

Chapter 7

K-Means Clustering in R

DATA ANALYSIS IN R

Outline

- Clustering
- K-Means Clustering
- K-Means Clustering in R

Clustering

- Clustering is a technique in machine learning that attempts to find clusters of observations within a dataset.
- The goal is to find clusters such that the observations within each cluster are quite similar to each other, while observations in different clusters are quite different from each other.
- Clustering is a form of unsupervised learning because we're simply attempting to find structure within a dataset rather than predicting the value of some response variable.
- One of the most common forms of clustering is known as k-means clustering.

K-Means Clustering

- K-means clustering is a technique in which we place each observation in a dataset into one of K clusters.
- The end goal is to have K clusters in which the observations within each cluster are quite similar to each other while the observations in different clusters are quite different from each other.

K-Means Clustering

In practice, we use the following steps to perform K-means clustering:

1. Choose a value for K.

First, we must decide how many clusters we'd like to identify in the data. Often we have to simply test several different values for K and analyze the results to see which number of clusters seems to make the most sense for a given problem.

2. Randomly assign each observation to an initial cluster, from 1 to K.

3. Perform the following procedure until the cluster assignments stop changing.

- For each of the K clusters, compute the cluster centroid. This is simply the vector of the p feature means for the observations in the kth cluster.
- Assign each observation to the cluster whose centroid is closest. Closest can be defined using Euclidean distance.

K-Means Clustering in R

Step 1: Load the Necessary Packages

First, we'll load two packages that contain several useful functions for k-means clustering in R.

```
library(factoextra)  
library(cluster)
```

K-Means Clustering in R

Step 2: Load and Prep the Data

For this example we'll use the USArrests dataset built into R, which contains the number of arrests per 100,000 residents in each U.S. state in 1973 for Murder, Assault, and Rape along with the percentage of the population in each state living in urban areas, UrbanPop.

```
#load data
```

```
df <- USArrests
```

```
#remove rows with missing values
```

```
df1 <- na.omit(df)
```

```
#scale each variable to have a mean of 0 and sd of 1
```

```
df1 <- scale(df)
```

```
#view first six rows of dataset
```

```
head(df1)
```

```
> head(df1)
```

	Murder	Assault	UrbanPop	Rape
Alabama	1.24256408	0.7828393	-0.5209066	-0.003416473
Alaska	0.50786248	1.1068225	-1.2117642	2.484202941
Arizona	0.07163341	1.4788032	0.9989801	1.042878388
Arkansas	0.23234938	0.2308680	-1.0735927	-0.184916602
California	0.27826823	1.2628144	1.7589234	2.067820292
Colorado	0.02571456	0.3988593	0.8608085	1.864967207

K-Means Clustering in R

Step 3: Perform K-Means Clustering with Optimal K=4

```
#perform k-means clustering with k = 4 clusters
km <- kmeans(df, centers = 4, nstart = 25)
```

```
#view results
km
```

From the results we can see that:

- 16 states were assigned to the first cluster
- 13 states were assigned to the second cluster
- 13 states were assigned to the third cluster
- 8 states were assigned to the fourth cluster

```
> km
```

K-means clustering with 4 clusters of sizes 13, 13, 16, 8

Cluster means:

	Murder	Assault	UrbanPop	Rape
1	-0.9615407	-1.1066010	-0.9301069	-0.96676331
2	0.6950701	1.0394414	0.7226370	1.27693964
3	-0.4894375	-0.3826001	0.5758298	-0.26165379
4	1.4118898	0.8743346	-0.8145211	0.01927104

Clustering vector:

	Alabama	Alaska	Arizona	Arkansas	California	Colorado
	4	2	2	4	2	2
Connecticut		Delaware	Florida	Georgia	Hawaii	Idaho
	3	3	2	4	3	1
Illinois		Indiana	Iowa	Kansas	Kentucky	Louisiana
	2	3	1	3	1	4
Maine		Maryland	Massachusetts	Michigan	Minnesota	Mississippi
	1	2	3	2	1	4
Missouri		Montana	Nebraska	Nevada	New Hampshire	New Jersey
	2	1	1	2	1	3
New Mexico		New York	North Carolina	North Dakota	Ohio	Oklahoma
	2	2	4	1	3	3
Oregon		Pennsylvania	Rhode Island	South Carolina	South Dakota	Tennessee
	3	3	3	4	1	4
Texas		Utah	Vermont	Virginia	Washington	West Virginia
	2	3	1	3	3	1
Wisconsin		Wyoming				
	1	3				

within cluster sum of squares by cluster:

```
[1] 11.952463 19.922437 16.212213 8.316061
(between_SS / total_SS = 71.2 %)
```

Available components:

```
[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss"
[7] "size" "iter" "ifault"
```

K-Means Clustering in R

- We can use the `aggregate()` function to find the mean of the variables in each cluster:

```
#find means of each cluster
```

```
aggregate(USArrests, by=list(cluster=km$cluster), mean)
```

```
> #find means of each cluster
```

```
> aggregate(USArrests, by=list(cluster=km$cluster), mean)
```

	cluster	Murder	Assault	UrbanPop	Rape
1	1	3.60000	78.53846	52.07692	12.17692
2	2	10.81538	257.38462	76.00000	33.19231
3	3	5.65625	138.87500	73.87500	18.78125
4	4	13.93750	243.62500	53.75000	21.41250

We interpret this output is as follows:

The mean number of murders per 100,000 citizens among the states in cluster 1 is 3.6.

The mean number of assaults per 100,000 citizens among the states in cluster 1 is 78.5.

The mean percentage of residents living in an urban area among the states in cluster 1 is 52.1%.

The mean number of rapes per 100,000 citizens among the states in cluster 1 is 12.2.

Questions