MACHINE LEARNING & DATA MINING

SAMPLE PROJECT

Project Description

Partitioning Clustering Part

In this assignment, we consider a set of observations on a number of silhouettes related to different type of vehicles, using a set of features extracted from the silhouette. Each vehicle may be viewed from one of many different angles. The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS, which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness. Four model vehicles were used for the experiment: a double decker bus, Chevrolet van, Saab and an Opel Manta. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

Objectives/Deliverables (partitioning clustering)

One dataset (vehicles.xls) is available and has 846 observations/vehicle samples. This dataset is defined by 18 attributes (i.e. input variables) and one output (i.e. class). There are 4 classes. This is a classic multi-dimensional, in terms of features, problem. For this clustering part, you need to use only the first 18 attributes to your calculations. Clustering is an unsupervised scheme, thus, the information included in the "class" attribute can't be used.

Description of attributes:

1. Comp: Compactness
2. Circ: Circularity
3. D.Circ: Distance Circularity
4. Rad.Ra: Radius ratio
5. Pr.Axis.Ra: pr.axis aspect ratio
6. Max.L.Ra: max.length aspect ratio
7. Scat.Ra: scatter ratio
8. 8. Elong: elongatedness
9. 9. Pr.Axis.Rect: pr.axis rectangularity
10. 10. Max.L.Rect: max.length rectangularity
11. 11. Sc.Var.Maxis: scaled variance along major axis
12. 12. Sc.Var.maxis: scaled variance along minor axis
13. 13. Ra.Gyr: scaled radius of gyration
14. 14. Skew.Maxis: skewness about major axis
15. 15. Skew.maxis: skewness about minor axis
16. 16. Kurt.maxis: kurtosis about minor axis
17. 17. Kurt.Maxis: kurtosis about major axis
18. 18. Holl.Ra: hollows ratio
19. 19. Class: type of cars (desired output)

The work in this part is divided into two subtasks:

• In the 1stsubtask, the analysis will be performed with all initial attributes, as the aim is to assess clusteringresults using all input variables.

• In the 2nd subtask, however, principal component analysis (PCA) will be applied to reduce the input dimensionality and the newly produced dataset will be again clustered. The aim in this 2nd subtask is to help students understand the principles and effects of reducing dimensionality in multi-dimensional problems.

1st Subtask Objectives:

a. Before conducting the k-means, perform the following pre-processing tasks: scaling and outliersdetection/removal and briefly justify your answer. (Suggestion: the order of scaling and outliers removal is important. The outlier removal topic is not covered in tutorials, so you need to explore it yourself). Obviously, you can implement this clustering task without exploring this "outlier" component, however, you will not be awarded the allocated marks for this component!

b. You need then to determine the number of cluster centres via four "automated tools". The "automated tools" should include NBclust, Elbow, Gap statistics and silhouette methods. You need to provide, in your report, the related R-outputs and your discussion on these outcomes.

c. The next step is the kmeans clustering investigation. Using, again, all input variables, perform a kmeans analysis using the most favoured "k" from those "automated" methods. For this k-means attempt, show the related R-based kmeans output, including information for the centres, clustered results, as well as the ratio of between_cluster_sums_of_squares (BSS) over total_sum_of_Squares (TSS). It is also important to calculate/illustrate the BSS and the within_cluster_sums_of_squares (WSS) indices (internal evaluation metrics).

d. Following the kmeans attempt, provide the silhouette plot (another internal evaluation metric) which displays how close each point in one cluster is to points in the neighbouring clusters. Provide the average silhouette width score and your discussion on this plot, which should include your comments on the level of "quality" of the obtained clusters.

2nd Subtask Objectives:

e. As this is a typical multi-dimensional, in terms of features problem, you need also to apply the PCA method to this vehicle dataset. You need to show all R-outputs related to PCA analysis, including eigenvalues/eigenvectors, cumulative score per principal components (PC). Create a new “transformed” dataset with principal components as attributes. Choose those PCs that provide at least cumulative score > 92%. Provide a brief discussion for your choice to choose specific number of PCs.

f. In reality, as we have practically a new dataset, we need to find an appropriate k for our “new” kmeans clustering attempt. Like previously, apply the same four “automated” tools to this new pca-based dataset. You need to provide, in your report, the related R-outputs and your discussion on these “new” outcomes.

g. Using this new pca-dataset, perform a kmeans analysis using the most favoured k from those “automated” methods. For this k-means attempt, show the related R-based kmeans output, including information for the centres, clustered results, as well as the ratio of between_cluster_sums_of_squares (BSS) over total_sum_of_Squares (TSS). It is also important to calculate/illustrate the BSS and the within_cluster_sums_of_squares (WSS) indices (internal evaluation metrics).

h. Following this “new” kmeans attempt, provide the silhouette plot which displays how close each point in one cluster is to points in the neighbouring clusters. Provide the average silhouette width score and your discussion on this plot, which should include your comments on the level of “quality” of the obtained clusters.

i. Following the kmeans analysis for this new “pca” dataset, implement and illustrate the Calinski-Harabasz Index. This is another well-known internal evaluation metric and it has not been covered in tutorial sessions. Provide, a brief discussion on the outcome of this index. Write a code in R Studio to address all the above issues (results/discussion need to be included in your report). At the end of your report, provide also as an Appendix, the full code developed by you for all these tasks. The usage of kmeans R function is compulsory.

Energy Forecasting Part

Buildings represent a large percentage of a country's energy consumption and associated greenhouse gas emissions. The energy needed in order to maintain internal conditions within buildings, is responsible for a significant portion of the overall energy usage and greenhouse emissions. Thus, improving energy efficiency in buildings is of great importance to our overall sustainability. Over the past few decades, a lot of research has been carried out in order to improve building energy efficiency through various techniques and strategies. The forecasting of energy usage in an existing building is essential for a variety of applications like demand response, fault detection & diagnosis, optimization and energy management. This is a typical time-series based application.
Data-driven forecasting models typically include two main approaches; statistical and machine learning based schemes. The statistical approach typically applies a pre-defined mathematical function and has shown good performance for medium to long term energy forecasting. In addition, such models have shown acceptable performance for short-term forecasting of consumption electricity loads. Machine learning approach in contrast, typically applies an algorithmic approach (which may non-linearly transform the data), in order to provide a forecast.

Objectives/Deliverables (Multi-layer Neural Network)

The provided (electricity_consumption.xlsx) file includes daily electricity consumption data for three hours (20:00, 19:00 & 18:00) for the 2018 and partly 2019 periods (in total 470 samples). The objective of this question is to use a multilayer neural network (MLP-NN) to predict the next stepahead (i.e. next day) electricity consumption for the 20:00 hour case. The first 380 samples will be used as the training data, while the remaining ones will be used as the testing set.
The work in this part is divided into two subtasks:

• In the 1st subtask, the one-step-ahead forecasting of electricity consumption will utilise only the "autoregressive" (AR) approach (i.e. time-delayed values of the 20th hour attribute as input variables).

• In the 2nd subtask, however, the one-step-ahead forecasting of electricity consumption will utilise additional input vectors by including information from the 19th and 18th hour attributes. In that case, your NN models could be considered as a "NARX" (nonlinear autoregressive exogenous) style models.

1st Subtask Objectives:

In this specific subtask, utilise only the "autoregressive" (AR) approach, i.e. time-delayed values of the 20th hour attribute as input variables. Experiment with various input vectors up to (t-4) level. According to literature, the electricity consumption forecast depends also on the (t-7) (i.e. one week before) value of the load. Thus, in your "AR" analysis, you need also to investigate the influence of this specific time-delayed load to the forecasting performance of your NN models.

a) As the order of this AR approach is not known, you need to experiment with various (time-delayed) input vectors and for each case chosen, you need to construct an input/output matrix (I/O) for the MLP training/testing (using "time-delayed" electricity loads)

b) Each one of these I/O matrices needs to be normalised, as this is a standard procedure especially for this type of NN. Explain briefly the rationale of this normalisation procedure for this specific type of NN (i.e. why do we need to normalise data before using them in an MLP structure?)

c) For the training phase, you need to experiment with various MLP models, utilising these different input vectors and various internal network structures (such as hidden layers, nodes, linear/nonlinear output, activation function, etc.). For each case, the testing performance (i.e. evaluation) of the networks will be calculated using the standard statistical indices (RMSE, MAE, MAPE and sMAPE – symmetric MAPE).

d) Briefly explain the meaning of these four stat. indices.

e) Create a comparison table of their testing performances (using these specific statistical indices). Add a column in this matrix, where you will provide a brief description of the specific NN structure. As, the number of potential NN structures (with various input vectors and internal structures) that can be created can be huge, in this exercise, restrict your total number of developed NNs to 12-15 models. Obviously, these models will have differences in terms of input vector and internal structure. The main aim of this task, by providing such different models, is to understand how such differences may have effect in the forecasting accuracy.

f) From this comparison table, check the "efficiency" of your best one-hidden layer and two-hidden layer networks, by checking the total number of weight parameters per network. Briefly, discuss which approach/structure is more preferable to you and why.