

# K-평균 군집분석

## 메뉴 호출하기

- 고급분석 > 분류분석 > 비지도 학습 > K-평균군집



K-평균 군집분석(cluster analysis)은 N개의 모든 데이터를 최종 K개의 군집으로 군집화시키는 방법입니다. 비지도학습 기법 중 하나로, 유사성을 기초로 한 고정된 수의 군집을 찾습니다. K-평균 군집분석은 양적변수만 가능하며, 분석이 이루어지는 동안 초기 군집에서 다른 군집으로 이동하는 재배치가 가능합니다. 초기 설정된 군집에 의하여 영향을 많이 받는 방법으로서, 초기치의 선택이 최종 군집에 영향을 주므로, 여러 개의 초기치를 대입 및, 여러 번 분석을 반복해야 할 필요가 있습니다.

• 변수설정 탭

K평균 군집분석

변수설정

분석옵션

출력옵션

데이터

전체변수

id

lowbw

preterm

matage

>

<

① 양적변수(선택-1개이상가능)

bweight

gestwks

>

<

② ☒ 양적변수 표준화

③ 질적변수(선택-1개이상가능)

sex

hyp

>

<

도움말

재설정

확인

취소

메뉴 요소	설명
① 양적변수	군집분석에 사용할 변수를 지정합니다. 질적변수는 선택할 수 없습니다. 적어도 하나 이상의 양적변수를 지정해야 합니다.
② 양적변수 표준화	양적변수가 1개 이상 선택된 경우 활성화됩니다. 군집분석 시, 표준화된 데이터 값을 사용합니다.
③ 질적변수	질적변수를 지정합니다. 질적변수로 선택한 변수들은 문자로 인식되어 분석에 사용됩니다. 질적변수로 양적변수를 선택할 수 없으며, 선택된 경우 분석에서 제외됩니다. 질적변수를 선택한 경우 거리행렬 계산 방법으로 Gower의 거리를 사용하게 됩니다.

• 분석옵션 탭

K평균 군집분석

변수설정

분석옵션

출력옵션

① 거리행렬 계산방법

☒ Euclidean
 ☐ Manhattan
 ☐ Maximum
 ☐ Minkowski
 ☐ Gower
 Minkowski power

② 군집의 수

③ 초기값 설정 횟수

④ 알고리즘

☒ Hartigan-Wong
 ☐ Forgy
 ☐ Lloyd
 ☐ MacQueen

⑤ 최대 반복 횟수

도움말

재설정

확인

취소

메뉴 요소	설명
① 거리행렬 계산방법	<p>관측값 간의 거리계산 방법을 지정합니다.</p> <ul style="list-style-type: none"> <li>Euclidean (Default) : 두 점 사이의 거리를 구할 때 가장 많이 쓰는 방식입니다. <math>d = \sqrt{\sum  P_i - Q_i ^2}</math></li> <li>Manhattan : 두 점 사이의 절대적 거리를 이용한 거리 계산 방식입니다. <math>d = \sum  P_i - Q_i </math></li> <li>Maximum : 두 점 사이의 거리가 좌표 차원에서의 가장 큰 벡터공간에서 정의됩니다.</li> <li>Minkowski : power에 입력된 값은 수식 상에 <math>p</math>로 반영됩니다. <math>d = (\sum  P_i - Q_i ^p)^{\frac{1}{p}}</math></li> <li>Minkowski power : Minkowski를 선택할 경우 활성화됩니다. 1 이상의 정수를 입력할 수 있으며, Default는 3입니다.</li> <li>Gower : 질적변수가 포함되어 있을 때 사용 가능한 방법입니다. 양적변수만 존재할 때에도 사용이 가능합니다. 선택된 변수들을 [0, 1] 사이의 값으로 표준화 시킨 후, 모든 변수들 간의 거리를 가중평균하여 합한 값을 사용합니다.</li> </ul>
② 군집의 수	<p>군집의 수를 지정해줍니다. 2 이상의 정수만 입력 가능하며, Default는 2입니다.</p>
③ 초기값 설정 횟수	<p>초기값 설정을 몇 번 재설정 할 것인지 지정합니다. 초기값 설정 횟수가 늘어날수록 분석 시간은 늘어납니다. 1 이상의 정수만 입력 가능하며, Default는 1입니다.</p>

• 분석옵션 탭

K평균 군집분석

변수설정

분석옵션

출력옵션

① 거리행렬 계산방법

☒ Euclidean
 ☐ Manhattan
 ☐ Maximum
 ☐ Minkowski
 ☐ Gower
 Minkowski power

② 군집의 수

③ 초기값 설정 횟수

④ 알고리즘

☒ Hartigan-Wong
 ☐ Lloyd
 ☐ Forgy
 ☐ MacQueen

⑤ 최대 반복 횟수

도움말

재설정

확인

취소

메뉴 요소	설명
④ 알고리즘	<p>거리계산 알고리즘을 지정합니다.</p> <ul style="list-style-type: none"> <li>Hartigan-Wong (Default) : 각 클러스터내의 제곱합이 최소가 되도록 군집을 할당해 주는 방법입니다.</li> <li>Lloyd : 각 클러스터의 무게중심(centroid) 을 이용한 방법입니다. k 개의 임의의 중심점을 할당한 후, 중심점과의 거리를 기준으로 군집을 할당해 줍니다. 그 후 형성된 군집의 무게중심 대응하는 벡터를 중심벡터로 다시 할당해 줍니다. 이 때 더 이상 중심점이 변화하지 않을 때까지 반복 해 줍니다.</li> <li>Forgy : 데이터 집합에서 임의의 k개의 데이터를 선택하여 각 클러스터의 초기값으로 설정합니다. Forgy 알고리즘은 데이터 순서에 대해 독립적입니다. Forgy 알고리즘의 경우 초기 클러스터가 임의의 k 개의 점들에 의해 설정되기 때문에 각 클러스터의 무게중심이 중심으로부터 퍼져 있는 경향을 씁니다.</li> <li>MacQueen : 대체적으로 Lloyd, Forgy와 비슷하나, 각 샘플벡터들이 새롭게 군집으로 할당될 때마다 무게중심이 업데이트 됩니다. MacQueen 알고리즘의 경우, 최종 수렴에 가까운 클러스터를 찾는 것은 비교적 빠르지만, 최종 수렴에 해당하는 클러스터를 찾는 것은 매우 느립니다.</li> </ul>
⑤ 최대 반복 횟수	<p>알고리즘의 최대 반복 횟수(분석 반복 횟수)를 지정합니다. 1 이상의 정수만 입력 가능하며, Default는 10입니다.</p>

• 출력옵션 탭

K평균 군집분석

변수설정 분석옵션 **출력옵션**

① 출력

☐ 기술통계량 ☒ 최종군집중심

☐ 분산분석표 ☒ 최종군집중심간 거리

☒ Silhouette plot

② 최적 군집수 탐색

☐ Within cluster sum of squares

☒ Silhouette

☐ Dunn Index

최대 군집의 수

③ 저장

☐ 최종군집

☐ 최종군집중심으로부터의 거리

도움말 재설정 **확인** 취소

메뉴 요소	설명
① 출력	<p>선택한 내용을 출력합니다.</p> <ul style="list-style-type: none"> <li>기술통계량 : 최종적으로 선정된 군집의 기술통계량을 출력합니다.</li> <li>분산분석표 : 분산분석표를 출력합니다.</li> <li>Silhouette plot : 실루엣 도표를 출력합니다.</li> <li>최종군집중심 : 최종적으로 선정된 군집의 중심점을 출력합니다.</li> <li>최종군집중심간 거리 : 최종적으로 선정된 군집의 중심 간의 거리를 출력합니다.</li> </ul>
② 최적 군집수 탐색	<ul style="list-style-type: none"> <li>Within cluster sum of squares : 군집내 제곱합을 계산하여 최적의 군집 수를 탐색합니다.</li> <li>Silhouette : 실루엣 지표를 계산해 최적의 군집 수를 탐색합니다.</li> <li>Dunn Index : 군집 간 거리의 최소값을 분자, 군집 내 요소 간 거리의 최대값을 분모로 하는 지표입니다. 군집화 결과가 좋을수록 Dunn Index는 커지게 됩니다.</li> <li>최대 군집의 수 : 최적 군집 수 탐색에 사용할 최대 군집 수를 지정합니다. 2 이상의 정수만 입력 가능하며, Default는 10입니다.</li> </ul>
③ 저장	<p>선택한 결과를 괄호 안의 변수명으로 저장합니다.</p> <ul style="list-style-type: none"> <li>최종군집 : 각 관측값이 최종적으로 할당된 군집을 출력한 후 저장합니다. (KMCluster)</li> <li>최종군집중심으로부터의 거리 : 각 관측값과 해당 관측값이 최종적으로 할당된 군집의 중심 사이의 거리를 출력한 후 저장합니다. (KMC_dist)</li> </ul>