited amount of time, cost and human resources.

(File Cup98LRN, whippet, directory Project – B - 2019)

Keywords: Decision tree, prediction, profit optimization

Main steps:

- Introduction
- The Data of KDD Cup 1998
- Data Exploration
- Training Decision Trees
- Model Evaluation
- Selecting the Best Tree
- Scoring
- Discussions and Conclusions

Case Study III: Predictive Modeling of Big Data with Limited Memory

This case study is on building a predictive model with limited memory. Because the training dataset was large and not easy to build decision trees within R, multiple subsets were drawn from it by random sampling, and a decision tree was built for each subset. After that, the variables appearing in any one of the built trees were used for variable selection from the original training dataset to reduce data size. In the scoring process, the scoring dataset was also split into subsets, so that the scoring could be done with limited memory. R codes for printing rules in plain English and in SAS format are also presented in this chapter.

(File Cup98LRN, whippet, directory Project – B - 2019)

Keywords: Predictive model, limited memory, large data, training, scoring

Main steps: