HW 4: Clustering Analysis

Name: Akshay Chaurasia
SUID: 309898873
Email: ahchaura@syr.edu
Date: 09/21/2019

In this exercise, we will be using Weka to perform clustering on Superstore.arff dataset and then use Tableau to answer questions.

Weka

Import Superstore1.arff into Weka. Use the SimplekMeans algorithm to perform a clustering analysis. Make sure to keep **nine** clusters.

Make sure to include following set of parameters while clustering:

A)

1) Profit.Ratio
2) Category
3) Region
4) Segment

Provide Screen Shot of Clusters (Example)

Solution:

kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 16299.0

Initial starting points (random):

Cluster 0: '31\%','Office Supplies',West,'Home Office'
Cluster 1: '49\%','Office Supplies',East,Consumer
Cluster 2: '-150\%','Office Supplies',Central,Corporate
Cluster 3: '-165\%','Office Supplies',Central,Consumer
Cluster 4: '48\%','Office Supplies',South,Consumer
Cluster 5: '16\%',Furniture,Central,Consumer
Cluster 6: '48\%','Office Supplies',West,Corporate
Cluster 7: '49\%','Office Supplies',West,'Home Office'
Cluster 8: '7\%','Office Supplies',South,Consumer

Missing values globally replaced with mean/mode

HW 4: Clustering Analysis

Name: Akshay Chaurasia
SUID: 309898873
Email: ahchaura@syr.edu
Date: 09/21/2019

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3	4	5	6	7	8
	(9994.0)	(2576.0)	(2420.0)	(1147.0)	(1217.0)	(909.0)	(928.0)	(738.0)	(50.0)	(9.0)
Profit.Ratio	48%	35%	49%	35%	35%	48%	11%	48%	49%	7%
Category	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Furniture	Office Supplies	Office Supplies	Office Supplies
Region	West	West	East	Central	Central	South	West	West	West	South
Segment	Consumer	Home Office	Consumer	Corporate	Consumer	Consumer	Consumer	Corporate	Home Office	Consumer

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3
	(9994.0)	(2576.0)	(2420.0)	(1147.0)	(1217.0)
Profit.Ratio	48%	35%	49%	35%	35%
Category	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies
Region	West	West	East	Central	Central
Segment	Consumer	Home Office	Consumer	Corporate	Consumer

4	5	6	7	8
(909.0)	(928.0)	(738.0)	(50.0)	(9.0)
48%	11%	48%	49%	7%
Office Supplies	Furniture	Office Supplies	Office Supplies	Office Supplies
South	West	West	West	South
Consumer	Consumer	Corporate	Home Office	Consumer

HW 4: Clustering Analysis

Name: Akshay Chaurasia
SUID: 309898873
Email: ahchaura@syr.edu
Date: 09/21/2019

Time taken to build model (full training data) : 0.14 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	2576 (26%)
1	2420 (24%)
2	1147 (11%)
3	1217 (12%)
4	909 (9%)
5	928 (9%)
6	738 (7%)
7	50 (1%)
8	9 (0%)

Now use Tableau to answer following Questions. To answer question:

1. Open Tableau Workbook
2. https://public.tableau.com/profile/prof.stephen.wallace4806#!/vizhome/IST407-707/Customers?publish=yes
3. Set filters to match Cluster
4. Answer questions (Hover over data point)

1) Person(s) with highest profit ratio?

➔ Russell Applegate (50.0%)

2) Person with lowest profit ratio?

➔ Beth Paige (-83.3%)

3) Which region had the highest profit ratio for all years? What was it?

➔ West (23.3%)

4) Which region had the highest profit ratio in 2016? What was it?

➔ West (29.8%)

5) Which region had the lowest profit ratio in 2016 in the technology and home office category? What was it?

➔ West (7.1%)

HW 4: Clustering Analysis

Name: Akshay Chaurasia
SUID: 309898873
Email: ahchaura@syr.edu
Date: 09/21/2019

B) Continuing to use the Superstore1.arff in Weka. Use the SimplekMeans algorithm to perform a clustering analysis. Make sure to keep **nine** clusters.

Create a new cluster - Make sure to include following set of parameters while clustering:

1) Sales
2) Category
3) Region
4) Segment

Provide Screen Shot of Clusters

Solution:

kMeans

======

Number of iterations: 3
Within cluster sum of squared errors: 16805.0

Initial starting points (random):

Cluster 0: 'Office Supplies',West,'$148 ','Home Office'
Cluster 1: 'Office Supplies',East,'$23 ',Consumer
Cluster 2: 'Office Supplies',Central,'$11 ',Corporate
Cluster 3: 'Office Supplies',Central,'$2 ',Consumer
Cluster 4: 'Office Supplies',South,'$58 ',Consumer
Cluster 5: Furniture,Central,'$342 ',Consumer
Cluster 6: 'Office Supplies',West,'$13 ',Corporate
Cluster 7: 'Office Supplies',West,'$36 ','Home Office'
Cluster 8: 'Office Supplies',South,'$17 ',Consumer

Missing values globally replaced with mean/mode

HW 4: Clustering Analysis

Name: Akshay Chaurasia
SUID: 309898873
Email: ahchaura@syr.edu
Date: 09/21/2019

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3	4	5	6	7	8
	(9994.0)	(2695.0)	(2339.0)	(1208.0)	(1197.0)	(857.0)	(904.0)	(743.0)	(20.0)	(31.0)
Category	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Furniture	Office Supplies	Office Supplies	Office Supplies
Region	West	West	East	Central	Central	South	West	West	West	South
Sales	$13	$6	$23	$11	$2	$13	$25	$13	$36	$17
Segment	Consumer	Home Office	Consumer	Corporate	Consumer	Consumer	Consumer	Corporate	Home Office	Consumer

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3	4	5	6	7	8
	(9994.0)	(2695.0)	(2339.0)	(1208.0)	(1197.0)	(857.0)	(904.0)	(743.0)	(20.0)	(31.0)
Category	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Office Supplies	Furniture	Office Supplies	Office Supplies	Office Supplies
Region	West	West	East	Central	Central	South	West	West	West	South
Sales	$13	$6	$23	$11	$2	$13	$25	$13	$36	$17
Segment	Consumer	Home Office	Consumer	Corporate	Consumer	Consumer	Consumer	Corporate	Home Office	Consumer

Time taken to build model (full training data) : 0.07 seconds
=== Model and evaluation on training set ===

Clustered Instances

0	2695 (27%)
1	2339 (23%)
2	1208 (12%)
3	1197 (12%)
4	857 (9%)
5	904 (9%)
6	743 (7%)
7	20 (0%)
8	31 (0%)