

Data Mining - Clustering

Decision Supports Systems 2017/18, Lecture 08

Marko Tkalčič

Alpen-Adria-Universität Klagenfurt

Other Data-driven Approaches

• we used data to fit the distributions of the unknown variables

Other Data-driven Approaches

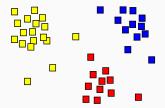
- we used data to fit the distributions of the unknown variables
- machine learning techniques are useful for making predictions
 - supervised:
 - regression (continuous)
 - classification (discrete)
 - unsupervised
 - clustering

Clustering with Weka

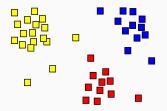
Clustering for decision making

- grouping objects together into clusters (groups)
- object in the same cluster should be **more similar** than objects outside of the cluster
- similarity is key and requires contextual knowledge

- grouping objects together into clusters (groups)
- object in the same cluster should be more similar than objects outside of the cluster
- similarity is key and requires contextual knowledge



- grouping objects together into clusters (groups)
- object in the same cluster should be **more similar** than objects outside of the cluster
- similarity is key and requires contextual knowledge



- there are several algorithms to achieve this task
 - k-means
 - hierarchical models
 - distribution-based models (GMM)

Clustering with Weka

Clustering for decision making

- iris.arff (https://archive.ics.uci.edu/ml/datasets/iris)
- 3 classes of 50 instances each, where each class refers to a type of iris plant.
- One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

- iris.arff (https://archive.ics.uci.edu/ml/datasets/iris)
- 3 classes of 50 instances each, where each class refers to a type of iris plant.
- One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Attribute Information:

- 1. sepal length in cm
- 2. sepal width in cm
- 3. petal length in cm
- 4. petal width in cm
- 5. class:
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica

Running Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- it has an easy GUI
- does not require knowledge of ML
- download Weka from https://www.cs.waikato.ac.nz/ml/weka/
- install

Running Weka

- Weka is a collection of machine learning algorithms for data mining tasks.
- it has an easy GUI
- does not require knowledge of ML
- download Weka from https://www.cs.waikato.ac.nz/ml/weka/
- install
- run and click Explorer



Explorer

- open file
- choose iris.arff

• • •	Weka E	xplorer		
Preprocess Classify Cluste	r Associate Select attributes V	isualize		
Open file Open UR	L Open DB Gener	rate Undo	Edit	Save
Filter				
Choose None				Apply
Current relation		Selected attribute		
Relation: None Instances: None	Attributes: None Sum of weights: None	Name: None Missing: None	Distinct: None	Type: None Unique: None
Attributes				
All None	Invert Pattern			Visualize All
Re	move			
Status				
Welcome to the Weka Explore	r			Log 🛷 x 0

- you can observe the histogram of the variables
- click on Cluster

● ● ● Weka	Explorer		
Preprocess Classify Cluster Associate Select attributes	/isualize		
Open file Open URL Open DB Gene	undo	Edit	Save
Choose None			Apply
Current relation	Selected attribute		
Relation: iris Attributes: 5 Instances: 150 Sum of weights: 150	Name: sepallength Missing: 0 (0%)	Distinct: 35	Type: Numeric Unique: 9 (6%)
Attributes	Statistic	Value	
All None Invert Pattern No. Name 1 sepallength	Minimum Maximum Mean StdDev	4.3 7.9 5.843 0.828	
2 sepalwidth 3 petallength 4 petalwidth 5 class	Class: class (Nom)	28 25	Visualize All
Remove	16 4.3	6.1	10 Z 7.5
ОК			Log 🛷 x 0

Cluster

• • •	Weka Explorer
Preprocess Classify Cluster Associate	Select attributes Visualize
Clusterer	
Choose EM -I 100 -N -1 -X 10 -max -:	1 -II-cv 1.0E-6 -II-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -5 100
Cluster mode	Clusterer output
 Use training set 	
O Supplied test set Set	
O Percentage split %	66
 Classes to clusters evaluation 	
(Nom) class	
Store clusters for visualization	
Ignore attributes	
Start Stop	
Result list (right-click for options)	
Status	
ОК	Log 🗸 x 0

- 1. Choose clustering algorithm (click *choose* -> k-means)
- 2. Choose number of clusters (click on the long list of parameters)
- 3. Remove attributes for clustering (click on Ignore attributes)
- 4. Start
- 5. Visualize cluster assignments (right-click on the chosen result in the resultlist and click *Visualize cluster assignments*)

Marko Tkalčič, DSS-201718-08-DataMining-Clustering

Algorithm

• • •	Weka Explorer
Preprocess Classify Cluster Associate	Select attributes Visualize
Clusterer	
CI Canopy	s 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A
Cobweb	Attribute Full Data 0 (150.0) (50.0)
 FarthestFirst FilteredClusterer HierarchicalClusterer MakeDensity8asedClusterer SimpleKMeans 	sepallength 5.8433 5.936 sepalwidth 3.054 2.77 petallength 3.7587 4.26 petalwidth 1.1987 1.326 class Iris-setosa Iris-versicolor Iris-
C Re	Time taken to build model (full training data) : 0.01 second === Model and evaluation on training set === Clustered Instances 0 50 (33%)
	0 50 (33%) 1 50 (33%) 2 50 (33%) (3%)
Status OK	Log 🛷 x

Parameters

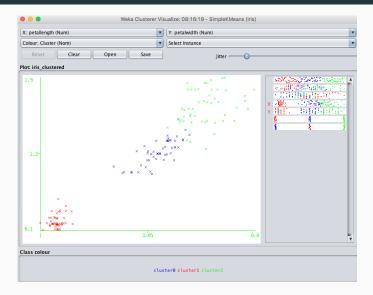
set number of clusters to three

		weka.gui.Gener	icObjectEdit	or		
eka.cluste About	erers.Simp	leKMeans				
toout						
Cluster	r data usin	g the k means algorith	m.		More	
					Capabilit	ies
canopyM	axNumCa	nopiesToHoldInMemor	y 100			
	canopy	MinimumCanopyDensit	v 2.0			
	can	opyPeriodicPruningRate	e 10000			
		canopyT	-1.25			
		canopyr	-1.23			
		canopyT	2 -1.0			
		4.6	False			
		debu	g			
		displayStdDev	False			
		distanceFunctio	n Choose			
		distanceFunctio	n Choose	Euclide	anDistance	-R fir
		doNotCheckCapabilitie	s False			-
		ntReplaceMissingValue	Ealse			v
	doi	ntkeplacemissingvalue	s			
		fastDistanceCal	c False			
		initializationMetho	Random			
		initializationmetrio	Kandom			
		maxiteration	s 500			
		numCluster	s 3			
						77+
			-			
Ope	n	Save	OK		Cance	1

Remove variables for clustering

🔴 🔘 🔵 Select items								
sepallength								
sepalwidth								
petallength								
petalwidth								
class								
Select Pattern Cancel								

Visualize



Clustering with Weka

Clustering for decision making

The management team of a large shopping mall would like to understand the types of people who are, or could be, visiting their mall. They have good reasons to believe that there are a few different **market segments**, and they are considering designing and positioning the shopping mall services better in order to attract mainly a few profitable market segments, or to differentiate their services (e.g. invitations to events, discounts, etc) across market segments.

- gather data
- market survey with potential customers

Name	Description	Scale
V1	Shopping is fun	1-7
V2	Shopping is bad for your budget	1-7
V3	I combine shopping with eating out	1-7
V4	I try to get the best buys while shopping	1-7
V5	l don't care about shopping	1-7
V6	You can save lot of money by comparing prices	1-7
Income	The household income of the respondent	Dollars
Mall.Visits	How often they visit the mall	1-7

Data

ID	V1	V2	V3	V4	V5	V6	Income	Mall.Visits
1	6	4	7	3	2	3	60000	3
2	2	3	1	4	5	4	30000	1
3	7	2	6	4	1	3	70000	3
4	4	6	4	5	3	6	30000	7
5	1	3	2	2	6	4	60000	1
6	6	4	6	3	3	4	50000	2
7	5	3	6	3	3	4	65000	3
8	7	3	7	4	1	4	55000	4
9	2	4	3	3	6	3	70000	0
10	3	5	3	6	4	6	25000	6

Steps in clustering

We will take the following steps

- Select Segmentation Variables
- Define similarity measure
- Method and Number of Segments
- Profile and interpret the segments

Select Segmentation Variables

- critically important decision
- exploratory research usually helps
 - visualization of distributions
 - contextual knowledge, creativity, and experimentation/iterations are needed.
- clustering we use only few variables (V1..V6)
- profiling we use the remaining ones (income, numOfVisits)

Define similarity measure

- clustering = grouping objects based on how similar they are
- similarities:
 - Euclidian
 - Manhattan
 - Cosine
 - • • •

• distance between two objects p and q, each with N variables

$$p = (p_1, p_2...p_N); q = (q_1, q_2, ...q_N)$$
$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_N - q_N)^2}$$

for first 10 subjects

	1	2	3	4	5	6	7	8	9	10
1	0									
2	8	0								
3	3	8	0							
4	6	6	7	0						
5	8	3	9	7	0					
6	2	7	3	4	7	0				
7	2	6	3	5	6	1	0			
8	2	9	2	6	9	3	3	0		
9	7	3	8	6	2	6	5	8	0	
10	7	4	7	2	6	6	6	7	5	0

Method and Number of Segments

- choosing the clustering method and number of clusters:
 - statistical reasoning,
 - judgment,
 - interpretability of the clusters,
 - actionable value of the clusters found,
- In practice different algorithms and numbers of segments should be explored, and the final choice should be made based on both statistical and qualitative criteria.

Method and Number of Segments

- choosing the clustering method and number of clusters:
 - statistical reasoning,
 - judgment,
 - interpretability of the clusters,
 - actionable value of the clusters found,
- In practice different algorithms and numbers of segments should be explored, and the final choice should be made based on both statistical and qualitative criteria.
- method:
 - kMeans
 - hierarchical
- number of clusters
 - 3

interpretation of the characteristics of the clusters

	Population	Cluster 1	Cluster 2	Cluster 3
V1	3.85	5.75	1.67	3.50
V2	4.10	3.62	3.00	5.83
V3	3.95	6.00	1.83	3.33
V4	4.10	3.12	3.50	6.00
V5	3.45	1.88	5.50	3.50
V6	4.35	3.88	3.33	6.00
Income	46000.00	60000.00	42500.00	30833.33
Mall.Visits	3.25	3.25	1.00	5.50

- File open: mall.csv
- repeat steps from the iris example

References

Part of the material has been taken from the following sources. The usage of the referenced copyrighted work is in line with fair use since it is for nonprofit educational purposes.

- http://inseaddataanalytics.github.io/INSEADAnalytics/ CourseSessions/Sessions45/ClusterAnalysisReading.html
- wikipedia.org
- https://archive.ics.uci.edu/ml/datasets/iris