

# DBSCAN 군집분석

DBSCAN 군집분석(cluster analysis)은 밀도기반 군집분석으로, 기준이 되는 반경 내에 최소한의 이웃 데이터가 있으면, 하나의 군집(cluster)으로 처리하는 방식입니다. 즉 데이터의 밀도가 높아지는 방향으로 데이터를 군집화 하는 방식으로, DBSCAN 군집분석(cluster analysis)을 위해서는 사용자가 군집 간 거리와, 군집 당 최소 샘플 수를 지정해줘야 합니다. 지정된 군집 간 거리를 이용해 주변공간을 정의한 후, 그 주변공간에 최소의 샘플수가 존재하면 하나의 군집으로 인식합니다. 생성된 군집내의 중심점은 핵심벡터 (core point)로 분류되며, 정의된 군집 간 거리 내에 최소한의 샘플 수 미만의 벡터들이 있어, 어떠한 군집으로도 분류되지 못하는 벡터들의 경우 노이즈 벡터로 분류됩니다.

## 메뉴 호출하기

- 고급분석 > 분류분석 > 비지도 학습 > DBSCAN군집



## • 변수설정 탭

DBSCAN군집

변수설정 분석옵션 출력옵션

① 입력 데이터 형식 (데이터 외의 경우 작업기록 기능에서 제외)

☒ 데이터 ☐ 거리행렬

데이터

전체변수

id  
lowbw  
preterm  
matage  
hyp  
sex  
HCluster  
HC\_dist

② 양적변수(선택-1개이상가능)

bweight  
gestwks

③ ☒ 양적변수 표준화

④ 질적변수(선택-1개이상가능)

도움말 재설정 확인 취소

메뉴 요소	설명
① 입력 데이터 형식	<p>데이터와 거리행렬 두 가지 중 하나를 택합니다.</p> <ul style="list-style-type: none"> <li>데이터 (Default) : 엑셀 스프레드 시트에 있는 데이터에서 변수를 택하여 분석하고자 할 때 선택합니다.</li> <li>거리행렬 : 데이터가 거리행렬인 경우에 사용합니다. 거리행렬을 선택할 경우, [데이터]-'양적변수', '질적변수'가 비활성화 됩니다.</li> </ul>
② 양적변수	<p>군집분석에 사용할 변수를 지정합니다. 질적변수는 선택할 수 없습니다. 적어도 하나 이상의 양적변수를 지정해야 합니다.</p>
③ 양적변수 표준화	<p>양적변수가 1개 이상 선택된 경우 활성화됩니다. 군집분석 시, 표준화된 데이터 값을 사용합니다.</p>
④ 질적변수	<p>질적변수를 지정합니다. 질적변수로 선택한 변수들은 문자로 인식되어 분석에 사용됩니다. 질적변수로 양적변수를 선택할 수 없으며, 선택된 경우 분석에서 제외됩니다. 질적변수를 선택한 경우 [거리행렬 계산 방법]-'Gower'의 거리만 활성화됩니다.</p>

## • 분석옵션 탭

DBSCAN군집

변수설정 [분석옵션] 출력옵션

① 거리행렬 계산방법

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

② 군집간 거리(epsilon)

③ 군집당 최소 샘플수

④ 최단 이웃점 탐색 방법

☒ Euclidean Distances  
☐ KD-tree search  
☐ Linear search

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 군집간 거리 (epsilon)	<p>핵심벡터를 중심으로 하는 반경거리를 지정합니다. 0을 초과한 수를 입력해야 하며, 값이 너무 작을 경우 군집이 하나도 형성되지 않을 수 있고, 값이 너무 클 경우 하나의 군집만 형성될 수 있습니다. Default는 0.15입니다.</p>

• 분석옵션 탭

DBSCAN군집

변수설정 분석옵션 출력옵션

① 거리행렬 계산방법

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

② 군집간 거리(epsilon)

③ 군집당 최소 샘플수

④ 최단 이웃점 탐색 방법

☒ Euclidean Distances  
☐ KD-tree search  
☐ Linear search

메뉴 요소	설명
③ 군집당 최소 샘플 수	한 군집 내에 최소 몇 개의 샘플이 있어야 하는지 지정합니다. 핵심벡터를 중심으로 군집 간 거리 내에 군집 당 최소 샘플 수 이상이 존재하면 하나의 군집으로 형성됩니다. 2 이상의 정수만 입력이 가능하며, Default는 5입니다.
④ 최단 이웃점 탐색 방법	<p>군집간 거리 측정 방법을 선택합니다.</p> <ul style="list-style-type: none"> <li>Euclidean Distances (Default) : 유클리드 거리를 이용해 계산합니다.</li> <li>KD-tree search : 이원 탐색 트리를 다차원 공간으로 확장한 것으로, 트리의 레벨에 따라 차원을 번갈아 가며 비교해 트리를 만듭니다.</li> <li>Linear search : 가장 가까운 지점을 찾기 위해서 항상 다른 모든 지점까지의 거리를 계산합니다.</li> </ul>

- 출력옵션 탭

DBSCAN군집

변수설정 분석옵션 **출력옵션**

**출력**

① ☒ Silhouette plot

② **최적 군집간 거리 탐색**

☒ K-최단 이웃 거리 그림  
k (군집당 최소 샘플수)

③ **저장**

☐ 최종군집  
☐ 최종군집중심으로부터의 거리

도움말 재설정 **확인** 취소

메뉴 요소	설명
① Silhouette plot	실루엣 도표를 출력합니다.
② 최적 군집간 거리 탐색	<ul style="list-style-type: none"> <li>K-최단 이웃 거리 그림 : 최적의 epsilon 값을 찾을 수 있는 K-최단 이웃 거리 도표를 그립니다. 도표에서 꺾이는 팔꿈치(elbow) 부분의 점을 기준으로 최적의 epsilon을 결정합니다.</li> <li>k (군집 당 최소 샘플 수) : 군집 당 최소 샘플 수를 지정합니다. 2 이상의 정수만 입력 가능하며, Default는 5입니다.</li> </ul>
③ 저장	<p>[변수설정]-'데이터' 선택 시 활성화 됩니다. 선택한 결과를 괄호 안의 변수명으로 저장합니다.</p> <ul style="list-style-type: none"> <li>최종군집 : 각 관측값이 최종적으로 할당된 군집을 출력한 후 저장합니다. (DBSCANclus)</li> <li>최종군집중심으로부터의 거리 : 각 관측값과 해당 관측값이 최종적으로 할당된 군집의 중심 사이의 거리를 출력한 후 저장합니다. (DBSCAN_dist)</li> </ul>