UOW-logo

Informatics Institute of Technology

Department of Computing

Machine Learning and Data Mining Coursework Report

Module : 5DATA001C.2: Machine Learning and Data Mining

Module Leader : Nipuna Senanayake

Date of submission : 04/05/2023

Student ID : 20212196 / w1902278

Student Name : Thejan Vithanage

Table of Content

[1. PARTITIONING CLUSTERING 1](#_Toc133866333)

[1.1. Pre-processing 1](#_Toc133866334)

[Outliers detection/removal 1](#_Toc133866335)

[Scaling 2](#_Toc133866336)

[1.2. Determining The Optimal Number of Clusters 4](#_Toc133866337)

[NbClust() function 4](#_Toc133866338)

[Elbow Method 6](#_Toc133866339)

[Gap statistic method 7](#_Toc133866340)

[Average silhouette method 7](#_Toc133866341)

[1.3. K-means analysis 8](#_Toc133866342)

[1.4. Silhouette plot 8](#_Toc133866343)

[1.5. PCA 9](#_Toc133866344)

[Eigenvectors 9](#_Toc133866345)

[Eigenvalues 11](#_Toc133866346)

[Cumulative score per principal components (PC) 11](#_Toc133866347)

[1.6. Determine the number of clusters. 11](#_Toc133866348)

[NBclust 11](#_Toc133866349)

[Elbow 11](#_Toc133866350)

[Gap Statistics 11](#_Toc133866351)

[Silhouette 11](#_Toc133866352)

[1.7. K-means analysis for this “PCA”-based dataset 11](#_Toc133866353)

[1.8. Silhouette 11](#_Toc133866354)

[1.9. Calinski-Harabasz index 11](#_Toc133866355)

[2. Multi-layer Neural Network 12](#_Toc133866356)

[2.1. Methods used for defining the input vector in electricity load forecasting 12](#_Toc133866357)

[2.2. Normalization 12](#_Toc133866358)

[2.3. Implement a number of MLPs for the “AR” approach. 12](#_Toc133866359)

[2.4. comparison matrix for the “AR” case 12](#_Toc133866360)

[2.5. issue of “efficiency” 12](#_Toc133866361)

[2.6. “NARX” approach 12](#_Toc133866362)

[References 13](#_Toc133866363)

[Appendices 14](#_Toc133866364)

[Appendix A: Code 14](#_Toc133866365)

[Appendix B: Screenshots 14](#_Toc133866366)

Table of Figures

[Figure 1: Shapiro-Wilk Test used code. 3](#_Toc133866367)

[Figure 2: Shapiro-Wilk Test Result 3](#_Toc133866368)

[Figure 3: Normalized and Non-Normalized data Box plot 3](#_Toc133866369)

# PARTITIONING CLUSTERING

## Pre-processing

Data pre-processing is used to prepare raw data for use in machine learning (ML) and data mining (DM). Because in ML and DM performance has an impact on the input data quality

### Outliers detection/removal

In data points there are significant different data points the can be a impact on a performance of the learning model therefore it need to detect and remove.

For outlier detection and removal, I chose boxplot method it gives an outlier from boxplot function easily. And it can be used in non-normal distribution data sets. After removing outliers and normalizing data then found another outlier.



### Scaling

Scaling is used to transform value of features to a common scale, in min max normalization its normally 0 to 1, This can improve performance in some algorithms.

For the scaling I used min-max normalization because for z-score normalization dataset need to be normal distribution data set, but this vehicle data set not a normal distributed dataset I check it by using Shapiro-Wilk Test (Gardener, 2012) therefore used min-max normalization.

Text

Description automatically generated

Figure 1: Shapiro-Wilk Test used code.



Figure 2: Shapiro-Wilk Test Result

But one of the drawbacks of the min max normalization is its respond to outliers but after removing outliers it is not an issue.



Figure 3: Normalized and Non-Normalized data Box plot

## Determining The Optimal Number of Clusters

### NbClust() function

\*\*\* : The Hubert index is a graphical method of determining the number of clusters.

In the plot of Hubert index, we seek a significant knee that corresponds to a

significant increase of the value of the measure i.e the significant peak in Hubert

index second differences plot.

\*\*\* : The D index is a graphical method of determining the number of clusters.

In the plot of D index, we seek a significant knee (the significant peak in Dindex

second differences plot) that corresponds to a significant increase of the value of

the measure.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 6 proposed 2 as the best number of clusters

\* 13 proposed 3 as the best number of clusters

\* 1 proposed 4 as the best number of clusters

\* 1 proposed 7 as the best number of clusters

\* 3 proposed 10 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 3

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*





Using NbClust function it suggests that 3 is the most favored from this function.

### Elbow Method

To identify the optimal number of clusters elbow method plot number of clusters against Total within sum of square. In this plot where the total within sum of square reduction slow down and flattened significantly that is the suggestion k number of elbow method.



In this chart it reduces slows significantly at k = 3, therefore elbow method suggests that optimal number of clusters 3

### Gap statistic method



In gap statics result its suggest k = 3

### Average silhouette method

In average silhouette method its suggest k = 2.

## K-means analysis

between\_cluster\_sums\_of\_squares (BSS) = 378.6251

total\_sum\_of\_Squares (TSS)= 314.8862

between\_SS / total\_SS = 54.6 %



## Silhouette plot



In silhouet method its average value range -1 to 1 if its 1 then cluster points vary simlir to own cluster -1 its opposite. In this result value its 0.29. Then it’s a cosiderabaly moderate to good cluster.

Silhouette of 812 units in 3 clusters from silhouette.default

(x = kmeans\_fit$cluster, dist = dist(normalized\_df)) :

Cluster sizes and average silhouette widths:

326 253 233

0.2185616 0.3776430 0.3051266

Individual silhouette widths:

Min. 1st Qu. Median Mean 3rd Qu. Max.

-0.04719 0.18871 0.31144 0.29297 0.41134 0.54793

## PCA



### Eigenvectors

> (phi <- vehicle.eigen$vectors)

[,1] [,2] [,3] [,4]

[1,] -0.23624665 -0.07363184 -0.06178410 -0.07375582

[2,] -0.31542815 0.14701255 -0.15377106 0.13105371

[3,] -0.30343195 -0.04445283 0.03935021 -0.09735182

[4,] -0.28321380 -0.19225397 0.08378814 0.15751956

[5,] -0.08980674 -0.20783258 0.04417990 0.40433417

[6,] -0.18493242 -0.08778341 -0.08901054 -0.08283361

[7,] -0.31608426 0.07269880 0.08118065 -0.04326184

[8,] 0.31637030 -0.01502041 -0.08498945 -0.02418334

[9,] -0.33259929 0.09024898 0.07815013 -0.06672657

[10,] -0.26515775 0.11524431 -0.14490752 0.07861075

[11,] -0.27036471 0.06307474 0.09733397 -0.01552691

[12,] -0.31048919 0.07622059 0.07995391 -0.04258020

[13,] -0.24687140 0.19042583 -0.14688582 0.12141565

[14,] 0.04767381 0.50194678 0.06059384 0.01059297

[15,] -0.04280132 -0.02281243 -0.81821562 -0.43341312

[16,] -0.06399448 -0.09591614 0.44473348 -0.74154445

[17,] -0.03608515 -0.49019439 -0.05897039 0.06977418

[18,] -0.09626900 -0.55513911 -0.03631322 -0.01214535

[,5] [,6] [,7] [,8]

[1,] -0.05254921 0.259637485 0.10279334 0.7535298986

[2,] 0.15922746 -0.096664210 -0.42335849 0.0158912733

[3,] 0.10283293 -0.001563963 0.21978435 -0.3052006421

[4,] -0.25747743 -0.192074499 0.17350590 -0.0338941786

[5,] -0.33354761 -0.668620543 0.07702210 0.1544997125

[6,] 0.66770983 -0.298114022 0.41401073 0.1283778683

[7,] -0.08839933 0.119474889 0.10034243 -0.0495882886

[8,] 0.11775123 -0.081292660 -0.13378256 0.1886547694

[9,] -0.07407334 0.142703538 0.09859514 -0.0004919574

[10,] 0.28272256 -0.100885166 -0.35013963 0.2106814391

[11,] -0.15843021 0.143995849 0.08147069 -0.1034792619

[12,] -0.11679801 0.157310396 0.06687641 -0.0054459687

[13,] 0.01785080 -0.094736094 -0.42966332 -0.2377131820

[14,] -0.19011978 -0.120490190 0.01917497 0.3537124618

[15,] -0.30418789 -0.149607205 0.11600908 -0.0624918130

[16,] -0.08423851 -0.359135826 -0.29403135 0.0659982047

[17,] -0.17310278 0.276944275 -0.26194756 0.1191672754

[18,] 0.16133742 0.044338665 -0.17190696 0.0060691552

[,9] [,10] [,11] [,12]

[1,] -0.437683995 -0.196523806 0.009705193 0.024401191

[2,] -0.081213882 0.195505630 0.065119368 -0.048232195

[3,] -0.240328077 -0.184111396 0.744445020 0.259255587

[4,] -0.188178449 0.069223502 -0.132848486 -0.200719863

[5,] 0.003174262 0.041415414 0.059323118 0.119397878

[6,] 0.184654565 -0.202100189 -0.194395173 -0.229643926

[7,] 0.146263284 0.116784591 -0.134867072 0.055655987

[8,] -0.309340881 0.122380102 -0.027654704 0.356466723

[9,] 0.095020283 0.253296857 -0.234387449 0.284280180

[10,] 0.149120445 0.398631123 0.279857255 -0.109780033

[11,] 0.063496698 -0.136385203 0.022367551 -0.161179091

[12,] 0.091589660 0.176688961 -0.187867592 0.182151142

[13,] -0.228480534 -0.605672084 -0.309638115 0.033385013

[14,] 0.517270295 -0.348766345 0.207111233 0.185529475

[15,] 0.065052561 0.051266792 -0.010240594 0.002041453

[16,] -0.040567367 -0.007672065 -0.011846600 -0.095106746

[17,] 0.302650361 -0.164690356 0.208405451 -0.386253691

[18,] 0.310714082 -0.148724957 -0.109793333 0.588741269

[,13] [,14] [,15] [,16]

[1,] -0.198605465 -0.102739808 -0.007299427 -0.004711970

[2,] -0.126959904 0.221139486 -0.535129713 0.431888371

[3,] 0.113765357 -0.002421975 -0.092062508 -0.079246476

[4,] 0.037038972 0.738466464 0.096922799 -0.248116591

[5,] 0.034996620 -0.400364548 -0.022666802 0.115250681

[6,] 0.207899628 -0.010096263 -0.083360677 0.069451642

[7,] -0.022611845 -0.065976317 -0.073351589 0.030024696

[8,] 0.654908076 0.217206226 0.126717929 0.195297190

[9,] 0.299422480 -0.184659005 -0.150587993 -0.357472590

[10,] -0.009730676 -0.051817028 0.513836145 -0.286614306

[11,] 0.107114578 -0.013397831 0.544425720 0.612249690

[12,] 0.235076405 -0.054132508 -0.054079934 0.206244860

[13,] 0.127738789 -0.125892746 0.106942011 -0.219750558

[14,] 0.068079654 0.295752559 -0.068323287 -0.047129078

[15,] -0.014341735 0.002989417 0.020755977 0.019976486

[16,] -0.001072782 -0.026549633 -0.021960166 0.002538822

[17,] 0.441844232 -0.062920666 -0.209427995 -0.069090916

[18,] -0.288775956 0.188130216 0.141090601 0.073811276

[,17] [,18]

[1,] 0.0028516938 -0.0003818185

[2,] 0.1853895567 -0.0149098297

[3,] -0.0455353177 0.0073277752

[4,] -0.0077394564 0.0263544030

[5,] -0.0026695123 -0.0192824509

[6,] 0.0007464104 0.0122779479

[7,] -0.4049486267 -0.7875783428

[8,] -0.0477185214 -0.2164563356

[9,] 0.5930866962 0.0138457572

[10,] -0.1053167211 0.0239983925

[11,] 0.3404048728 -0.0462582325

[12,] -0.5546056439 0.5721621839

[13,] -0.0901525866 -0.0032686836

[14,] 0.0167022331 0.0073378957

[15,] 0.0052689429 0.0025648010

[16,] 0.0002344129 0.0076745736

[17,] -0.0220477318 -0.0305177668

[18,] 0.0446584089 0.0059490489

### Eigenvalues

[1] 4.538704e-01 1.702475e-01 6.583635e-02 5.466125e-02

[5] 4.143633e-02 2.824432e-02 1.488878e-02 8.852277e-03

[9] 5.088701e-03 3.344046e-03 2.701453e-03 2.081054e-03

[13] 1.330168e-03 1.015179e-03 6.711608e-04 5.514777e-04

[17] 2.939714e-04 1.664103e-05

### Cumulative score per principal components (PC)

[1] 0.5307612 0.7298506 0.8068404 0.8707619 0.9192180

[6] 0.9522472 0.9696583 0.9800103 0.9859610 0.9898716

[11] 0.9930307 0.9954643 0.9970198 0.9982070 0.9989919

[16] 0.9996368 0.9999805 1.0000000

No of PCA attributes used for cumulative score more than 92% 6. Therefore, need to use PCA1 to PCA 6.

## Determine the number of clusters.

### NBclust

### Elbow

### Gap Statistics

### Silhouette

## K-means analysis for this “PCA”-based dataset

## Silhouette

## Calinski-Harabasz index

# Multi-layer Neural Network

## Methods used for defining the input vector in electricity load forecasting

There are various methods used for defining the input vectors in electricity load forecasting, and they depend on the type of forecasting model being used. Some common input variables used for electricity load forecasting include historical load data, weather data, calendar data, and time-of-day and day-of-week indicators.

For example, in regression analysis, the input vector typically consists of past observations of the load variable, as well as other variables such as temperature, humidity, and time indicators. In artificial neural network (ANN) models, input vectors may include lagged values of the load variable and other relevant variables, such as temperature, humidity, and calendar variables. In support vector regression (SVR) models , input vectors consist of the same set of variables as in regression analysis. (Groß, et al., 2021)

## Normalization



## Implement a number of MLPs for the “AR” approach.

## comparison matrix for the “AR” case

## issue of “efficiency”

## “NARX” approach

# References

Gardener, M., 2012. *Chapter 5: Data: Distribution | Beginning R: The Statistical Programming Language.* [Online]   
Available at: https://learning.oreilly.com/library/view/beginning-r-the/9781118239377/164303c05.html#x164303c05-para-0107  
[Accessed 30 04 2023].

Groß, A. et al., 2021. Comparison of short-term electrical load forecasting methods for different building types. *Energy Informatics,* 4(3), p. 13.

# Appendices

## Appendix A: Code

Provide a full listing of your files, with well indented and commented code in the following format:

* Filename
* Code

## Appendix B: Screenshots

Provide a screenshot of all your web pages, in the following format:

* Filename
* Screenshot(s)