Big Data Analytics

Prof. Giacomo di Tollo et al.
Summer Semester 2024
University of Luxembourg
Exercise Sheet #3 – Due on May 30th, 2024, 4PM

Incremental Exercise Sheet

Raw Data Processing **2 Points**

Problem 1.
Let's assume that the file –ClusterBicluster.zip– on Moodle contains -csv files coming from the execution of a parallel task on a cluster. For each file, rows 0 to the one labelled *Ratio Zero to* represent the raw output of our runs.
Each row identifies a city; each column identifies a variable. Our goal is to resume the outcomes of the experiments similarly to the image *ClusterPlot*, in which, for each subplot (representing a file), x-axis represent variables, and y-axis represent variables' values. Each plot line connects variables values representing the variables values associated to a city.
The goal of the exercise is to produce plots from the image *ClusterPlot*, associated to the new set of experiments.

Decision Trees **8 Points**

Problem 2.
The two Excel files named –Before-Pre-Processing– and –After-Pre-Processing– contain information coming from the same set of attributes, whose *raw data* are firstly manually inspected and explored (file –Before-Pre-Processing–) and then prepared and transformed into training data format (file –After-Pre-Processing–).
Please implement a Decision Tree to predict variables $S1Q061$ and $S1Q06P1$ (one tree for each predicted variable) by using all all other variables in the predictor set. Then, please focus on a smaller predictor set, to be decided by the lecturer. Please report the *precision*, *recall*, and other performance metrics discussed during the lectures.
Last, please determine a more autonomous method to determine what variables have to be contained in the predictor set.

Decision Trees (updated) **2 Points**

Problem 3. Please repeat the afore mentioned experiments on the following sets of data:

- Before-Pre-Processing (b), in which $S1Q061$ and $S1Q06P1$ are NOT pre-processed;
- After-Pre-Processing (anotp), in which $S1Q061$ and $S1Q06P1$ are NOT PRE-PROCESSED;
- After-Pre-Processing (ap), in which $S1Q061$ and $S1Q06P1$ are PRE-PROCESSED.

Please perform experiments over 30 different partitions of *training-test* sets. Please provide a table with the following statistics of the performance metrics discussed during the lecture: *min, max, avg, stddev.*
Then please produce two more sets of experiments, by using the following variables in the predictor set: the first is referred to as INVBOOL

- S2Q01-a-201801-19
- S1Q05
- S5Q05-a01 to S5Q05-b22
- S4Q011AL

the second is referred to as INVCONT

- S2Q01-a-201801-19

- S1Q05
- S5Q05-a01 - S5Q05-b22
- S4Q021 to S4Q026

Please summarize the experiments statistics in a table. Then, define a regression model to predict the same variables, but without partitioning data at hand in training and test set. To assess the goodness of the fit, please use the R^2, predicted R^2, adjusted R^2, and $F-test$. Please also assess the significance of the predictors. **Please summarize the experiments statistics in a table**.

Performance metrics **4 Points**

Problem 4. Please read the paper —-Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45, p. 427-437— and understand how the performance measures seen in class may be tailored to multi-class classification. Please redifine all the previously implemented metrics to take into account this concept and compare your previously introduced metrics relative to Problems 2 and 3 (with actual values) to the ones outlined in the paper.

K-Means **14 Points**

Problem 5.
Read the paper —A Similarity Measure for Clustering and Its Applications Guadalupe J. Torres, Ram B. Basnet, Andrew H. Sung, Srinivas Mukkamala, Bernardete M. Ribeiro— in order to assess how different clustering output may be compared. Then, please download the set of data -Dataset for Exercise Sheet- from Moodle. It consists of a set of csv files with the same structure. Each file contains a set of variables that can provide interestings insights for a clustering exercise. You have to perform the following tasks:

1. The variable $Area-layer-name$ can be instantiated only to three values. This implicitly defines partition P_{1aln} composed of three clusters. Please use K-means (k=3) to partition the dataset using variables E,F,J ..., but excluding $Area-layer-name$, to obtain the partition $P_{1kmeans}$. Please compute the similarity between partitions P_{1aln} and $P_{1kmeans}$. Then, repeat this procedure using different variables sets as inputs of the k-means algorithm, and reports the obtained values FOR ALL FILES IN THE REPOSITORY.
2. The variable $period-layer-name-it$ can be instantiated only to three values. This implicitly defines partition P_{2aln} composed of three clusters. Please use K-means (k=3) to partition the dataset using variables E,F,J ..., but excluding $period-layer-name-it$, to obtain the partition $P_{2kmeans}$. Please compute the similarity between partitions P_{2aln} and $P_{2kmeans}$. Then, repeat this procedure using different variables sets as inputs of the k-means algorithm, and reports the obtained values FOR ALL FILES IN THE REPOSITORY.
3. The variable $customer-class-id$ can be instantiated only to three values. This implicitly defines partition P_{3aln} composed of three clusters. Please use K-means (k=3) to partition the dataset using variables E,F,J ..., but excluding $customer-class-id$, to obtain the partition $P_{3kmeans}$. Please compute the similarity between partitions P_{3aln} and $P_{3kmeans}$. Then, repeat this procedure using different variables sets as inputs of the k-means algorithm, and reports the obtained values FOR ALL FILES IN THE REPOSITORY.
4. The variable $visitor-class-id$ can be instantiated only to four values. This implicitly defines partition P_{4aln} composed of three clusters. Please use K-means (k=4) to partition the dataset using variables E,F,J ..., but excluding $visitor-class-id$, to obtain the partition $P_{4kmeans}$. Please compute the similarity between partitions P_{4aln} and $P_{4kmeans}$. Then, repeat this procedure using

different variables sets as inputs of the k-means algorithm, and reports the obtained values FOR ALL FILES IN THE REPOSITORY.

5. The variable $detailed - visitor - class - id$ can be instantiated only to five values. This implicitly defines partition P_{5aln} composed of three clusters. Please use K-means (k=5) to partition the dataset using variables E,F,J ..., but excluding $detailed - visitor - class - id$, to obtain the partition $P_{5kmeans}$. Please compute the similarity between partitions P_{5aln} and $P_{5kmeans}$. Then, repeat this procedure using different variables sets as inputs of the k-means algorithm, and reports the obtained values FOR ALL FILES IN THE REPOSITORY.

6. Please use the decision tree approaches you have devised in Problems $2 - -4$ to predict the variables $Area - layer - name$, $period - layer - name - it$, $customer - class - id$, $visitor - class - id$, $detailed - visitor - class - id$. Please takle two cases: 1) Tackling the overall data set; 2) Use 30 different partitions of train and test sets. Please provide the evaluations statistics you have already used in Problems $2 - -4$ to assess the performances.

7. Please use the regression approaches you have devised in Problem 3 to predict the variables $Area - layer - name$, $period - layer - name - it$, $customer - class - id$, $visitor - class - id$, $detailed - visitor - class - id$. Please takle two cases: 1) Tackling the overall data set; 2) Use 30 different partitions of train and test sets. Please provide the evaluations statistics you have already used in Problems $2 - -4$ to assess the performances.

8. Please perform a correlation (Spearman, Ranked Based, and Mutual Information, also tailored to categorical data), analysis over all the data taken into account in points 6 and 7. Please provide a table with the reported correlations. Please re-make the experiments required in points 7. and 8., deleting the most correlated variables (i.e., if two variables are highly correlated, remove one of them from the predictors set) from the predictor set, and assess whether some improvements are recorded.